МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(национальный исследовательский университет)

ФИЗТЕХ-ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Квалификационная работа на соискание степени магистра
по направлению 01.04.02 «Прикладная математика и информатика»,
магистерская программа «Комбинаторный анализ»

Метод дифференциальной кросс-валидации для выбора уровня сложности обобщенных линейных моделей зависимостей

Выполнил:
студент группы М05-874в
Ангуло Яури Бриан Флориан

Научный руководитель:
д.т.н., профессор
Моттль Вадим Вячеславович

Москва, 2020

Аннотация

Задача восстановаления зависимостей обучающей совокупности в классе моделей возрастающей сложности неизбежно связана с выбором значения структурного параметра, определяющего сложность модели. Наиболее популярные подходы кросс-валидации, в частности Leave-One-Out, страдают от необходимости многократного повторения обучения модели на разных подвыборках обучающей совокупности. В дипломе предложен метод дифференциальной поэлеметной кросс-валидации для обобщенных линейных моделей произвольных зависимостей, который позволяет оценивать модель только один раз с каждым предварительным значением структурного параметра. Идея предложенного метода состоит в удалении бесконечно малой части одного объекта из обучающей совокупности, вместо его полного удаления, как это происходит в Leave-One-Out. Показатель качества модели вычисляется как среднее частных производных ошибок на каждом из отдельных объектов по весам их вхождения в обучающую совокупность. Вычисление показателя качества модели не увеличивает вычислительную сложность процедуры восстановаления зависимости.

Ключевые слова: задача восстановаления зависимостей, класс моделей возрастающей сложности, обобщающая способность, Leave-One-Out, кросс-валидация, отбор признаков, вычислительная сложность.

Оглавление

Обозначения и сокращения

$z(\mathbf{x}\|\mathbf{a})$	обобщенные линейные признаки
$\mathbf{a}$	направляющий вектор
$q(y, z)$	функция связи (функция потерь)
$J(\mathbf{a}\|\gamma, \mu)$	обобщенная линейная модель
$(\mathbf{x}_j, y_j)_{j=1}^N$	обучающая выборка (совокупность)
$\mathbf{x}_j$	вектор признакого описания
y_j	значение целевой переменной
$\mathbf{y}$	вектор целевой переменной
$\mathbf{X}^T$	матрица объекты-признаки
n	число исходных признаков
N	число обучающих объектов
$\mathbf{V}$	матрица вторичных признаков
$\mathbb{I}_{\gamma,\mu}$	подмножество активных признаков
$\mathbb{R}$	множество действительных чисел
$\mathbb{R}^+$	множество неотрицательных действительных чисел
$\boldsymbol{\lambda}$	вектор множителей Лагранжа

$EmpR(\mathbf{a})$	эмпирический риск
LOO	Leave-One-Out
DiffLOO	Дифференциаьная поэлементная кросс-валидация
SVM	Support Vector Machine
L1	LASSO regularization
L2	Ridge regularization

Введение

Селективность признаков — важнейшее свойство в задачах линейных моделей восстановления зависимостей, которое предоставляет возможность отбирать самые релевантные признаки. Особенно это важно в задачах, где число признаков намного превышает числа обучающих объектов $n \gg N$, так как существует континуум моделей, которые одинаково хорошо решают задачу обучения. При этом нахождение оптимальных гиперпараметров крайне необходимо, поскольку от этого зависит качество обучения.

В работе рассматривается именно этот случай, когда число признаков намного больше числа объектов $n \gg N$. Существующие методы для выбора оптимальных гиперпараметров модели, обладающие свойством селективности, страдают от недостатка L1-регуляризации, который приводит к удалению значимых признаков, когда еще не все шумовые признаки отброшены. Для решения этой проблемы используется новый вид регуляризации, предложенный в [4], которая позволяет решать недостатки L1 [11] и L2 [12] регуляризаций по отдельности.

Также, существующие методы имеют полиномиальную вычислительную сложность по числу признаков и линейную вычислительную сложность по числу объектов. Это плохо может сказываться на вычислительной сложности метода, с учетом предположения, что $n \gg N$. Для решения этой проблемы в работе рассматривается алгоритм, который имеет полиномиальную вычислительную сложность по наименьшему числу из числа объектов и числа признаков $\min(n, N)$, что является целесообзраным.

Кроме того, из-за селективности признаков нельзя напрямую использовать процедуру Leave-One-Out с применением формулы Вудбери для одноразового обучения в рамках данного исследования. Поскольку есть предположение, что полное удаление одного объекта, приведет к изменению подмножества активных признаков.

В силу необходимости постоянства подмножества активных признаков, возникает необходимость в ограничении удаления одного объекта. При удалении одного объекта с весом p, стремляющимся к нулю, возникает предположение, что такое подмножество не изменится.

Таким образом, в данном исследовании предлагается новый метод для выбора гиперпараметров модели μ и γ, используя процедуру LOO в измененном виде вместе с применением формулы Вудбери для задач линейной регрессии и логистической регрессии. Что является очень эффективным так как, во-первых, это позволяет провести обучение один раз, а, во-вторых, — подобрать оптимальные значения гиперпараметров модели, в первую очередь гиперпараметр селективности.

Цель работы: разработать новый метод для подбора гиперпараметров модели μ и γ, который

- применим при селективном отборе признаков;
- не увеличивает вычислительную сложность обучения;
- гарантирует подбор оптимальных значений μ и γ для рассматриваемой модели.

Для достижения цели исследования поставлены следующие задачи:

1. дать описание выбора структурных параметров в терминах восстановления зависимости обобщенной линейной модели с использованием селективной Ридж-регуляризации;
2. дать описание понятия взвешенной задачи восстановления зависимости обобщенной линейной модели и вместе с тем, привести вспомогательные понятия и определения, на которые будет опираться метод дифферениальной поэлеметной кросс-валидации;

3. сформулировать и подобрать алгоритм численного решения для выбора активных признаков, на результатах которых будет проводиться метод DiffLOO.

4. сформулировать метод дифференциальной поэлементной кросс-валидации в терминах взвешенной задачи на базе алгоритма численного решения, а также, получить ее критерий качества решения;

5. провести экспериментальное исследование на примере прикладных задач в рамках задач линейной и логистической регрессии, чтобы продемонстрировать эффективность метода и правильность его работы.

Основные положения, выносимые на защиту

1. Математическая формулировка взвешенной обобщенной задачи восстановления зависимостей в линейном пространстве.

2. Алгоритм численного решения для отбора признаков при заданной регуляризации.

3. Концепция дифференциальной поэлементной кросс-валидации для выбора оптимальных гиперпараметров модели.

4. Метод DiffLOO, применимый при селективном отборе признаков, и его критерий качества.

5. Экспериментальное исследование разработанного алгоритма DiffLOO на прикладных задачах.

1. Задача восстановления зависимостей и функция регуляризации

1.1. Обобщенная задача восстановления зависимостей

Классическая задача восстановаления зависимостей заключается в том, что необходимо восстановить неизвестную зависимость скрытой переменной $y \in \mathbb{Y}$, связанной с каким-то объектом реального мира, по наблюдаемому вектору его числовых признаков $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, когда достуна только обучающая совокупность объектов реального мира.

$$\{(\mathbf{x}_j, y_j)_{j=1}^N\},\ \mathbf{x}_j = (x_{j,1}, \ldots, x_{j,n})^T \in \mathbb{R}^n, y \in \mathbb{Y} \tag{1}$$

Единственная разница между регрессией и распознаванием образов состоит в том, что в задаче регрессии целевая переменная — это действительное число $y \in \mathbb{R}$, а в распозновании — категориальное, например, принимает одно из двух значений $y \in \mathbb{Y} = \{-1, 1\}$.

В данной работе будем придерживаться обобщенного линейного подхода для восстановаления зависимостей, под которым понимается комбинация принципа минимизации регуляризованного эмпирического риска Вапника, и идеи обобщенной линейной модели John-a Nelder-a. Понятие обобщенной линейной модели нами будет использоваться как средства, чтобы искомая математическая модель восстановаления исследуемой зависимости не зависела от конкретного масштаба целевой переменной $y \in \mathbb{Y}$.

Следуя [3], математическая модель восстановаления зависимостей произвольного вида рассматривается как пара:

$$\begin{cases} z(\mathbf{x}|\mathbf{a}) = \mathbf{a}^T\mathbf{x} : \mathbb{R}^n \to \mathbb{R} \\ q(y, z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+ \end{cases} \tag{2}$$

Вещественную переменную $z(\mathbf{x}|\mathbf{a}) = \mathbf{a}^T\mathbf{x} \in \mathbb{R}$ будем называть обобщенным линейным признаком объекта реального мира, представленного как числовой вектор признаков $\mathbf{x} \in \mathbb{R}^n$ относительно гиперплоскости в признаковом пространстве $\{\mathbf{x}' \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x}' = 0\} \subset \mathbb{R}^n$, определяемой направляющим вектором $\mathbf{a} \in \mathbb{R}^n$. Обобщенный линейный признак носит смысл положительного или отрицательного расстояния между точкой $\mathbf{x}$ и гиперплоскостью относительно ее положительной или отрицательной стороны.

Функция связи $q(y, z)$ (функция потерь в терминах Вапнкика [1]) выбирается наблюдателем и носит характер штрафа, т.е. как природа будет штрафовать оценку неизвестной переменной $y \in \mathbf{Y}$ для каждого объекта $\mathbf{x} \in \mathbb{R}^n$, представленного как обобщенный линейный признак $z(\mathbf{x}|\mathbf{a})$.

Поскольку фукция связи выбирается наблюдателем, то направляющий вектор гиперплоскости полностью определяет правило принятия решения:

$$\hat{y}(\mathbf{x}, \mathbf{a}) = \arg\min_{y \in \mathbb{Y}} q(y, z(\mathbf{x}|\mathbf{a})) = \arg\min_{y \in \mathbb{Y}} q(y, \mathbf{a}^T\mathbf{x}) \tag{3}$$

Конкретные задачи восстановаления зависимостей различаются между собой только выбором функции связи, а именно:

- Линейная регрессия:

$$y \in \mathbb{R}, q(y, z) = (y - z)^2,\ \hat{y}(\mathbf{x}|\mathbf{a}) = \mathbf{a}^T\mathbf{x} \tag{4}$$

- Логистическая регрессия:

$$y = \pm 1, q(y, z) = \ln[1 + \exp(-yz)],\ \hat{y}(\mathbf{x}|\mathbf{a}) = \begin{cases} 1, \mathbf{a}^T\mathbf{x} \geq 1, \\ -1, \mathbf{a}^T\mathbf{x} < 1. \end{cases} \tag{5}$$

- Метод опорных векторов(SVM):

$$y = \pm 1, q(y, z) = \max(0, 1 - yz),\ \hat{y}(\mathbf{x}|\mathbf{a}) = \begin{cases} 1, \mathbf{a}^T\mathbf{x} \geq 1, \\ -1, \mathbf{a}^T\mathbf{x} < 1. \end{cases} \tag{6}$$

Функция связи для перечисленных задач выпуклая и непрерывная. Для линейной регрессии и логистической регрессии фунция ещё и гладкая, а для SVM — кусочно-гладкая. А это значит, что существуют простые выражения для производных и, почти везде (6), вторые производные

$$q'(y,z) = \frac{\partial}{\partial z} q(y,z), q''(y,z) = \frac{\partial^2}{\partial z^2} q(y,z) \tag{7}$$

1.2. Селективная регуляризация и принцип минимизации эмпирического риска

С точки зрения обобщенного линейного подхода для восстановаления зависимостей, качество параметра гиперплоскости $\mathbf{a} \in \mathbb{R}^n$ это — среднее значение функции потерь $q(y, \mathbf{a}^T\mathbf{x})$ по всем объектам реального мира $(\mathbf{x}, y) \in \mathbb{R} \times \mathbb{Y}$, которую обычно называют средним риском ошибки $AvR(a)$. Однако проблематично посчитать минимизацию среднего риска $AvR(a) \to \min$, потому что гипотетическая вселенная недостижима для непосредственного наблюдения.

Вместо предыдущего подхода, обычно принято приближенно оценивать средний риск по обучающей выборке как среднее арифметическое доступных значений ошибок потерь. Этот известный критерий называется минимизацией эмпирического риска [1], который в наших терминах имеет следующий вид:

$$EmpR(\mathbf{a}) = \frac{1}{N} \sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j) \to \min_{a \in \mathbb{R}} \tag{8}$$

Данная задача оптимизации является выпуклой, если функция связи q(y, z) выбрана также выпуклой.

Когда практическая задача возникает из медицинской или индустриальной областей, доступный объем данных N ограничен, в то время как наблюдатель пытается придумать как можно больше признаков n, опасаясь потерять важные внешние проявления объектов в виде признаков. Таким

образом количество признаков часто превосходит количество обучающих объектов $n \gg N$. Если это так, то задача минимизации эмпирического риска становится некорректной — существует континуум моделей с направляющим вектором $\mathbf{a} \in \mathbb{R}^n$, которые польностью апроксиммируют обучающие данные. Следовательно, для ограничения сложности моделей в работе применяется принип минимизации регуляризованного эмпирического риска с селективной Ридж регуляризацией

$$J(\mathbf{a}|\gamma,\mu) = \gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \le \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum_{j=1}^{N} q\Big(y_j, \sum_{i=1}^{n} a_i x_{j,i}\Big) \to$$

$$\to \min(\mathbf{a} \in \mathbb{R}^n) \quad (9)$$

где $\gamma > 0$ — коэффициент Ридж регуляризации. Что касается $\mu \ge 0$ — это параметр селективности, впервые предложенный в [4], который должен быть подобран по обучающей совокупности (1) вместе с восстановлением направляющего вектора обобщенной линейной модели восстановаления зависимостей. Такая функция регуляризации является выпуклой и гладкой.

Ключевым свойством данной селективной регуляризации состоит в том, что направляющий вектор, оцененный как минимальную точку из (9), $\hat{\mathbf{a}}_{\gamma,\mu} \in \mathbb{R}^n$ будет содержать некоторые элементы, которые в точности равняются нулю.

$$\hat{\mathbf{a}}_{\gamma,\mu} = (\hat{a}_{\gamma,\mu,i}, i = 1,\dots,n) = \arg\min J(\mathbf{a}|\gamma,\mu) \quad (10)$$

Подмножество из $\hat{n}_{\gamma,\mu} \le n$ ненулевых компонентов является подмножеством активных признаков.

$$\hat{\mathbb{I}}_{\gamma,\mu} = i : |\hat{a}_{\gamma,\mu,j}| > 0 \subseteq \{1,\dots,n\}, \hat{n}_{\mathbb{I}} = |\mathbb{I}_{\gamma,\mu}| \quad (11)$$

Это является результатом применения селективной регуляризации. Если $\mu = 0$, то функция регуляризации в (9) совпадает в точности с

обычной Ридж регуляризацией, $\gamma \mathbf{a}^T \mathbf{a} + EmpR(\mathbf{a}) \to \min$ и все элементы направляющего вектора остаются ненулевыми. Но когда параметр селективности возрастает $\mu > 0$, штраф $\mu|a_i|$ в (9) приводит к нулю коэффициенты признаков, то есть элементы направляющего вектора, которые слабо способствуют уменьшению эмпирического риска. А бесконечный рост селективной минимизации $\mu \to \infty$ приводит к обнулению всех коэффициентов.

Необходимость в процессе обучения найти золотую середину делает неизбежным создание инструмента для подбора структурных параметров (γ, μ) в (9), в первую очередь, селективности μ. Опыт показывает, что наиболее популярный метод кросс-валидации LOO отлично подходит для этой цели, но страдает от необходимости многократного обучения, когда каждый объект поочередно удаляется из обучающей совокупности.

В наших интересах целесообразно будет использовать тот факт, что критерий обучения (9) является выпуклым и гладким. Таким образом он допускает точное квадратичное представление в виде двух членов разложения ряда Тейлора при некоторых ограничениях в бесконечно малой окрестности точки минимума $\hat{\mathbf{a}}_{\gamma,\mu} \in \mathbb{R}$. Этот факт позволяет использовать формулу Шермана-Моррисона-Вудбери [6], чтобы избежать многократного решения квадратичной оптимизации при удалении одного объекта из обучающей совокупности.

Кроме того, гладкость функции регуляризации в (9) допускает небольшие изменения направляющего вектора модели зависимости без изменений структуры регуляризованного штрафа. На основе данного факта возникает идея дифференциальной кросс-валидации, когда только небольшая часть объекта исключается из обучения, в отличие от традиционной дискретной версии кросс-валидации LOO.

Алгоритмическая реализация обобщенной селективной задачи минимизации регуляризованного эмпирического риска (9) рассматривалась в [3]. Поскольку и функция регуляризации, и функции связи (4), (5) и (6) явля-

ются выпуклыми и кусочно-гладкими, задача легко поддается численному решению с помощью алгоритмов ньютоновского типа. В [3] показано, что вычислительная сложность таких алгоритмов линейна по числу признаков и полиномиальна по числу объектов обучающей совокупности. Это свойство особенно благоприятно для типичной ситуации, когда количество доступных признаков намного больше количества обучающих объектов.

В рамках исследования рассматриваются задачи восстановления зависимости в задачах линейной регрессии и логистической регрессии. Соответствующие функции связи (4) и (5) являются полностью гладкими и, следовательно, допускают особенно простые алгоритмы DiffLOO. Что касается SVM, функция связи (6) имеет интервал нулевых значений и одну точку негладкости. В результате обучающая совокупность разбивается на три подмножества (периферийные, опорные-граничные, опорные-нарушители) [5], и процедура LOO усложняется, хотя принцип остается точно таким же. В работе не будет рассмотрена задача SVM.

2. Основные идеи и понятия дифференциальной поэлементной кросс валидации

2.1. Взвешенная задача восстановления зависимостей и идея дифференциальной кросс-валидации

Основное отличие дифференциальной поэлементной кросс-валидации от LOO основано на несколько более общей формулировке задачи восстановления зависимостей (9), в которой предполагается, что объекты реального мира встречаются в обучающей совокупности с некоторыми весами $\mathbf{r} = (r_1, \ldots, r_N), 0 \leq r_t \leq 1$:

$$J^r(\mathbf{a}|\gamma,\mu) = \gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \leq \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum_{j=1}^{N} q\Big(y_j, \sum_{i=1}^{n} r_j a_i x_{j,i}\Big) \rightarrow$$

$$\rightarrow \min(\mathbf{a} \in \mathbb{R}^n) \quad (12)$$

При этом в данной работе будет рассматриваться случай, когда веса всех объектов будут равны единице, кроме веса j-го объекта, у которого вес $\mathbf{r}_j = 1 - p$. Такую задачу будем называть взвешенную задачу с заглушенным j-ым объектом.

Лемма 1. *Взвешенную задачу с j-ым заглушенным объектом $r_j = 1 - p$, со значением p, стремляющимся к нулю, можно представить в следующем виде.*

$$J^r(\mathbf{a}|\gamma,\mu) = \gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \leq \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum_{j=1}^{N} r_j q\Big(y_j, \sum_{i=1}^{n} a_i x_{j,i}\Big) \rightarrow$$

$$\rightarrow \min(\mathbf{a} \in \mathbb{R}^n) \quad (13)$$

Направляющий вектор $\hat{\mathbf{a}}^r_{\gamma,\mu} \in \mathbb{R}^n$, оцененный по взвешенной (заглушенной) обучающей выборке $\hat{\mathbf{a}}^r_{\gamma,\mu} \in \mathbb{R}^n = \arg\min J^r(\mathbf{a}|\gamma,\mu) : \mathbb{R}^N \rightarrow \mathbb{R}^n$, является функцией аргумента вектора $r = (r_1, \ldots, r_N) \in \mathbb{R}^N$.

Если $r = \mathbf{1} = (1, \ldots, 1)$, т.е. вектор весов равен единичному вектору, то это — оригинальная формулировка задачи селективной оценки зависи-

мости (9), предложенная в [3]. В этом тривиальном случае все объекты в равной степени участвуют в процессе обучения. Полученные ошибки $q(y_t, \mathbf{x_t}^T \hat{\mathbf{a}}^1_{\gamma,\mu})$ равномерно распределены по обучающей совокупности как лучший компромисс модели.

Но если все веса равны единице кроме веса j-го объекта, для которого вес меньше одного $r_j = 1 - p,\ p \to 0$, то это значит, что этот данный j-ый объект будет меньше учитываться в процессе обучения. Пусть такой вектор весов будет обозначен как $\mathbf{r} = \mathbf{1}^p_j \in \mathbb{R}^N$. Назовем значение p темпом заглушения соответствующего обучающего объекта. Следует ожидать, что ошибка j-го заглушенного объекта будет расти по сравнению со случаем однородных единичных весов $q(y_j, \mathbf{x}_j^T \hat{\mathbf{a}}^{1^p_j}_{\gamma,\mu}) - q(y_j, \mathbf{x}_j^T \hat{\mathbf{a}}_{\gamma,\mu}) > 0$, в то время как $\|\hat{\mathbf{a}}^{1^p_j}_{\gamma,\mu} - \hat{\mathbf{a}}^1_{\gamma,\mu}\| \to 0$, в силу бесконечно малого уменьшения веса. Чем хуже показатель качества модели со структурными параметрами (γ, μ), тем больше должна быть эта разница.

Идея критерия DiffLOO для подбора гиперпараметров модели γ и μ заключается в минимизации среднего скоростей роста ошибки $q(y_j, \hat{\mathbf{x}}_j^T \hat{\mathbf{a}}^{1^p_j}_{\gamma,\mu})$ по каждому объекту из обучающей выборки.

$$\begin{cases} DiffLOO(\gamma,\mu) = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N}\left[\frac{\partial}{\partial p} q(y_j, \hat{\mathbf{x}}_j^T \hat{\mathbf{a}}^{1^p_j}_{\gamma,\mu})\right]_{p=0} \to \min(\gamma,\mu) \\ \hat{z}^{\mathbf{1}^p_j}_{\gamma,\mu,j} = \mathbf{x}_j^T \hat{\mathbf{a}}^{\mathbf{1}^p_j}_{\gamma,\mu},\ j = 1,\ldots,N. \end{cases} \tag{14}$$

Следовательно, требуется найти способ быстрого вычисления этих производных по критерию селективной задачи восстановления зависимостей (13).

В крайнем частном случае, когда $p = 1$ и $r_j = 0$, мы получаем обычное LOO обучение. Однако, такой способ верификации структурного параметра плохо совместим с селективной регуляризацией в (9), поскольку полное удаление объекта может привести к изменению подмножества активных признаков и сделать неправильным применение формулы Шермана-Моррисона-

Вудбери. Вместо этого потребуется многократно повторить процедуру обучения для каждого из удаленных объектов.

2.2. Основные и вторичные признаки

Во-первых, для удобства дальнейшего изложения в этом подразделе будут приведены несколько вспомогательных понятий.

В (1) было принято обозначить символом $\mathbf{x}_t \in \mathbb{R}^n$ вектор-столбец признакого описания t-го объекта. Эти признаки следует называть основными потому что потребуется внести еще и понятие вторичных признаков. В наших терминах вектор-строка $x_i \in \mathbb{R}^N$ будет обозначать i-ый основной признак для всех N объектов. Соответственно, матрица $\mathbf{X}$ в свою очередь будет обозначена для обучающей выборки:

$$\mathbf{X}_{n\times N} = (\mathbf{x}_1, \ldots, \mathbf{x}_N),\ \mathbf{X}^T_{N\times n} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \tag{15}$$

Вектор-столбец скалярных произведений вектора основных признаков j-ого объекта и всех остальных векторов признаков из обучающей выборки будет называться вектором-столбцом вторичных признаков $\mathbf{v}_j$. Тогда матрица, которая получается при умножении двух матриц из (14) является матрицей вторичных признаков всех объектов из обучающей выборки:

$$\mathbf{v}_j = \begin{pmatrix} \mathbf{x}_j^T\mathbf{x}_1 \\ \vdots \\ \mathbf{x}_j^T\mathbf{x}_N \end{pmatrix} = \begin{pmatrix} v_{j,1} \\ \vdots \\ v_{j,N} \end{pmatrix} \in \mathbb{R},\ \mathbf{V} = \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \cdots & \mathbf{x}_1^T\mathbf{x}_N \\ \vdots & \ddots & \vdots \\ \mathbf{x}_N^T\mathbf{x}_1 & \cdots & \mathbf{x}_N^T\mathbf{x}_N \end{pmatrix} \tag{16}$$

Эти символы должны быть снабжены дополнительными элементами, которые будут представлять эффект селективности признаков. Решение задачи (12) с тривиальным единичным весом $r_t = 1$ включает в себя поиск подмножества активных признаков $\mathbb{I}_{\gamma,\mu}$ (11), а также как последствие по-

иск усеченного направляющего вектора. Количество активных признаков обозначено в (11) как $n_{\mathbb{I}} < n$, это также размерность вектора активных признаков, который обозначим через $\mathbf{x}_{\mathbb{I},t} \in \mathbb{R}^{n_{\mathbb{I}}}$. Специальное обозначение $\mathbf{X}_{\mathbb{I}}$ требуется и для усеченной матрицы, которая состоит не из всех признаков $\mathbf{x}$, как в (13) в обучающей выборке, а только из активных признаков:

$$\mathbf{X}_{n_{\mathbb{I}} \times N} = (\mathbf{x}_{\mathbb{I},1}, \ldots, \mathbf{x}_{\mathbb{I},N}),\ \mathbf{X}^T_{N \times n_{\mathbb{I}}} = \begin{pmatrix} \mathbf{x}^T_{\mathbb{I},1} \\ \vdots \\ \mathbf{x}^T_{\mathbb{I},N} \end{pmatrix} \tag{17}$$

Аналогично (16) вектор-столбец скалярных произведений j-го вектора активных основных признаков и остальных векторов активных основных признаков - это вектор-столбец вторичных признаков $\mathbf{v}_j$ относительно соответствующего j-го объекта. После произведения матриц $X^T_{\mathbb{I}} X_{\mathbb{I}}$ получается симметричная матрица, которая имеет ту же самую размерность $N \times N$ независимо от результата селективности признаков. Такая матрица — это множество векторов вторичных признаков:

$$\mathbf{v}_j = \begin{pmatrix} \mathbf{x}^T_{\mathbb{I},j}\mathbf{x}_{\mathbb{I},1} \\ \vdots \\ \mathbf{x}^T_{\mathbb{I},j}\mathbf{x}_{\mathbb{I},N} \end{pmatrix} = \begin{pmatrix} v_{j,1} \\ \vdots \\ v_{j,N} \end{pmatrix} \in \mathbb{R}^N,\ \mathbf{V}_{\mathbb{I}} = \begin{pmatrix} \mathbf{x}^T_{\mathbb{I},1}\mathbf{x}_{\mathbb{I},1} & \cdots & \mathbf{x}^T_{\mathbb{I},1}\mathbf{x}_{\mathbb{I},N} \\ \vdots & \ddots & \vdots \\ \mathbf{x}^T_{\mathbb{I},N}\mathbf{x}_{\mathbb{I},1} & \cdots & \mathbf{x}^T_{\mathbb{I},N}\mathbf{x}_{\mathbb{I},N} \end{pmatrix} \tag{18}$$

2.3. Обобщенные линейные признаки для объектов из обучающей выборки и обратная функция связи

Рассмотрим формулировку взвешенной задачи, эквивалетную задаче (13), в разделенной записи через промежуточное понятие обобщенных линейных признаков $z_t \in \mathbb{R}$ (2) из обучающей совокупности.

$$\begin{cases} \gamma \sum\limits_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \le \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum\limits_{t=1}^{N} q(y_t, z_t) \to min(a \in \mathbb{R}^n, z_1, \ldots, z_N) \\ z_t = \mathbf{a}^T \mathbf{x_t}, t = 1, \ldots, N \end{cases} \tag{19}$$

Понятие обратной функции связи, введенное в [3], лежит в основе необходимого и достаточного условия минимума для задачи оптимизации (13) как седловая точка двойственной задачи выпуклой оптимизации в терминах множителей Лагранжа $\lambda_t \in \mathbb{R},\ t = 1, \ldots, N$, с ограничениями $z_t = \mathbf{a}^T \mathbf{x_t}$ в (19). Обратная функция связи определяется следующим образом:

$$\varphi(y_t, \lambda_t|\gamma) = -\min_{z_t \in \mathbb{R}} \Big(\frac{1}{2\gamma} q(y_t, z_t) + \lambda_t z_t \Big),\ \lambda_t \in \mathbb{R} \tag{20}$$

В [3] доказано, что обратная функция связи выпуклая в $\lambda_t \in \mathbb{R}$, если функция связи $q(y_t, z_t)$ является выпуклой в $z_t \in \mathbb{R}$

Теорема 1. *В частном случае линейной регрессии фукнция связи* (4) $q(y_t, z_t) = (y_t - z_t)^2,\ y_t \in \mathbb{R}$

$$\varphi(y_t, \lambda_t|\gamma) = \frac{1}{2}\gamma\lambda_t^2 - y_t\lambda_t \tag{21}$$

В частном случае логистической регрессии в двух классовом распознавании образов (5) $q(y_t, z_t) = \ln[1 + \exp(-y_t z_t)], y_t = \pm 1$

$$\varphi(y_t, \lambda_t|\gamma) = \frac{1}{2\gamma}\Big[\big(2\gamma y_t\lambda_t\big) \ln\big(2\gamma y_t\lambda_t\big) + \big(1 - 2\gamma y_t\lambda_t\big) \ln\big(1 - 2\gamma y_t\lambda_t\big) \Big] \tag{22}$$

2.4. Двойственная запись взвешенной селективной задачи восстановления зависимостей

Задача выпуклого программирования (17) с $n+N$ переменными может быть сформулировано в двойственной форме относительно N переменных Лагражна.

Это следует из теоретической структуры, разработанной в [3], что значения переменных Лагранжа непосредственно определяют подмножество активных признаков (11) относительно $\mathbf{x}_i \in \mathbb{R}^N$ (15):

$$\mathbb{I}(\gamma, \mu) = \big\{ i : |\mathbf{x}_i^T \lambda| > \mu \big\} \subseteq 1, \ldots, n \tag{23}$$

Теорема 2. *Двойственная запись критерия селективной задачи восстановления зависимостей* (13) *и* (19) *относительно понятий* (17), (20) *и* (23) *— это дважды дифференцируемая выпуклая задача*

$$W(\lambda|\gamma,\mu) = \frac{1}{2}\boldsymbol{\lambda}^T\Big(\sum_{i\in\mathbb{I}_{(\gamma,\mu)}}\mathbf{x}_i\mathbf{x}_i^T\Big)\boldsymbol{\lambda} + \sum_{t=1}^{N}\varphi(y_t,\lambda_t|\gamma) \to \min(\boldsymbol{\lambda}\in\mathbb{R}^N) \quad (24)$$

Решение задачи $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\in\mathbb{R}^N, \hat{\mathbf{z}}_{\gamma,\mu}\in\mathbb{R}^N$ определяет:

- каждый компонент из направляющего вектора $\hat{\mathbf{a}}_{\gamma,\mu} = (\hat{a}_{\gamma,\mu,1},\ldots,\hat{a}_{\gamma,\mu,n})^T = \arg\min J(\mathbf{a}|\gamma,\mu)$ (13), независимо от остальных:

$$\hat{a}_{\gamma,\mu,i} = \begin{cases} 0, & |\mathbf{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}| \le \mu \\ \mathbf{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}, & |\mathbf{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}| > \mu \end{cases} \quad (25)$$

- подмножество активных признаков: $\hat{\mathbb{I}}_{\gamma,\mu} = \{i : |\mathbf{x}_i^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}| > \mu\}$

- модели линейных объектов $\hat{\mathbf{z}}_{\gamma,\mu} = (\hat{z}_{\gamma,\mu,1},\ldots,\hat{z}_{\gamma,\mu,n})^T \in \mathbb{R}^N$ в искомой зависимости (2) в соответствие с (17) и (18):

$$\hat{\mathbf{z}}_{\gamma,\mu} = \mathbf{X}_{\hat{I}_{\gamma,\mu}}^T\mathbf{X}_{\hat{I}_{\gamma,\mu}}\hat{\boldsymbol{\lambda}}_{\gamma,\mu} = \mathbf{V}_{\hat{I}_{\gamma,\mu}}^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu} \in \mathbb{R}^N \quad (26)$$

Необходимо отметить, что, из (26) очевидно, что вектор множителей Лагранжа — нечто иное, как направляющий вектор линейной модели зависимости в N-мерном линейном пространстве вторичных признаков $\mathbf{V}_{\hat{I}_{\gamma,\mu}} \in \mathbb{R}^N$ (18) объектов обучающей совокупности.

Размерность вторичного линейного пространства, которому принадлежат объекты, всегда остается неизменной. Однако количество основных признаков, которые образуют вторичные признаки как скалярное произведение между собой, зависит от результата селективности признаков.

Полное доказательство опущено, поскольку эта теорема фактически содержится в более общей теореме, доказанной в [3]. Двойственная запись задачи (22) подходит для регрессии (4) и для логистической регрессии (5), но

недостаточно для SVM (6). Соответствующее условие в [3] содержит также систему ограничений неравенств $\lambda_{min,t} \leq \lambda_t \leq \lambda_{max,t}$, которая покрывает все типы выпуклых функций связи $q(y_t, z_t)$

Мы приведем только лемму, которая существенным образом лежит в основе доказательства и непосредственно показывает итерационную процедуру численного решения двойственной задачи.

Лемма 2. *Задача* (24) *эквивалетна поиску седловой точки Лагранжа при ограничениях задачи* (19)*:*

$$-\frac{1}{2}\boldsymbol{\lambda}^T\Big(\sum_{i\in\mathbb{I}_{(\boldsymbol{\lambda}|\mu)}}\mathbf{x}_i\mathbf{x}_i^T\Big)\boldsymbol{\lambda}+\frac{1}{2\gamma}\sum_{t=1}^{N}r_t q(y_t,z_t)+\mathbf{z}^T\lambda\to\begin{cases}\nabla_\lambda=0\in\mathbb{R}^N\\ \min(\mathbf{z}\in\mathbb{R}^N)\end{cases}\tag{27}$$

2.5. Итерационный алгоритм Ньютона для численного решения задачи восстановления зависимостей

Пусть первый член в (24) всюду выпуклый в $\mathbb{R}^N$ как сумма квадратичных функций $\sum_{i\in\mathbb{I}(\boldsymbol{\lambda},\mu)}\boldsymbol{\lambda}^T\Big(\mathbf{x}_i\mathbf{x}_i^T\Big)\boldsymbol{\lambda}$, но он не квадратичный поскольку диапазон суммирования завивист от вектора-решения множителей Лагранжа $\boldsymbol{\lambda}$. Обратная функция связи $\varphi(y_t, \lambda_t|\gamma)$ также является выпуклой, но не обязательно является квадратичной. В [3] показано, что обе эти функции дважды дифференцируемы. Таким образом, мы получили выпуклую и дважды дифференцируемую двойственную селективную задачу восстановления зависимостей, которая может быть численно решена с помощью итерационного алгоритма Ньютона.

Пусть $\hat{\boldsymbol{\lambda}}^k \in \mathbb{R}^N$ — текущая точка в итерационном алгоритме, направленная на решение выпуклой задачи двойственной оптимизации $\mathbf{W}(\boldsymbol{\lambda}|\gamma,\mu)\to\min(\boldsymbol{\lambda})$ с подмножеством активных признаков $\mathbb{I}^k=\mathbb{I}(\boldsymbol{\lambda}^k|\mu)$. Тогда согласно (26), на к-ой итерации у нас есть $\mathbf{z}^k=\big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k}\big)\boldsymbol{\lambda}^k$.

Предположим, что лучшее решение на следующем шаге итерации $\boldsymbol{\lambda}^{k+1}\in\mathbb{R}^N$ следует искать из условия минимума квадратичного разложения ряда Тейлора $\tilde{\mathbf{W}}^k(\boldsymbol{\lambda}|\gamma,\mu)$ из критерия $\mathbf{W}^k(\boldsymbol{\lambda}|\gamma,\mu)$ в небольшой окрестности

из $\mathbb{R}^N$ вокруг точки $\boldsymbol{\lambda}^k$:

$$\tilde{\boldsymbol{\lambda}}^{k+1} = \arg\min \tilde{\mathbf{W}}^k(\boldsymbol{\lambda}|\gamma, \mu) \tag{28}$$

Матрица в первом пункте в (24) — это сумма матриц $\mathbf{X}^T_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k} = \sum_{i\in\mathbb{I}_{(\boldsymbol{\lambda}|\mu)}} \mathbf{x}_i\mathbf{x}_i^T$, определенная в (15) и (17), где диапазон суммирования зависит от подмножества активных признаков $i \in \mathbb{I}(\boldsymbol{\lambda}|\mu)$. Поскольку это подмножество определяется строгим неравенством в (23), то диапазон суммирования остается фиксированным в окрестности текущей точки $\hat{\boldsymbol{\lambda}}^k$. Разложение Тейлора для квадратичного представления с изменяющейся матрицей сводится к фиксации подмножества $i \in \mathbb{I}(\boldsymbol{\lambda}|\mu) = \mathbb{I}^k(\boldsymbol{\lambda}^k|\mu)$.

$$\boldsymbol{\lambda}^T\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda} \cong \boldsymbol{\lambda}^T\Big(\mathbf{X}^T_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}\Big)\boldsymbol{\lambda} \tag{29}$$

Каждое слагаемое второго члена в (24) будет квадратичной функцией, если подставить вместо обратных функций связи $\varphi(y_t, \lambda_t|\gamma), t = 1, \ldots, N$, их разложения ряда Тейлора $\tilde{\varphi}(y_t, \lambda_t|\gamma)$ в небольшой окрестности текущей точки $\boldsymbol{\lambda}^k \in \mathbb{R}^N$. В свою очередь, каждое разложение $\tilde{\varphi}(y_t, \lambda_t|\gamma)$ может быть получено путем применения определения (20) к разложению Тейлора $\tilde{q}^k(y_t, z_t)$ для исходной функции связи $q^k(y_t, z_t)$ вокруг соответствующей точки $z_t^k = \mathbf{x}^T_{t,\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}\boldsymbol{\lambda}^k$ (26):

$$\begin{aligned}
&\tilde{q}^k(y_t, z_t) \cong q(y_t, z_t) \\
&\tilde{q}^k(y_t, z_t) = \frac{1}{2}q''(y_t, z_t^k)(z_t - z_t^k)^2 + q'(y_t, z_t^k)(z_t - z_t^k) + \text{const.}
\end{aligned} \tag{30}$$

Итак, сначала мы должны найти разложение ряда Тейлора для каждой функции связи. В исследовании рассматриваются только случаи линейной регрессии (4) и логистической регрессии (5).

Теорема 3. *Для квадратичной функции линейной регрессии* $q(y_t, z_t) = (y_t - z_t)^2$*, разложение инвариатно относительно* $\hat{z}_t^k$*:*

$$q'(y_t, \hat{z}_t^k) = -2(y_t - \hat{z}_t^k), q''(y_t, \hat{z}_t^k) = 2, \tilde{q}^k(y_t, z_t) = (y_t - z_t)^2 \tag{31}$$

В случае логистической регрессии $q(y_t, z_t) = \ln[1 + \exp(-y_t z_t)]$*, выбор центра* $\hat{z}_t^k$ *очень важен для разложения Тейлора.*

$$q'(y_t, \hat{z}_t) = -y_t \frac{\exp(-y_t \hat{z}_t)}{1 + \exp(-y_t \hat{z}_t)}, q''(y_t, z_t) = \frac{\exp(-y_t \hat{z}_t)}{(1 + \exp(-y_t \hat{z}_t))^2},$$
$$\tilde{q}^k(y_{t,t}) = \Big\{ \frac{1}{2} \frac{\exp(-y_t \hat{z}_t^k)}{(1 + \exp(-y_t \hat{z}_t^k))^2} (z_t - \hat{z}_t^k))^2 - y_t \frac{\exp(-y_t \hat{z}_t^k)}{1 + \exp(-y_t \hat{z}_t^k)} (z_t - \hat{z}_t^k) \Big\} + \\ + const \tag{32}$$

Для упращения будет удобно представить функцию как полный квадрат относительно ее минимальной точки:

$$\tilde{q}^k(y_t, z_t) \cong \tilde{g}_t^k (z_t - \tilde{y}_t^k)^2 + \text{const}, \tilde{g}_t^k = \frac{1}{2} \frac{\exp(-y_t \hat{z}_t^k)}{(1 + \exp(-y_t \hat{z}_t^k))^2}$$
$$\tilde{y}_t^k = \hat{z}_t^k + y_t (1 + \exp(-y_t \hat{z}_t^k)) \tag{33}$$

В случае логистической регрессии, если $y_t \hat{z}_t^k \to \infty$ функция связи стремится к нулю $\ln[1 + \exp(-y_t z_t)] \to 0$, а если $y_t \hat{z}_t^k \to -\infty$ ее форма становится более линейной $\ln[1 + \exp(-y_t z_t)] \to -y_t z_t^k$. В обоих пределах его квадратичное представление разложение ряда Тейлора почти вырождено $g_t^k \to 0$ (32). Чтобы избежать вырожденности, поставим:

$$\tilde{q}^k(y_t, z_t) = \tilde{g}_t^k \big(z_t - \tilde{y}_t^k\big), \ \tilde{g}_t^k = \tilde{g}_t^k + \varepsilon, \ \varepsilon > 0 \ \text{— маленькое число} \tag{34}$$

Вид линейной регрессии (31) легко получить из вида логистической регрессии (32), если учесть (35). Значит, достаточно рассмотреть логистическую регрессию.

$$\tilde{g}_t^k = 1, \ \tilde{y}_t^k = y_t. \tag{35}$$

Таким образом, как для регрессии, так и для логистической регрессии разложение функции Тейлора выражается квадратичной функцией той же

структуры (33).

Теорема 4. *Если разложение Тейлора для функции связи квадратично* $\tilde{q}^k(y_t, z_t) = \tilde{g}_t^k\big(z_t - \tilde{y}_t^k\big)^2 + const$, *то разложение Тейлора для обратной функции связи также квадратично:*

$$\tilde{\varphi}(y_t, \lambda_t|\gamma) = -\min_{z_t}\left(\frac{1}{2\gamma}\tilde{q}^k(y_t, z_t) + \lambda_t z_t\right) = \frac{\gamma}{2}\frac{1}{\tilde{g}_t^k}\lambda_t^2 - \tilde{y}_t^k\lambda_t \tag{36}$$

Таким образом, k-й шаг итерации состоит в решении задачи (28) в отношении (24):

$$\tilde{\mathbf{W}}^k(\boldsymbol{\lambda}|\gamma,\mu) = \frac{1}{2}\boldsymbol{\lambda}^T\Big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k}\Big)\boldsymbol{\lambda} + \frac{1}{2\gamma}\sum_{t=1}^{N}\frac{\gamma}{2}\frac{1}{\tilde{g}_t^k}\lambda_t^2 - \tilde{y}_t^k\lambda_t \to \min(\boldsymbol{\lambda}\in\mathbb{R}^N) \tag{37}$$

Примем следующие обозначения:

$$\begin{aligned}\tilde{G}^k &= \mathrm{Diag}(\tilde{g}_1^k, \ldots, \tilde{g}_N^k)\\ \mathbf{y}^k &= (\tilde{y}_1^k, \ldots, \tilde{y}_N^k)\end{aligned} \tag{38}$$

Тогда $\mathbf{W}$ можно представить в следующем виде:

$$\tilde{\mathbf{W}}^k(\boldsymbol{\lambda}|\gamma,\mu) = \frac{1}{2}\boldsymbol{\lambda}^T\Big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k} + \gamma\big(\hat{\mathbf{G}}^k\big)^{-1}\Big)\boldsymbol{\lambda} + (\tilde{y}^k)^T\boldsymbol{\lambda} \to \min(\boldsymbol{\lambda}\in\mathbb{R}^N) \tag{39}$$

Теорема 5. *Требования седловой точки* (27) *на к-том шаге итерации с учетом* (39) *эквивалентны соответствующей системе линейных уравнений для линейной и логистической регрессии, соответственно:*

$$\left(\Big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k}\Big) + \gamma E\right)\hat{\boldsymbol{\lambda}} = \mathbf{y} \tag{40}$$

$$\left(\Big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k}\Big) + \gamma\Big(\hat{\mathbf{G}}^k\Big)^{-1}\right)\hat{\boldsymbol{\lambda}} = \tilde{\mathbf{y}}^k \tag{41}$$

Однако, решение соответствующей системы лйнейных уравнений показывает только наилучшее направление шага Ньютона.

$$\tilde{\boldsymbol{\lambda}}^{k+1} = \left(\Big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k}\Big) + \gamma E\right)^{-1}\mathbf{y} \tag{42}$$

$$\tilde{\boldsymbol{\lambda}}^{k+1} = \left(\Big(\mathbf{X}_{\mathbb{I}^k}^T\mathbf{X}_{\mathbb{I}^k}\Big) + \gamma\Big(\hat{\mathbf{G}}^k\Big)^{-1}\right)^{-1}\tilde{\mathbf{y}}^k \tag{43}$$

Может случиться так, что длина на данном шаге слишком большая, и ее следует сократить. Чтобы проверить необходимо ли это, достаточно сравнить значения двойственного критерия (37) в точках $\hat{\boldsymbol{\lambda}}^k$ и $\hat{\boldsymbol{\lambda}}^{k+1}$ относительно разложения функции в ряд Тейлора (32) и (36):

$$\text{если } W(\tilde{\boldsymbol{\lambda}}^{k+1}|\gamma,\mu,\mathbf{r}) \leq W(\tilde{\boldsymbol{\lambda}}^{k}|\gamma,\mu,\mathbf{r}),\ \boldsymbol{\lambda}^{k+1} = \tilde{\boldsymbol{\lambda}}^{k+1}; \tag{44}$$

$$\text{если } W(\tilde{\boldsymbol{\lambda}}^{k+1}|\gamma,\mu,\mathbf{r}) > W(\tilde{\boldsymbol{\lambda}}^{k}|\gamma,\mu,\mathbf{r}),\ \text{длину следует сократить}; \tag{45}$$

Для нахождения подходящей длины шага Ньютона, следует применить одномерную оптимизацию в (37). В качестве алгоритма будет использоваться алгоритм золотого сечения.

$$\begin{aligned}\tau^{k+1} &= \arg\min W[(1-\tau)\hat{\boldsymbol{\lambda}}^k + \tau\tilde{\boldsymbol{\lambda}}^{k+1}|\gamma,\mu], 0<\tau<1, \\ \hat{\boldsymbol{\lambda}}^{k+1} &= \tau^{k+1}\hat{\boldsymbol{\lambda}}^k + (1-\tau^{k+1})\tilde{\boldsymbol{\lambda}}^{k+1}\end{aligned} \tag{46}$$

На самом деле, алгоритм итеративно запускает подмножества объектов $\mathbb{I}^k = \mathbb{I}(\hat{\boldsymbol{\lambda}}^k) \subset \mathbb{I} = \{1,\ldots,n\}$ (23) без циклов, потому что $W(\tilde{\boldsymbol{\lambda}}^{k+1}|\gamma,\mu,\mathbf{r}) \leq W(\tilde{\boldsymbol{\lambda}}^{k}|\gamma,\mu,\mathbf{r})$ на каждом шаге. Таким образом, условие остановки (47) будет достигнуто после конечного числа шагов:

$$\mathbb{I}(\boldsymbol{\lambda}^{k+1}) = \mathbb{I}(\boldsymbol{\lambda}^k),\ \text{то есть},\ \mathbb{I}^{k+1} = \mathbb{I}^k, \tag{47}$$

Также может оказаться так, что на k-ом шаге число признаков меньше числа объектов $n < N$ в силу свойства селективности. Тогда матрица $\left(\mathbf{X}^T_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}\right)$ размерности $N \times N$ будет иметь ранг равен числу признаков n, а значит, она будет плохо обусловлена. В этом случае обращение матрицы будет численно неустойчиво.

Лемма 3. *В случае когда матрица* $\left(\mathbf{X}^T_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}\right)$ *плохо обусловлена, для численной устойчивости следует применить ее эквивалентную запись:*

$$\tilde{\boldsymbol{\lambda}}^{k+1} = \frac{1}{\gamma}\left[I_{N\times N} - \mathbf{X}^T_{\mathbb{I}^k}\left(\gamma I_{n\times n} + \mathbf{X}_{\mathbb{I}^k}\mathbf{X}^T_{\mathbb{I}^k}\right)^{-1}\mathbf{X}_{\mathbb{I}^k}\right]\mathbf{y} \tag{48}$$

Таким образом Лемма 3 решает проблему численной неустойчивости обращения матрицы в (42). Полученная эквивалентная запись с помощью формулы Вудбери теперь обращает матрицу $\left(\mathbf{X}_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}^T\right)$ полного ранга n (число признаков).

Вектор множителей Лагранжа (43) будет в точности решением взвешенной задачи восстановления зависимости в прямой записи (13) и в двойственной записи (19).

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu} = \hat{\boldsymbol{\lambda}}^{k+1} \in \mathbb{R}^N \tag{49}$$

Что касается задачи логистической регрессии, это условие необходимое, но недостаточное, так как разложение ряда Тейлора для обратной функции связи может измениться ($g_t^{k+1} \neq g_t^k,\ y_t^{k+1} \neq y_t^k$) согласно (33) и (34). Итерационная процедура должна продолжиться с тем же подмножеством активных объектов, пока не будет выполнено достаточное условие:

$$||\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k|| < \varepsilon \text{ — заданное небольшое значение} \tag{50}$$

В итоге, вектор-решение $\hat{\boldsymbol{\lambda}}_{\gamma,\mu} \in \mathbb{R}^N$ сразу определяет направляющий вектор обобщенной линейной модели в прямой записи $\hat{\mathbf{a}}_{\gamma,\mu}$ (10) через двойственное решение (24) и обобщенные линейные признаки всех обучающих объектов (26).

Почти очевидно, что вычислительная сложность решения двойственной задачи (24) полиномиальная по числу обучающих объектов N и линейная по числу признаков n, а это всего лишь вычислительная сложность всей задачи восстановления зависимости.

2.6. Решение задачи восстановления зависимостей

Пусть двойственная задача (24) решена, и $\hat{\boldsymbol{\lambda}}_{\gamma,\mu} \in \mathbb{R}^N, \hat{\mathbf{z}}_{\gamma,\mu} \in \mathbb{R}^N$ — решения задачи. Здесь особенно важен факт, сформулированный равенством (25) в теореме 2, а именно, что направляющий вектор в пространстве

основных признаков $\hat{\mathbf{a}}_{\gamma,\mu}$ является функцией вектора множителей Лагранжа $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}$.

В соответствии с исходной идеей обобщенной линейной модели искомой зависимости (3), направляющий вектор $\hat{\mathbf{a}}_{\gamma,\mu}$ определяет решающее правило в пространстве основных признаков $\hat{y}(\mathbf{x}|\mathbf{a}) = \arg\min_{y\in\mathbb{Y}} q(y, \hat{z}_{\gamma,\mu,t}) = \arg\min_{y\in\mathbb{Y}} q(y, \hat{\mathbf{a}}_{\gamma,\mu}^T\mathbf{x})$, применимое для любого нового объекта, представленного как вектор, составленный из их основных признаков $\mathbf{x} \in \mathbb{R}^n$. В свою очередь согласно (23), (25) и (17), направляющий вектор в пространстве основных признаков выражается через вектор множителей Лагранжа как: $\hat{\mathbf{a}}_{\gamma,\mu} = \mathbf{X}_{\hat{\mathbb{I}}_{\gamma,\mu}}\hat{\boldsymbol{\lambda}}_{\gamma,\mu}$

Это означает, что модель зависимости, содержащаяся в классе обобщенных линейных моделей (2) из обучающей выборки (1), имеет вид:

$$\hat{y}_{\gamma,\mu}(\mathbf{x}|\hat{\boldsymbol{\lambda}}_{\gamma,\mu}) = \arg\min_{y\in\mathbb{Y}} q(y, \hat{z}_{\gamma,\mu,t}) = \arg\min_{y\in\mathbb{Y}} q(y, \mathbf{v}_{\mathbb{I}_{\gamma,\mu},t}^T\hat{\boldsymbol{\lambda}}_{\gamma,\mu}) \qquad (51)$$

Можно сделать следующие промежуточные выводы: во-первых, только вторичные признаки объектов реального мира в подпространстве их основных признаков $\mathbb{R}^{\hat{n}_{\gamma,\mu}} \subseteq \mathbb{R}^n, \hat{n}_{\gamma,\mu} = |\hat{\mathbb{I}}_{\gamma,\mu}|$, имеют значение, когда хотим восстановить скрытую линейную зависимости модели. Во-вторых, только вектор множителей Лагранжа $\hat{\boldsymbol{\lambda}}_{\gamma,\mu} \in \mathbb{R}^N$ определяет линейную модель как направляющий вектор в линейном пространстве вторичных признаков объектов реального мира $\mathbf{v}_{\mathbb{I}_{\gamma,\mu},t}$

3. Дифференциальная кросс-валидация для верификации гиперпараметров модели

3.1. Квадратичное представление взвешенной задачи восстановления зависимостей

Пусть решение задачи восстановления зависимости итерационно найдено, и $\hat{\boldsymbol{\lambda}}_{\gamma,\mu} \in \mathbb{R}^N$ (49) — результат. В случае логистической регрессии, коэффициенты разложения Тейлора для неквадратичной обратной функции связи в точке решения будут определены (33)—(38) и пусть они обозначены как

$$\tilde{\mathbf{G}}_{\gamma,\mu} = \tilde{\mathbf{G}}^k = \tilde{\mathbf{G}}^{k+1}\ (N \times N),\ \tilde{\mathbf{y}}_{\gamma,\mu} = \tilde{\mathbf{y}}^k = \tilde{\mathbf{y}}^{k+1} \in \mathbb{R}^N \tag{52}$$

В случае регрессии, функция связи и обратная функция связи квадратичны, и тогда:

$$\tilde{\mathbf{G}}_{\gamma,\mu} = \mathbf{I}_N,\ \tilde{\mathbf{y}}_{\gamma,\mu} = \mathbf{y} \in \mathbb{R}^N \tag{53}$$

Очевидно, что $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}$ является фиксированной точкой итерационного алгоритма Ньютона на последнем шаге (43):

$$\tilde{\boldsymbol{\lambda}}^{(j,0)}_{\gamma,\mu} = \left(\left(\mathbf{X}^T_{\mathbb{I}_{\gamma,\mu}}\mathbf{X}_{\mathbb{I}_{\gamma,\mu}}\right) + \gamma\left(\hat{\mathbf{G}}_{\gamma,\mu}\right)^{-1}\right)^{-1}\tilde{\mathbf{y}}_{\gamma,\mu} \tag{54}$$

Принцип DiffLOO, изложенный в разделе II, основан на взвешенной формулировке задачи восстановления зависимости (13), в которой функции связи объектов обучающей выборки снабжаются некоторыми весами $r_t q(y_t, z_t)$. Веса изначально равны единице $r_t = 1$, но на каждом шаге процедуры LOO один из них слегка уменьшается $r_j = 1 - p$. Рассмотрим диагональную матрицу:

$$\begin{aligned}
&\mathbf{R}_j(p) = Diag(r_1 \dots r_N),\ r_t = 1, t \neq j,\ r_j = 1 - p,\\
&\mathbf{R}_j(p) = \mathbf{I}_N - Diag(0 \dots 0 \underbrace{p}_{j} 0 \dots 0),\ p \to 0.
\end{aligned} \tag{55}$$

Поскольку $\mathbf{R}_j(p)$ мало отличается от единичной матрицы $\mathbf{I}_N$, равенство (52) остается верным, если вставить $\mathbf{R}_j(p)$ между любыми двумя матрицами соответствующих размерностей. Обозначим через $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}$ вектор множителей Лагранжа (направляющий вектор обобщенной линейной модели зависимости в пространстве вторичных признаков), оцененный по взвешенной обучающей выборке с заглушенным слегка j-ым объектом:

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)} = \left(\mathbf{R}_j(p)\left(\mathbf{X}_{\mathbb{I}_{\gamma,\mu}}^T \mathbf{X}_{\mathbb{I}_{\gamma,\mu}}\right) + \gamma\left(\hat{\mathbf{G}}_{\gamma,\mu}\right)^{-1}\right)^{-1} \mathbf{R}_j(p)\tilde{\mathbf{y}}_{\gamma,\mu} \tag{56}$$

В этих терминах направляющий вектор невзвешенной модели с $p = 0$ будет обозначен как $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,0)}$. Таким образом, следует ожидать, что эффект заглушения j-го объекта в обучающей выборке проявится как небольшое увеличение соответствующей ошибки функции потерь (28)

$$q(y_j, \hat{z}_{\gamma,j}^{j,p}) = q(y_j, \mathbf{v}_{j,\mathbb{I}_{\gamma,\mu}}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}) \cong q(y_j, \mathbf{v}_{j,\mathbb{I}_{\gamma,\mu}}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,0)}) \tag{57}$$

Как было сказано в разделе II, в алгоритме вычисляются скорости роста $(\partial/\partial p)q(y_j, \hat{z}_{\gamma,j}^{j,p}) \geq 0$ для всех объектов из обучающей выборки в точке с весом $p = 0$:

$$\frac{\partial}{\partial p} q(y_j, \hat{z}_{\gamma,j}^{j,p})|_{p=0} = q'(y_j, \hat{z}_{\gamma,j}^{j,p})\left(\frac{\partial}{\partial p}\hat{z}_{\gamma,j}^{j,p}\right)_{p=0} \tag{58}$$

Здесь $\hat{z}_{\gamma,\mu,j}^{(j,p)} = \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}$. Чтобы найти эти скорости роста, мы должны сначала выразить вектор множителей Лагранжа, как N различных функций $\hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,p)}$ с темпом заглушения p для каждого обучающего объекта. Обратная матрица в этом равенстве является непрерывной функцией переменной p, поэтому для нее требуется найти аналитическое выражение.

3.2. Обращение возмущенной матрицы с помощью формулы Вудбери

Рассмотрим разницу между (54) и (56) через понятие вторичных признаков (18). Обе матрицы вторичных признаков имеют схожий вид:

$$\left(\mathbf{X}^T_{\mathbb{I}_{\gamma,\mu}}\mathbf{X}_{\mathbb{I}_{\gamma,\mu}}\right) = \mathbf{V}_{\mathbb{I}_{\gamma,\mu}},\ \mathbf{R}_j(p)\left(\mathbf{X}^T_{\mathbb{I}_{\gamma,\mu}}\mathbf{X}_{\mathbb{I}_{\gamma,\mu}}\right) = \mathbf{V}_{\mathbb{I}_{\gamma,\mu}} - p\mathbf{1}^t\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},t} \tag{59}$$

где $\mathbf{1}^t \in \mathbb{R}^N$ — строка где все элементы нули, кроме t-го.

Как видно частичное исключение j-го объекта математически представляет с собой вычитание из исходной матрицы $\left(\mathbf{X}^T_{\mathbb{I}_{\gamma,\mu}}\mathbf{X}_{\mathbb{I}_{\gamma,\mu}}\right)$ матрицу ранга один $p\mathbf{1}^j\mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}$. Это также верно для столбцов $\tilde{\mathbf{y}}_{\gamma,\mu}$ и $\mathbf{R}_j(p)\tilde{\mathbf{y}}_{\gamma,\mu}$:

$$\mathbf{R}_j(p)\tilde{\mathbf{y}}_{\gamma,\mu} = \tilde{\mathbf{y}}_{\gamma,\mu} - p\tilde{y}_j\mathbf{1}^j \tag{60}$$

С учетом (51), (52) и (50) получаем структуру для нового решения:

$$\hat{\lambda}^{(j,p)}_{\gamma,\mu} = \left[\left(\left(\mathbf{X}^T_{\mathbb{I}_{\gamma,\mu}}\mathbf{X}_{\mathbb{I}_{\gamma,\mu}}\right) + \gamma\left(\tilde{\mathbf{G}}_{\gamma,\mu}\right)^{-1}\right) - p\mathbf{1}^j\mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\right]^{-1}\left[\tilde{\mathbf{y}}_{\gamma,\mu} - p\tilde{y}_j\mathbf{1}^j\right] \tag{61}$$

Следует отметить, что поскольку предполагается, что начальная задача без пропущенных объектов будет решена (49), матрица в скобках уже обращена. Пусть такой результат будет обозначен как

$$\hat{\mathbf{D}}_{\gamma,\mu} = \left(\left(\mathbf{X}^T_{\mathbb{I}^{\mathbf{1}}_{\gamma,\mu}}\mathbf{X}_{\mathbb{I}^{\mathbf{1}}_{\gamma,\mu}}\right) + \gamma\left(\tilde{\mathbf{G}}^{\mathbf{1}}_{\gamma,\mu}\right)^{-1}\right)^{-1} \tag{62}$$

Последние аргументы наряду с рассуждениями в [6] лежат в основе следующей теоремы.

Теорема 6. *Решение задачи* (50) *выражается через следующее равенство:*

$$\hat{\boldsymbol{\lambda}}^{(j,p)}_{\gamma,\mu} = \hat{\boldsymbol{\lambda}}_{\gamma,\mu} - p\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}\frac{(y_j - \mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\boldsymbol{\lambda}}_{\gamma,\mu})}{1 - p\mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}} \tag{63}$$

$$\hat{\mathbf{z}}^{(j,p)}_{\gamma,\mu,j} = \hat{\mathbf{z}}_{\gamma,\mu,j} - \frac{p\mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}{1 - p\mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}(\tilde{y}_{\gamma,\mu,j} - \hat{z}_{\gamma,\mu,j}) \tag{64}$$

3.3. Критерий диференциального LOO

Минимум достижимого эмпирического риска в селективной задаче восстановления зависимостей из заданного набора данных (1) обеспечивается условием (8) как сумма ошибок функции потерь $\sum_{j=1}^{N} q(y_j, \hat{z}_{\gamma,\mu,j}),\ \hat{z}_{\gamma,\mu,j} = \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^{T}\hat{\boldsymbol{\lambda}}_{\gamma,\mu},\ \hat{\boldsymbol{\lambda}}_{\gamma,\mu} = \hat{\boldsymbol{\lambda}}_{\gamma,\mu}^{(j,0)}$.

Это неподходящий показатель качества для выбора гиперпараметров (γ, μ). Показатель качества обычного LOO это сумма ошибок функции потерь $\sum_{j=1}^{N} q\Big(y_j, \hat{z}_{\gamma,\mu,j}^{(j,1)}\Big)$ (54) со значением $p = 1$ для параметра частичного удаления, следовательно, полное исключение каждого следующего объекта $r_j = 1$, $p = 0$. Однако, в этом случае правило (63) быстрого решения задачи LOO является полностью правильным, только если параметр селективности признаков отключено $\mu = 0$, потому что в противном случае существует риск того, что удаление объекта приведет к изменению подмножества активных признаков $\hat{\mathbb{I}}_{\gamma,\mu}^{(j,p)} = \hat{\mathbb{I}}_{\gamma,\mu}$. Конечно, способ многократного численного решения взвешенной задачи (23) остается всегда верным, но он чрезвычайно трудоемкий. Следовательно, чтобы выразить идею оценки скоростей роста ошибок на объектах из обучающей выборки как применимый критерий DiffLOO, необходимо найти формулу для производных.

$$\text{DifLOO}(\gamma, \mu) = \frac{1}{N}\sum_{j=1}^{N}\Big[\frac{\partial}{\partial p} q(y_j, \hat{z}_{\gamma,\mu,j}^{(j,p)})\Big]_{p=0}; \tag{65}$$

Теорема 7. *Критерий дифференциального LOO полностью определяется решением задачи оценки начальной зависимости*

$$\mathit{DifLOO}(\gamma, \mu) = \frac{1}{N}\sum_{j=1}^{N}\Big(y_j - \hat{z}_{\gamma,\mu,j}\Big)^2\big(\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^{T}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}\big) \tag{66}$$

С функциями связи для линейной регрессии и логистической регрессии, соответственно:

$$\begin{cases} \tilde{y}_{\gamma,\mu,j} = y_j \\ \tilde{y}_{\gamma,\mu,j} = \hat{z}_{\gamma,\mu,j} + y_j(1 + \exp(-y_j\hat{z}_{\gamma,\mu,j})) \end{cases} \tag{67}$$

4. Оптимизация гиперпараметров через критерий DiffLOO

В рамках постановки задачи, рассматриваемой в данном исследовании, вывод модели восстановления зависимости из заданной обучающей выборки понимается как минимизация регуляризованного эмпирического риска с помощью селективной Ридж-регуляризации. Соответствующая целевая функция, которую нужно минимизировать (9), содержит два вещественных структурных гиперпараметра γ и μ.

Первый из них $0 \leq \gamma < \infty$ — это обычный коэффициент Ридж-регуляризации, который служит как решение, чтобы избежать континуум моделей, когда матрица может быть вырожденной, как в (54) и (56). Часто достаточно принять достаточно малое значение $\gamma \approx 0$.

Что касается гиперпараметра селективности $0 \leq \mu < \infty$, то это основной гиперпараметр задачи восстановления зависимости в прямой (9) или двойственной записи (13), который существенно влияет на результат обучения. Если $\mu = 0$, критерий вообще не обладает свойством селективности, и все оцененные компоненты вектора направления остаются активными (25). Наоборот, когда гиперпараметр селективности достаточно велик $\mu \to \infty$, все компоненты направляющего вектора обнуляются. Максимальное значение селективности, которое полностью подавляет все признаки будем обозначать как $\mu_{\max}$.

Предположим, что подмножество активных признаков пустое в (25) $\mathbb{I} = \varnothing$:

$$\hat{\boldsymbol{\lambda}}^* = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^N} \mathbf{W}_0(\boldsymbol{\lambda}|\gamma, \mu) = \sum_{t=1}^{N} \varphi_t(y_t, \lambda_t|\gamma) \tag{68}$$

Это выпуклая функция, поскольку каждая из обратных функций связи $\varphi_t(y_t, \lambda_t|\gamma)$ (24) является выпуклой.

В случае линейной регрессии все они квадратичны (24), квадратичная

функция (68) имеет замкнутое решение $\hat{\boldsymbol{\lambda}}^* = (1/\gamma)\mathbf{y}$

$$\mathbf{W}^*(\boldsymbol{\lambda}|\gamma,\mu) = \left(\frac{1}{2}\gamma\boldsymbol{\lambda}^T\boldsymbol{\lambda} - \mathbf{y}^T\boldsymbol{\lambda}\right) \to \min(\boldsymbol{\lambda} \in \mathbb{R}^N) \tag{69}$$

В случае логистической регрессии обратные функции связи неквадратичные и решение (68) должно быть найдено с помощью того же алгоритма Ньютона, что и основная двойственная задача, с существенным следствием того, что подмножество активных признаков остается пустым $\mathbb{I} = \varnothing$

$$\mathbf{W}^*(\boldsymbol{\lambda}|\gamma,\mu) = \sum_{t=1}^{N} \frac{1}{2\gamma}\Big[\big(2\gamma y_t\lambda_t\big)\ln\big(2\gamma y_t\lambda_t\big) + \big(1-2\gamma y_t\lambda_t\big)\ln\big(1-2\gamma y_t\lambda_t\big)\Big] \to$$

$$\to \min(\boldsymbol{\lambda} \in \mathbb{R}^N) \tag{70}$$

Пусть $\boldsymbol{\lambda}^*$ – решение задачи (68), и $\mu_{\max}$ определена как

$$\mu_{\max} = \max_{i=1,\ldots,n} \big(\boldsymbol{\lambda}^*\mathbf{x}_i\big) \tag{71}$$

Тогда, если $\mu = \mu_{\max}$, то матрица $V_{\mathbb{I}} = \mathbf{0}$ $(N \times N)$ в (26), и направляющий вектор $\hat{\mathbf{a}} = \mathbf{0}$ в (25). Таким образом, адаптивный диапазон гиперпараметра селективности в (68) $0 \le \mu \le \mu_{\max}$.

Число активных признаков будет расти от 0 до n, пока μ уменьшается от $\mu_{\max}$ до 0. Это просто точное представление о полном пути регуляризации [8, 9]. Теоретически, число точек верификации, где меняется число активных признаков, будет не меньше n, но в действительности оно будет намного больше n, поскольку этот процесс далек от монотонности. В результате, такая процедура была бы слишком трудоемкой в случае большого количества признаков.

Мы ограничимся поиском по сетке в пределах конечного множества «разумных» значений

$$\mu \in \{0 < \mu_1 < \cdots < \mu_m = \mu_{\max}\} \tag{72}$$

Опыт показывает, что целесообразно делить интервал $[10^{-8}\mu_{\max} \approx 0,\ \mu_{\max}]$ в логарифмическом масштабе на $m \leq n$ подинтервалов.

$$\mu_l = 10^{-8(l/m)}\mu_{\max}, l = 0, 1, \ldots, m; \quad \mu_0 = 10^{-0}\mu_{\max} = \mu_{\max},$$
$$\mu_m = 10^{-8}\mu_{\max} \approx 0 \quad (73)$$

Грубая регуляризация начинается с $l = 0$, что соответствует $\mu = \mu_{\max}$ и тривиальная двойственная задача (63), которая приводит $\hat{\mathbf{a}}$ к нулевому вектору (27).

Тем не менее, результат итерационного процесса $\hat{\boldsymbol{\lambda}}_{\mu_0}$ должен быть сохранен. Каждое следующее значение параметра селективности $\mu = \mu_l$ будет практически совпадать с предыдущим значением $\mu = \mu_{l-1}$, и итерационный процесс, начатый с предыдущим решением $\hat{\boldsymbol{\lambda}}_{\mu_{l-1}}$, будет сходиться только после нескольких итераций. В большинстве случаев, после одной или двух итераций. Число ненулевых компонент направляющего вектора будет постепенно расти (25). Наконец, на последнем шаге $\mu = \mu_m \approx 0$, направляющий вектор будет состоять почти из всех компонентов.

5. Экспериментальное исследование метода оценки дифференциальной кросс-валидации

Целью экспериментального исследования заключается в том, чтобы продемонстрировать правильность работы методологии выбора признаков в сочетании с селективной оптимизацией через метод DiffLOO для восстановления зависимостей крайне разреженных моделей. Необходимость таких моделей вытекает из прикладных задач, естественно удовлетворяемых предположению о том, что коэффициенты признаков в обобщенной линейной модели отличаются от нуля только в пределах реально существующего небольшого подмножества из большого множества доступных признаков. Поиск этого подмножества (Factor Search) является основной целью обработки данных.

Рассмотрим две прикладные задачи — линейную и логистическую регрессии, в которых заранее известно оптимальное подмножество признаков. Задача заключается в том, чтобы по заданным выборкам восстановить это оптимальное подмножество при оптимальных значений гиперпараметров $\hat{\gamma}$ и $\hat{\mu}$. Целью эксперимента в задаче линейной регрессии является нахождение подмножества активных ценных бумаг $\hat{n} = 13$ из набора регрессоров $n = 650$, а в задаче логистической регрессии — нахождение заданного оптимального подмножества признаков $\hat{n} = 2$ из заданных признаков $n = 400$.

5.1. Линейная регрессия

В качестве яркого примера мы рассмотрим практическую задачу восстановления скрытого состава инвестиционного портфеля, представленного временным рядом его периодических доходностей (значений относительной доходности). А в качестве набора регрессоров применяются $n = 650$ временных рядов месячных доходностей биржевых ценных бумаг на Нью-Йорской

фондовой бирже, каждый длиной чуть больше 20 лет.

$$\mathbf{X}_t = (x_{t,i=1,\ldots,N}) \in \mathbb{R}^n,\ n = 650 \tag{74}$$

Наблюдаемый сигнал, состоящий из $N = 251$.

$$y_t,\ t = 1, \ldots, N = 251 \tag{75}$$

Временной ряд доходностей инвестиционного портфеля, построенного как вложение капитала в равных долях в $n^* = 13$ неизвестных ценных бумаг во множестве $\{1, \ldots, n\}$. В качестве модели используется регрессионая модель Шарпа

$$y_t \cong \sum_{i=1}^{n} a_i x_{t,i} \tag{76}$$

где a_i - это доли вложения капитала.

Кроме того, рассматривается дополнительная задача Factor Search, заключающаяся в оценивании фактического состава инвестиционного портфеля в большом множестве биржевых активов.

Результат верификации селективной модели по критерию DiffLOO показан на рис. 1 в зависимости от μ. Для каждого значения μ итерационный алгоритм зафиксировал подмножества активных признаков, на которых проводится метод DiffLOO (23). Результаты показывают, что оптимальное подмножество признаков $n* = 13$ содержится в подмножестве активных признаков $\hat{n} = 16$ при оптималном значении $\hat{\mu} = \exp(-4)$. Элементы данных подмножеств показаны в табл. 1.

Таблица 1 – Оптимальные помножества регрессоров

μ	92, 155, 164, 184, 243, 263, 288, 332, 347, 399, 412, 421, 430, 507, 522, 626
$\hat{\mu}$	92, 164, 184, 243, 263, 288, 332, 347, 412, 421, 507, 522, 626

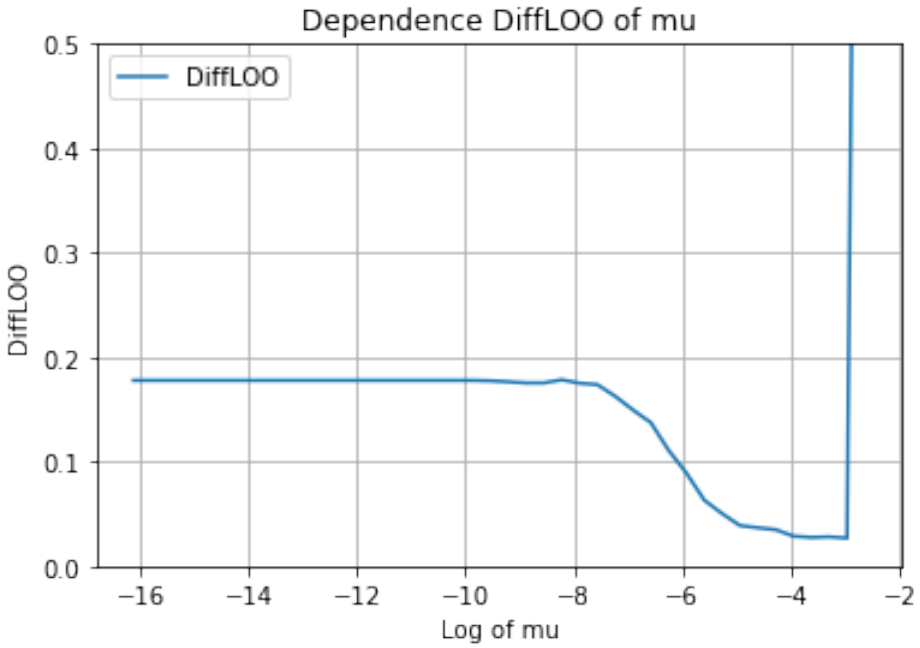

Рисунок 1 – Зависимость дифференциального LOO от параметра μ

5.2. Логистическая регрессия

В качестве выборки вязта искуственно сгенерированная выборка, состоящая из 400 признаков и 200 объектов, в которой только 2 признака образуют оптимальное подмножество признаков, а остальные признаки — шумовые.

$$\mathbf{X}_t = (x_{t,i=1,\ldots,N}) \in \mathbb{R}^n,\ n = 400 \tag{77}$$

Целевая переменная, состоящая из $N = 200$.

$$y_t \in \ \{+1, -1\},\ t = 1, \ldots, N = 200 \tag{78}$$

Оптимальные признаки заданы из нормального распределения с $\mathbb{E} = 0$ и $\mathbb{D} = 1$. Остальные признаки шумовые. Направляющий вектор задается единичным, и с его помощью получена целевая переменная для каждого объекта.

Результат верификации селективной модели для задачи логистической регрессии по критерию DiffLOO показан на рис. 2 в зависимости от μ. Результаты показывают, что оптимальное подмножество признаков $N = 2$

в точности совпадает с $\hat{N} = 2$ при оптималном значении $\hat{\mu} = \exp(-3)$.

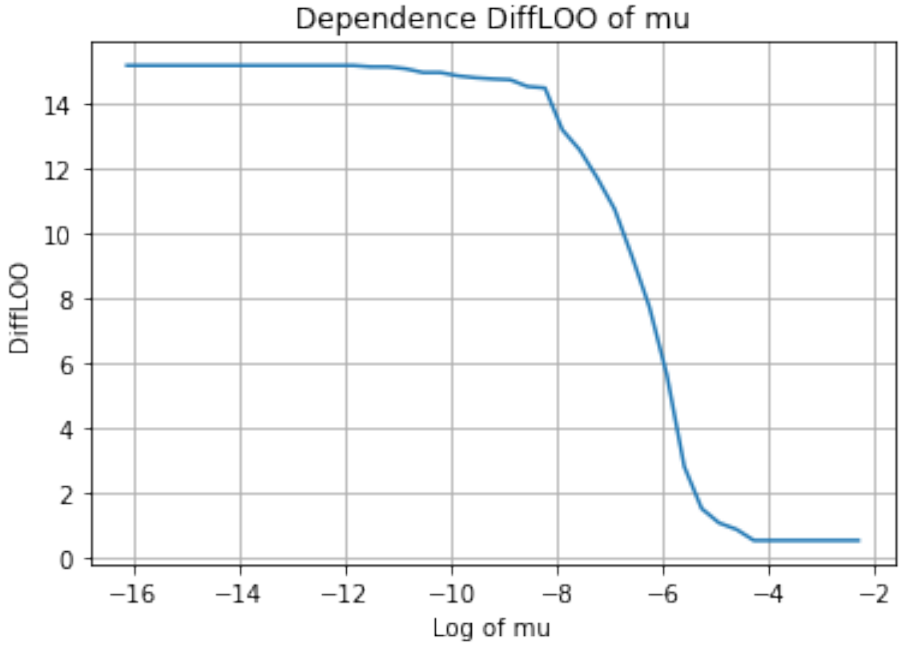

Рисунок 2 – Зависимость дифференциального LOO от параметра μ

6. Пропущенные доказательства

6.1. Доказательство леммы 1

Так как функция регуляризация одна и та же, то достаточно показать:

$$\sum_{j=1}^{N} r_j q\Big(y_j, \sum_{i=1}^{n} a_i x_{j,i}\Big) \approx \sum_{j=1}^{N} q\Big(y_j, \sum_{i=1}^{n} r_j a_i x_{j,i}\Big) \tag{79}$$

Для линейной регрессии в матричном виде проверим следующее соотношение:

$$\begin{gathered}\Big(\mathbf{Y} - \mathbf{X}^T\mathbf{a}\Big)^T \mathbf{R}\Big(\mathbf{Y} - \mathbf{X}^T\mathbf{a}\Big) \approx \Big(\mathbf{Y} - \mathbf{R}\mathbf{X}^T\mathbf{a}\Big)^T\Big(\mathbf{Y} - \mathbf{R}\mathbf{X}^T\mathbf{a}\Big) \\ \Big(\mathbf{Y} - \mathbf{Z}\Big)^T \mathbf{R}\Big(\mathbf{Y} - \mathbf{Z}\Big) \approx \Big(\mathbf{Y} - \mathbf{R}\mathbf{Z}\Big)^T\Big(\mathbf{Y} - \mathbf{R}\mathbf{Z}\Big) \\ \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{R}\mathbf{Z} + \mathbf{Z}^T\mathbf{R}\mathbf{Z} \approx \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{R}\mathbf{Z} + \mathbf{Z}^T\mathbf{R}\mathbf{R}\mathbf{Z} \\ \mathbf{Y}^T\mathbf{R}\mathbf{Y} + \mathbf{Z}^T\mathbf{R}\mathbf{Z} \approx \mathbf{Y}^T\mathbf{Y} + \mathbf{Z}^T\mathbf{R}\mathbf{R}\mathbf{Z}\end{gathered} \tag{80}$$

Представим R как разность $E - 0_j^1$, где E - единичная матрица, а 0_j^1 это диагональная матрица, у которой на диагонали все элементы нулевые, кроме j-го, который равен $p \to 0$

$$\begin{gathered}\mathbf{Y}^T\mathbf{Y} - p\mathbf{Y}^T\mathbf{0}_j^1\mathbf{Y} + \mathbf{Z}^T\mathbf{Z} - p\mathbf{Z}^T\mathbf{0}_j^1\mathbf{Z} \approx \mathbf{Y}^T\mathbf{Y} + \mathbf{Z}^T\mathbf{Z} + (p^2 - 2p)\mathbf{Z}^T\mathbf{0}_j^1\mathbf{Z} \\ -py_j^2 - pz_j^2 \approx (p^2 - 2p)z_j^2\end{gathered} \tag{81}$$

При p стремляющимся к нулю ($p \to 0$), выражения в (81) асимптотически равны.

6.2. Доказательство леммы 2

Рассмотрим взвешенную задачу (13)

$$\begin{cases}\gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \le \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum_{t=1}^{N} r_t q(y_t, z_t) \to min(\mathbf{a} \in \mathbb{R}^n, \mathbf{z}) \\ z_t = \mathbf{a}^T\mathbf{x_t}, t = 1, \ldots, N \end{cases} \tag{82}$$

Соответствующая функция Лагранжа для взвешенной задачи (82):

$$L(\mathbf{a}, z_1, \dots, z_N, \lambda_1, \dots, \lambda_N) = \frac{1}{2}\sum_{i=1}^{n}\begin{pmatrix} 2\mu|a_i|, |a_i| \le \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \\ + \frac{1}{2\gamma}\sum_{t=1}^{N} r_t q(y_t, z_t) - \sum_{j=t}^{N}\lambda_t\Big(\sum_{i=i}^{n} a_i x_{t,i} - z_t\Big) = \\ = \frac{1}{2}\sum_{i=1}^{n}\begin{pmatrix} 2\mu|a_i| - 2\Big(\sum_{t=1}^{N}\lambda_t x_{t,i}\Big)a_i, |a_i| \le \mu \\ \mu^2 + {a_i}^2 - 2\Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)a_i, |a_i| > \mu \end{pmatrix} + \\ + \sum_{t=1}^{N}\Big(\frac{1}{2\gamma} r_t q(y_t, z_t) + \lambda_t z_t\Big) \to \begin{cases} min(\mathbf{a}, z_1, \dots, z_N) \\ \partial/\partial\lambda_t, t = 1, \dots, N. \end{cases} \tag{83}$$

Обозначим через L_i первый член функции (83)

$$L_i(a_i, \lambda_1, \dots, \lambda_N) = \frac{1}{2}\sum_{i=1}^{n}\begin{pmatrix} 2\mu|a_i| - 2\Big(\sum_{t=1}^{N}\lambda_t x_{t,i}\Big)a_i, |a_i| \le \mu \\ \mu^2 + {a_i}^2 - 2\Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)a_i, |a_i| > \mu \end{pmatrix} \tag{84}$$

Тогда представим функцию Лагранжа в следующем виде:

$$L(\mathbf{a}, \mathbf{z}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} L_i(a_i, \boldsymbol{\lambda}) + \sum_{t=1}^{N}\Big(\frac{1}{2\gamma} r_t q(y_t, z_t) + \lambda_t z_t\Big) \to \\ \to \begin{cases} min(a_1, \dots, a_n, z_1, \dots, z_N) \\ \partial/\partial\lambda_t, t = 1, \dots, N. \end{cases} \tag{85}$$

Минимизация по направляющему вектору $\mathbf{a}$ эквивалента выражению $\arg\min_{a_i} L_i(a_i, \boldsymbol{\lambda})$. Значит,

$$\hat{a}_i(\boldsymbol{\lambda}) = \arg\min_{a_i}\begin{pmatrix} 2\mu|a_i| - 2\Big(\sum_{t=1}^{N}\lambda_t x_{t,i}\Big)a_i, |a_i| \le \mu \\ \mu^2 + {a_i}^2 - 2\Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)a_i, |a_i| > \mu \end{pmatrix} \tag{86}$$

Для удобства представим функцию ${a_i}^2 - 2\Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)a_i$ из (84) в

виде полного квадрата

$$a_i{}^2 - 2\Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)a_i = \Big(a_i - \sum_{t=i}^{N}\lambda_t x_{t,i}\Big)^2 - \Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)^2 \qquad (87)$$

Следовательно получаем:

$$L_i(a_i,\boldsymbol{\lambda}) = \frac{1}{2}\begin{pmatrix} 2\mu|a_i| - 2\Big(\sum_{t=1}^{N}\lambda_t x_{t,i}\Big)a_i & ,|a_i|\le\mu \\ \mu^2 + \Big(a_i - \sum_{t=i}^{N}\lambda_t x_{t,i}\Big)^2 - \Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)^2 & ,|a_i|>\mu \end{pmatrix} \to$$

$$\to \min(a_i) \qquad (88)$$

Значит,

$$\hat{a}_i(\boldsymbol{\lambda}) = \begin{pmatrix} 0, & |\sum_{t=i}^{N}\lambda_t x_{t,i}| \le \mu \\ \sum_{t=i}^{N}\lambda_t x_{t,i}, & |\sum_{t=i}^{N}\lambda_t x_{t,i}| > \mu \end{pmatrix} \qquad (89)$$

С учетом (87) и (89) получаем:

$$\hat{L}_i(\lambda) = \min_{a_i} L_i(a_i,\lambda) = \frac{1}{2}\begin{pmatrix} 0 & ,|\sum_{t=i}^{N}\lambda_t x_{t,i}| \le \mu \\ \mu^2 - \big(\sum_{t=i}^{N}\lambda_t x_{t,i}\big)^2 & ,|\sum_{t=i}^{N}\lambda_t x_{t,i}| > \mu \end{pmatrix} =$$

$$= \frac{1}{2}\min\Big[0, \mu^2 - \big(\sum_{t=i}^{N}\lambda_t x_{t,i}\big)^2\Big] \qquad (90)$$

Подстановка в функцию Лагранжа

$$\min_{\mathbf{a}} L(\mathbf{a},\mathbf{z},\lambda) = \sum_{t=1}^{n}\frac{1}{2}\min\Big[0,\mu^2 - \big(\sum_{t=i}^{N}\lambda_t x_{t,i}\big)^2\Big] + \sum_{t=1}^{N}\Big(\frac{1}{2\gamma}r_t q(y_t,z_t) + \lambda_t z_t\Big) \to$$

$$\to \begin{cases} min(\mathbf{a}, z_1, \ldots, z_N) \\ \partial/\partial\lambda_t, t = 1,\ldots,N. \end{cases} \qquad (91)$$

Для удобства, используем обозначение из (23) для подмножества активных признаков $\mathbb{I}(\lambda|\mu) = \{i : |\mathbf{x}_i\lambda| > \mu\} \subseteq 1, \ldots, n$

Рассмотрим двойственную задачу

$$-\frac{1}{2}\sum_{i\in\mathbb{I}(\lambda|\mu)}\Big(\sum_{t=i}^{N}\lambda_t x_{t,i}\Big)^2 + \sum_{t=1}^{N}\Big(\frac{1}{2\gamma}r_t q(y_t, z_t) + \lambda_t z_t\Big) \to$$

$$\to \begin{cases} min(z_1, \ldots, z_N) \\ \partial/\partial\lambda_t, t = 1, \ldots, N. \end{cases} \tag{92}$$

Двойственная задача в матричном виде

$$-\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda} + \frac{1}{2\gamma}\sum_{t=1}^{N} r_t q(y_t, z_t) + \mathbf{z}^T\boldsymbol{\lambda} \to \begin{cases} min(\mathbf{z}) \\ \nabla_{\boldsymbol{\lambda}} = 0 \end{cases} \tag{93}$$

6.3. Доказательство леммы 3

Необходимо представить эквивалентную запись для следующей формулы из (49) с учетом $\tilde{\mathbf{G}}^{-1}_{\gamma,\mu} = \mathbf{I}_{N\times N}$:

$$\hat{\boldsymbol{\lambda}}_{\mathbb{I}(\lambda|\mu)} = \left(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big) + \gamma I\right)^{-1}\mathbf{y} \tag{94}$$

Формула Вудбери:

$$\big(A + BC\big)^{-1} = A^{-1} - A^{-1}B\big(I + CA^{-1}B\big)^{-1}CA^{-1} \tag{95}$$

Для получения эквивалетной формулы (49) для решения вектора множителей Лагранжа применим формулу Вудбери, представленную в (95)

$$\left(\underbrace{\gamma I_{N\times N}}_{A}+\Big(\underbrace{\mathbf{X}_{\mathbb{I}^k}^T}_{B}\underbrace{\mathbf{X}_{\mathbb{I}^k}}_{C}\Big)\right)^{-1}=$$

$$=\frac{1}{\gamma}I_{N\times N}-\frac{1}{\gamma}I_{N\times N}\mathbf{X}_{\mathbb{I}^k}^T\left(I_{n\times n}+\mathbf{X}_{\mathbb{I}^k}\Big(\frac{1}{\gamma}I_{N\times N}\Big)\mathbf{X}_{\mathbb{I}^k}^T\right)^{-1}\mathbf{X}_{\mathbb{I}^k}\Big(\frac{1}{\gamma}I_{N\times N}\Big)=$$

$$=\frac{1}{\gamma}I_{N\times N}-\frac{1}{\gamma^2}\mathbf{X}_{\mathbb{I}^k}^T\left(I_{n\times n}+\frac{1}{\gamma}\mathbf{X}_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}^T\right)^{-1}\mathbf{X}_{\mathbb{I}^k}=$$

$$=\frac{1}{\gamma}I_{N\times N}-\frac{1}{\gamma}\mathbf{X}_{\mathbb{I}^k}^T\left(\gamma I_{n\times n}+\mathbf{X}_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}^T\right)^{-1}\mathbf{X}_{\mathbb{I}^k}=$$

$$=\frac{1}{\gamma}\left[I_{N\times N}-\mathbf{X}_{\mathbb{I}^k}^T\left(\gamma I_{n\times n}+\mathbf{X}_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}^T\right)^{-1}\mathbf{X}_{\mathbb{I}^k}\right] \quad (96)$$

Тогда подставляем в (94) и получаем эквивалентную запись для случая когда число признаков меньше числа объектов $n < N$.

$$\hat{\lambda}=\frac{1}{\gamma}\left[I_{N\times N}-\mathbf{X}_{\mathbb{I}^k}^T\left(\gamma I_{n\times n}+\mathbf{X}_{\mathbb{I}^k}\mathbf{X}_{\mathbb{I}^k}^T\right)^{-1}\mathbf{X}_{\mathbb{I}^k}\right]\mathbf{y} \quad (97)$$

6.4. Доказательство теоремы 5

Задача поиска седловой точки в (82) состоит из задач оптимизации (98) и (103).

$$\frac{1}{2\gamma}\sum_{t=1}^{N}r_t q(y_t,z_t)+\mathbf{z}^T\boldsymbol{\lambda}\to\min(\mathbf{z}) \quad (98)$$

6.4.1. Линейная регрессия

$$\sum_{t=1}^{N}r_t q(y_t,z_t)=\sum_{t=1}^{N}r_t(z_t-y_t)^2=(\mathbf{z}-\mathbf{y})^T\mathbf{R}(\mathbf{z}-\mathbf{y}) \quad (99)$$

В (93) заменим $q(y_t, z_t)$ на ее соответствующую функцию связи

$$\hat{\mathbf{z}}^r=\arg\min_z\left(\frac{1}{2\gamma}(\mathbf{z}-\mathbf{y})^T\mathbf{R}(\mathbf{z}-\mathbf{y})+\mathbf{z}^T\boldsymbol{\lambda}\right) \quad (100)$$

Приравняем к нулю градиент по $\mathbf{z}$

$$\frac{1}{\gamma}\mathbf{R}(\mathbf{z}-\mathbf{y})+\boldsymbol{\lambda}=0,\ \ \frac{1}{\gamma}R(\mathbf{z}-\mathbf{y})+\boldsymbol{\lambda}=0,\ \ R(\mathbf{z}-\mathbf{y})=-\gamma\boldsymbol{\lambda}, \tag{101}$$
$$R\mathbf{z}=R\mathbf{y}-\gamma\boldsymbol{\lambda},\ \ \mathbf{z}=R^{-1}(R\mathbf{y}-\gamma\boldsymbol{\lambda}),\ \ \mathbf{z}=\mathbf{y}-\gamma R^{-1}\boldsymbol{\lambda}$$

Подставим $\mathbf{z}$ в функции Лагранжа (93)

$$\begin{aligned}
&-\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\frac{1}{2\gamma}(\mathbf{z}-\mathbf{y})^T\mathbf{R}(\mathbf{z}-\mathbf{y})+\mathbf{z}^T\boldsymbol{\lambda}=\\
&=-\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\frac{1}{2\gamma}(\mathbf{y}-\gamma R^{-1}\boldsymbol{\lambda}-\mathbf{y})^T\mathbf{R}(\mathbf{y}-\gamma R^{-1}\boldsymbol{\lambda}-\mathbf{y})+\mathbf{z}^T\boldsymbol{\lambda}=\\
&=-\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\frac{1}{2\gamma}(-\gamma R^{-1}\boldsymbol{\lambda})^T\mathbf{R}(-\gamma R^{-1}\boldsymbol{\lambda})+(\mathbf{y}-\gamma R^{-1}\boldsymbol{\lambda})^T\boldsymbol{\lambda}=\\
&=-\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\frac{1}{2}\gamma R^{-1}\boldsymbol{\lambda}^T\boldsymbol{\lambda}+\mathbf{y}^T\boldsymbol{\lambda}-\gamma R^{-1}\boldsymbol{\lambda}^T\boldsymbol{\lambda}=\\
&=-\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}-\frac{1}{2}\gamma R^{-1}\boldsymbol{\lambda}^T\boldsymbol{\lambda}+y^T\boldsymbol{\lambda}
\end{aligned} \tag{102}$$

Получаем функцию W, которую необходимо минимизировать по $(\boldsymbol{\lambda})$

$$W(\mathbf{z}|\gamma,\mu)=\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\frac{1}{2}\gamma R^{-1}\boldsymbol{\lambda}^T\boldsymbol{\lambda}-\mathbf{y}^T\boldsymbol{\lambda}\to\min(\boldsymbol{\lambda}) \tag{103}$$

Приравниваем к нулю градиент в (103)

$$\begin{aligned}
\nabla_{\boldsymbol{\lambda}}&\left(\frac{1}{2}\boldsymbol{\lambda}\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\frac{1}{2}\gamma R^{-1}\boldsymbol{\lambda}^T\boldsymbol{\lambda}-\mathbf{y}^T\boldsymbol{\lambda}\right)\\
&\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)\boldsymbol{\lambda}+\gamma R^{-1}\boldsymbol{\lambda}-\mathbf{y}=0\\
&\left(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma R^{-1}\right)\boldsymbol{\lambda}=\mathbf{y}\\
&\boldsymbol{\lambda}=\left(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma R^{-1}\right)^{-1}\mathbf{y}\\
&\boldsymbol{\lambda}=\left(R\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma E\right)^{-1}R\mathbf{y}
\end{aligned} \tag{104}$$

6.4.2. Логистическая регрессия

Из теоремы 3 известно приближенное квадратичное представление логистической функции потерь $\ln(1+\exp(-y_t z_t))$, формула которого дана

в формуле (32)

$$\tilde{q}(y_t, z_t) \cong \tilde{g}_t(z_t - \tilde{y}_t)^2 \tag{105}$$

$$\hat{\mathbf{z}}^r = \arg\min_z \left(\frac{1}{\gamma}(\mathbf{z} - \tilde{\mathbf{y}})^T \mathbf{R}\tilde{\mathbf{G}}(\mathbf{z} - \tilde{\mathbf{y}}) + \mathbf{z}^T \left(\mathbf{X}^T\mathbf{X} \right)^{-1} \mathbf{z} \right) \tag{106}$$

Задача линейной регрессии и логистической регрессии отличаются только, тем что, в логистической регрессии появляется новая целевая переменная $\tilde{\mathbf{y}}$ и каждый штраф взвешивается с коэффициентов $\tilde{g}_t$. Поэтому повторяя те же шаги из расчетов для линейной регрессии получаем:

$$\hat{\boldsymbol{\lambda}}^r = \left(\mathbf{R}\left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)} \mathbf{X}_{\mathbb{I}(\lambda|\mu)} \right) + \gamma \tilde{\mathbf{G}}^{-1} \right)^{-1} \mathbf{R}\tilde{\mathbf{y}} \tag{107}$$

6.5. Доказательство теоремы 6

Из формул (93) и (107), если в качестве матрицы $\mathbf{R}$ взять единичную матрицу

$$\hat{\mathbf{z}} = \left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)} \mathbf{X}_{\mathbb{I}(\lambda|\mu)} \right) \left(\left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)} \mathbf{X}_{\mathbb{I}(\lambda|\mu)} \right) + \gamma E \right)^{-1} \mathbf{y} \tag{108}$$

Для упрощения формулы (108) примем следующее обозначение

$$D_{\mathbb{I}(\lambda|\mu)} = \left(\left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)} \mathbf{X}_{\mathbb{I}(\lambda|\mu)} \right) + \gamma E \right)^{-1} \tag{109}$$

Теперь в наших терминах для удаления одного объекта с весом p матрицу $\mathbf{R}$ заменяем на единичную, но на месте j-ой позиции будет элемент $1 - p$, таким образом, получаем:

$$\hat{\mathbf{z}}^{1^p_j} = \left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)} \mathbf{X}_{\mathbb{I}(\lambda|\mu)} \right) \left[\left(\left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)} \mathbf{X}_{\mathbb{I}(\lambda|\mu)} \right) + \gamma E \right) - p\mathbf{1}^j \mathbf{v}_j^T \right]^{-1} \times \\ \times \left[\mathbf{y} - p y_j \mathbf{1}^j \right] \tag{110}$$

В формуле (110) для обращения матрицы примем следующие обозначения для применения формулы Вудбери

$$\Big[\underbrace{\Big(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma E\Big)}_{A}+\underbrace{(-p\mathbf{1}^j)}_{B}\underbrace{\mathbf{v}_j^T}_{C}\Big]^{-1} \tag{111}$$

Тогда применяем формулу Вудбери (110)

$$\Big(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma E\Big)^{-1}-\Big(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma E\Big)^{-1}\big(-p\mathbf{1}^j\big)\times$$
$$\times\Big[1+\big(\mathbf{v}_j^T\big)\Big(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma E\Big)^{-1}\big(-p\mathbf{1}^j\big)\Big]^{-1}\big(\mathbf{v}_j^T\big)\Big(\Big(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\Big)+\gamma E\Big)^{-1} \tag{112}$$

Обозначим обратную матрицу A^{-1} как D. С учетом принятого обозначения для вторичных признаков (18), в (109) производим замену и получаем следущее равенство

$$\begin{aligned}\hat{\boldsymbol{\lambda}}^{1^p_j}_{\gamma,\mu}&=\Big[D-D\big(-p\mathbf{1}^j\big)\Big(1+\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D\big(-p\mathbf{1}^j\big)\Big)^{-1}\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D\Big]\big[\mathbf{y}-py_j\mathbf{1}^j\big]=\\&=\left(D+\frac{pD_j\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D}{1-p\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D_j}\right)\big[\mathbf{y}-py_j\mathbf{1}^j\big]=D\mathbf{y}+\frac{pD_j\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D\mathbf{y}}{1-p\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D_j}-\\&\quad-\frac{py_jD_j}{1-p\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D_j}=\hat{\boldsymbol{\lambda}}_{\gamma,\mu}-\frac{pD_j(y_j-\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\lambda}_{\gamma,\mu})}{1-p\mathbf{V}^T_{\mathbb{I}(\lambda|\mu),j}D_j}\end{aligned} \tag{113}$$

В конечном итоге получаем:

$$\hat{\lambda}^{\mathbf{1}^p_j}_{\gamma,\mu}=\hat{\lambda}_{\gamma,\mu}-\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}\frac{p(y_j-\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\lambda}_{\gamma,\mu})}{1-p\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}} \tag{114}$$

6.6. Доказательство теоремы 7

Из (114) и равенства $z_j=\mathbf{v}_j^T\boldsymbol{\lambda}$ получаем следующее соотношение.

$$\begin{aligned}\hat{\mathbf{z}}^{\mathbf{1}^p_j}_{\gamma,\mu}&=\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\lambda}^{\mathbf{1}^p_j}_{\gamma,\mu}=\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\lambda}_{\gamma,\mu}-\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}\frac{p(y_j-\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\lambda}_{\gamma,\mu})}{1-p\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}=\\&=\hat{\mathbf{z}}_{\gamma,\mu}-\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}\frac{p(y_j-\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\hat{\lambda}_{\gamma,\mu})}{1-p\mathbf{V}^T_{\mathbb{I}_{\gamma,\mu},j}\mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}\end{aligned} \tag{115}$$

6.6.1. Линейная регрессия

Для линейной регрессии и с учетом формулы (115)

$$q(y_j, \hat{\mathbf{z}}_{\gamma,\mu}^{\mathbf{1}_j^p}) = (y_j - \hat{\mathbf{z}}_{\gamma,\mu}^{\mathbf{1}_j^p})^2 \tag{116}$$

С учетом (114) и (115) получаем

$$\begin{aligned} y_j - \hat{\mathbf{z}}_{\gamma,\mu}^{\mathbf{1}_j^p} = y_j - \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu} + \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j} \frac{p(y_j - \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \hat{\lambda}_{\gamma,\mu})}{1 - p\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}} = \\ = \left(y_j - \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right)\left(1 + \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j} \frac{p}{1 - p\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}\right) = \\ = \frac{y_j - \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}}{1 - p\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}} \end{aligned} \tag{117}$$

Тогда функцию потерь примет вид:

$$q(y_j, \hat{\mathbf{z}}_{\gamma,\mu}^{\mathbf{1}_j^p}) = \left(\frac{y_j - \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}}{1 - p\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}\right)^2 \tag{118}$$

DiffLOO в свою очередь примет следующий вид

$$\text{DiffLOO}(\gamma, \mu) = \frac{1}{N} \sum_{j=1}^{N} \frac{\partial}{\partial p} q(y_j, \hat{\mathbf{z}}_{\gamma,\mu,j}^{\mathbf{1}_j^p}) \tag{119}$$

Имеем

$$\begin{aligned} \frac{\partial}{\partial p} q(y_j, \hat{\mathbf{z}}_{\gamma,\mu}^{\mathbf{1}_j^p}) = 2\big(y_j - \hat{\mathbf{z}}_{\gamma,\mu,j}\big)^2 \frac{\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}}{\big(1 - p\mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j}\big)^3} = \\ = 2\big(y_j - \hat{\mathbf{z}}_{\gamma,\mu,j}\big)^2 \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j} \end{aligned} \tag{120}$$

Подстановка в DiffLOO

$$\text{DiffLOO}(\gamma, \mu) = \frac{1}{N} \sum_{j=1}^{N} \big(y_j - \hat{\mathbf{z}}_{\gamma,\mu,j}\big)^2 \mathbf{v}_{\mathbb{I}_{\gamma,\mu},j}^T \mathbf{D}_{\mathbb{I}_{\gamma,\mu},j} \tag{121}$$

6.6.2. Логистическая регрессия

Для логистической регрессии примем как матрицу $\mathbf{D}_{\mathbb{I}_{\gamma,\mu}}$ следующее выражение, которое вложено в формулу (106):

$$\tilde{\mathbf{D}}_{\mathbb{I}_{\gamma,\mu}} = \left(\mathbf{R}\left(\mathbf{X}^T_{\mathbb{I}(\lambda|\mu)}\mathbf{X}_{\mathbb{I}(\lambda|\mu)}\right) + \gamma\tilde{\mathbf{G}}^{-1}\right)^{-1} \tag{122}$$

С учетом измененной матрицы $\mathbf{D}_{\mathbb{I}_{\gamma,\mu}}$ и того факта, что новая целевая переменная равна $\tilde{y}_j$ получаем

$$\text{DiffLOO}(\gamma,\mu) = \frac{1}{N}\sum_{j=1}^{N}\left(\tilde{y}_j - \hat{\mathbf{z}}_{\gamma,\mu,j}\right)^2\mathbf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\tilde{\mathbf{D}}_{\mathbb{I}_{\gamma,\mu},j} \tag{123}$$

Заключение

В данном исследовании определена взвешенная задача в прямой и в двойственной форме записи как некоторое обобщение задачи восстановления зависимостей обобщенных линейных моделей, в которой объекты встречаются с некоторыми весами в обучающей совокупности. Рассматриваются две обобщенные линейные модели, а именно, линейную и логистическую регрессию вместе с селективной Ридж-регуляризацией, которая обладает свойством селективности.

На основе взвешенной задаче для рассматриваемых моделей получены вспомогательные алгоритмы численного решения для выбора активных признаков. Решение производится с помощью итерационного алгоритма Ньютона с переменным шагом при некоторых заданных ограничениях. Учтена неустойчивость, появляющаяся когда число объектов превышает число активных признаков в силу свойства селективности моделей, посредством изменения формулы для решения вектора множителей Лагранжа с помощью формулы Вудбери.

Разработан метод дифференциальной поэлеметной кросс-валидации для задач восстановления зависимостей обобщенных линейных моделей, который применим при селективном отборе признаков, и в то же время не увеличивает вычислительную сложность задачи обучения. А также получен критерий качества решения метода DiffLOO, как среднее скоростей роста ошибки, который гарантирует подбор оптимальных гиперпараметров μ и γ.

Результаты экспериментальных исследований соответствуют поставленным задачам, показывая эффективность алгоритма. Кроме того, в рассматриваемых прикладных задачах оптимальные значения гиперпараметров модели, на которых применен разработанный метод DiffLOO, приводят к отбору истинных оптимальных подмножеств признаков.

Список использованных источников

[1] Vapnik, V. *Estimation of Dependences Based on Empirical Data*, Springer Velarg New York, 1982.

[2] Nelder, J, Wedderburn, R. *Generalized Linear Models. Journal of the Royal*, Statistical Society. Series A (General), 1972, Vol. 135, Issue. 3, pp. 370—384.

[3] V. Mottl, V. Sulimova, O. Krasotkina, A. Morozov, A. Tatarchuk. *Computational Complexity of Dependence Estimation via Generalized Linear Models in Multidimensional Feature Spaces*, 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russia, 2019, 719-724.

[4] Tatarchuk, A., Mottl, V., Eliseyev, A., Windridge, *Selectivity supervision in combining pattern recognition modalities by feature- and kernel-selective Support Vector Machines*, Proceedings of the 19th International Conference on Pattern Recognition ICPR, 2008, Vol 1-6, 2336-2339.

[5] C. Cortes, V. Vapnik, *Support-Vector Networks. Machine Learning*, 1995, 273-297.

[6] Goldberg M.A. , Cho H.A., *Introduction to Regression Analysis*, WIT Press, 2004.

[7] M. Park, T. Hastie. *L1-Regularization path algorithm for generalized linear models.* Journal of the Royal Statistical Society, Series B (Statistical Methodology), 2007 Vol. 69, Part 4, pp. 659–677.

[8] J. Friedman, T. Hastie, R. Tibshirani. *Regularization paths for generalized linear models via coordinate descent.* Journ. of Stat.Soft, Series B (Statistical Methodology), 2010 , Vol. 33.

[9] Sharpe W.F. *Determining a fund's effective asset mix.* Investment Management Review, September/October 1988.

[10] Sharpe W.F. *Asset allocation: Management style and performance measurement.* The Journal of Portfolio Management, Winter 1992 , pp. 7-19.

[11] Tibshirani R. *Regression shrinkage and selection via the lasso* Journal of the Royal Statistical Society. Series B (Methodological) 1996, 267-288.

[12] A.E. Hoerl, R.W. Kennard, *Ridge regression: biased estimation for nonorthogonal problems.* Technometrics, 1970 , 12(1), 55–67.