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

Ангуло Яури Бриан Флориан

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: д.т.н. профессор В.В.Моттль

17 июня 2020 г.

Цель исследования

Цель

Разработать подходящий метод выбора оптимальных гиперпараметров для обобщенной линейной модели восстановления зависимостей при селективном отборе признаков.

Задача

Требуется найти оптимальные гиперпараметры модели в случае когда число признаков больше числа объектов.

Предложение

Удалить бесконечно малую часть из одного объекта обучающей совокупности, и измерить скорость роста потерь.

Задача восстановления зависимости

$$X = \{(x_j, y_j)_{j=1}^N\},\ x_j \in \mathbb{R}^n,\ y_j \in \mathbb{Y}$$

Требуется сконструировать решающее правило, применимое к каждому $x \in \mathbb{R}^n$

$$\hat{y} = f(x) : \mathbb{R}^n \to \mathbb{Y}$$

Обобщенная линейная модель

$z(x|a, b) = a^T x + b : \mathbb{R}^n \to \mathbb{R}$ – обобщенная линейная модель

где a – направляющий вектор, b – сдвиг

$q(y, z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+$ – функция связи

$\hat{y}(x|a, b) = \arg\min_{y \in \mathbb{Y}} q(y, z(x|a, b))$ – решающее правило

Обобщенный линейный подход

Минимум регуляризованного эмпирического риска

$$J(\mathrm{a}|\gamma,\mu) = \gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \le \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum_{j=1}^{N} q\left(y_j, \mathrm{x}_j^T \mathrm{a}\right) \to \min(\mathrm{a})$$

где μ — гиперпараметр селективности;
γ — гиперпараметр Ридж-регуляризации;

В качестве регуляризации применяется селективная Ридж-регуляриза--ция[1].
Результат — подмножество активных (информативных) признаков:
$\hat{\mathbb{I}}_{(\gamma,\mu)} = \{i : a_{\gamma,\mu} \neq 0\} \subseteq \{1,\ldots,n\}$

Обобщенный линейный подход

С учетом вышесказанного, что число признаков больше числа объектов, целесообразно использовать LOO.

Предположение

Из-за селективности признаков нельзя напрямую использовать LOO. Поскольку удаление одного объекта целиком, может привести к изменению подмножества активных признаков.

Идея

Удаление бесконечно малой части из одного объекта обучающей совокупности, не приведет к изменению подмножества активных (информативных) признаков. Доля объекта, которую удаляем, будет равна p, $p \to 0$.

Обобщенный линейный подход

Взвешенная задача

Рассмотрим немного более общую задачу.

Взвешенная задача

Во взвешенной задаче предполагается, что объекты встречаются в обучающей выборке с некоторыми весами $r_j = 1,\ r_t = 1 - p,\ p \to 0$

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

Можно показать, что следующая задача эквивалентна предыдущей:

Взвешенная задача

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

Для решения $J(\mathrm{a}|\gamma,\mu)$ удобно представить задачу в разделенной записи, следующим образом:

$$\begin{cases} \gamma \sum\limits_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, |a_i| \le \mu \\ \mu^2 + {a_i}^2, |a_i| > \mu \end{pmatrix} + \sum\limits_{t=1}^{N} r_j q(y_j, z_j) \to min(a \in \mathbb{R}^n, z_j) \\ z_j = \mathrm{a}^T \mathrm{x_j}, j = 1, \ldots, N \end{cases}$$

Решение задачи сводится к поиску седловой точки функции Лагранжа

$$-\frac{1}{2}\boldsymbol{\lambda}^T \left(\mathrm{X}^T_{\mathbb{I}_{\gamma,\mu}} \mathrm{X}_{\mathbb{I}_{\gamma,\mu}} \right) \boldsymbol{\lambda} + \frac{1}{2\gamma} \sum_{t=1}^{N} r_j q(y_j, z_j) + \mathrm{z}^T \boldsymbol{\lambda} \to \begin{cases} \min(\mathrm{z}) \\ \nabla_\lambda = 0 \end{cases}$$

Численное решение задачи

Итерационный алгоритм Ньютона

Предыдущая задача является выпуклой гладкой задачей оптимизации, если функция потерь выбрана также выпуклой и гладкой.

$$\hat{\boldsymbol{\lambda}}_{\gamma,\mu} = \arg\min \left\{ \frac{1}{2}\boldsymbol{\lambda}^T \left(X_{\mathbb{I}_{\gamma,\mu}}^T X_{\mathbb{I}_{\gamma,\mu}} \right) \boldsymbol{\lambda} + \sum_{j=1}^{N} \left[- \inf_{z \in \mathbb{R}} \left(\frac{1}{2\gamma} q(y_j, z_j) + \lambda_j z \right) \right] \right\}$$

Итерационный алгоритм Ньютоновского типа $\boldsymbol{\lambda}^k$, $k = 0, 1, \ldots$ основан на разложении выпуклых функции в ряд Тейлора на каждом шаге во крестности очередной точки $\boldsymbol{\lambda}^k$.
Когда решение двойственной задачи найдено $\hat{\boldsymbol{\lambda}}_{\gamma,\mu} \in \mathbb{R}^N$ компоненты направляющего вектора $\hat{a}_{\gamma,\mu}$ находятся независимо друг от друга:

$$\hat{a}_{\gamma,\mu,i} \begin{cases} 0, & |x_i{}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}| \leq \mu, \\ x_i{}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}, & |x_i{}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}| < \mu. \end{cases} \qquad \hat{\mathbb{I}}_{(\boldsymbol{\lambda}|\mu)} = \{ i : |x_i{}^T \hat{\boldsymbol{\lambda}}_{\gamma,\mu}| > \mu \} \subseteq \{1, \ldots, n\}$$

В случае линейной регрессии можно получить следующее выражение:

$$W(\boldsymbol{\lambda}|\gamma,\mu) = \boldsymbol{\lambda}^T\left(X^T_{\mathbb{I}_{\gamma,\mu}}X_{\mathbb{I}_{\gamma,\mu}}\right)\boldsymbol{\lambda} + \frac{\gamma}{2}\boldsymbol{\lambda}^T\boldsymbol{\lambda} - Y^T\boldsymbol{\lambda} \to \min(\boldsymbol{\lambda})$$

Для решения задачи воспользуемся итерационным алгоритмом Ньютона с переменным шагом.

$$\begin{cases} \tilde{\boldsymbol{\lambda}}^{k+1} = \arg\min\limits_{\boldsymbol{\lambda}} \tilde{W}^k(\boldsymbol{\lambda}|\gamma,\mu) \\ \tilde{W}^k(\boldsymbol{\lambda}|\gamma,\mu) = \boldsymbol{\lambda}^T\left(X^T_{\mathbb{I}^k}X_{\mathbb{I}^k}\right)\boldsymbol{\lambda} + \frac{\gamma}{2}\boldsymbol{\lambda}^T\boldsymbol{\lambda} - Y^T\boldsymbol{\lambda} \to \min(\boldsymbol{\lambda}) \end{cases}$$

Критерий остановки: $\hat{\mathbb{I}}(\boldsymbol{\lambda}^{k+1}) = \hat{\mathbb{I}}(\boldsymbol{\lambda}^k)$

Численное решение задачи

Итерационный алгоритм Ньютона

Поиск седловой точки сводится к системе линейных уравнений.
Решение системы показывает только наилучшее направление шага

Линейная регрессия

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

Логистическая регрессия

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

где $\tilde{\mathrm{y}}^k$ и $\tilde{\mathrm{G}}^k$ — переменные, полученные при квадратичном представлении логистической фунции потерь в окрестности точки z^k ряда Тейлора.

Может случиться так, что длина шага очень большая, тогда следует ее сократить

Проверка

если $W(\tilde{\lambda}^{k+1}|\gamma,\mu,\mathsf{r}) \leq W(\tilde{\lambda}^{k}|\gamma,\mu,\mathsf{r})$, $\lambda^{k+1} = \tilde{\lambda}^{k+1}$;

если $W(\tilde{\lambda}^{k+1}|\gamma,\mu,\mathsf{r}) > W(\tilde{\lambda}^{k}|\gamma,\mu,\mathsf{r})$, $-$длину следует сократить;

Для нахождения подходящей длины шага Ньютона используем метод золотого сечения

Одномерная оптимизация

$$\tau^{k+1} = \arg\min W[(1-\tau)\hat{\lambda}^{k} + \tau\tilde{\lambda}^{k+1}|\gamma,\mu], 0 < \tau < 1,$$

$$\hat{\lambda}^{k+1} = \tau^{k+1}\hat{\lambda}^{k} + (1-\tau^{k+1})\tilde{\lambda}^{k+1}$$

Дифференциальный LOO

Формула Вудбери

Пусть исходная двойственная задача решена:

Вектор-решение

$$\hat{\lambda}^{1}_{\gamma,\mu} = \underbrace{\left(\left(\mathrm{X}^{T}_{\mathbb{I}_{\gamma,\mu}}\mathrm{X}_{\mathbb{I}_{\gamma,\mu}}\right) + \gamma E\right)}_{D}^{-1} \mathrm{y}$$

Решение возмущенной двойственной задачи когда вес j-го объекта $r_j = 1 - p$

Вектор-решение с $R^{1^p_j}$

$$\hat{\lambda}^{1^p_j}_{\gamma,\mu} = \left[\left(\mathrm{X}^{T}_{\mathbb{I}_{\gamma,\mu}}\mathrm{X}_{\mathbb{I}_{\gamma,\mu}} + \gamma E\right) - p\, 1^j\, \mathrm{v}^{T}_{\mathbb{I}_{\gamma,\mu},j}\right]^{-1}\left[\mathrm{y} - p\, \mathrm{y}_j\, 1^j\right]$$

где 1^j - вектор-столбец: $1_t = 0, t \in \{1,\ldots,N\}\backslash j, 1_j = 1$

Решение задачи с заглушенным объектом, вес у которого равен $1 - p$, выражается следующим образом.

Решение задачи с заглушенным объектом

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

В терминах z решение следующее

Решение в терминах обобщенных признаков

$$\hat{\mathrm{z}}^{1_j^p}_{\gamma,\mu,j} = \hat{\mathrm{z}}_{\gamma,\mu,j} - \frac{p\mathrm{v}^T_{\mathbb{I}_{\gamma,\mu},j} \mathrm{D}_{\mathbb{I}_{\gamma,\mu},j}}{1 - p\mathrm{v}^T_{\mathbb{I}_{\gamma,\mu},j} \mathrm{D}_{\mathbb{I}_{\gamma,\mu},j}} (y_{\gamma,\mu,j} - \hat{z}_{\gamma,\mu,j})$$

С учетом формулы Вудбери получаем

Отклонение $\hat{z}_j^{1_j^p}$ от целевой переменной y_j

$$y_j - \hat{z}_{\gamma,\mu}^{1_j^p} = \frac{y_j - \hat{z}_{\gamma,\mu,j}}{1 - p v_{\mathbb{I}_{\gamma,\mu},j}^T D_{\mathbb{I}_{\gamma,\mu},j}}$$

Линейную регрессию можно представить как

Линейная регрессия

$$q(y_j, \hat{z}_{\gamma,\mu}^{1_j^p}) = \left(\frac{y_j - \hat{z}_{\gamma,\mu,j}}{1 - p v_{\mathbb{I}_{\gamma,\mu},j}^T D_{\mathbb{I}_{\gamma,\mu},j}} \right)^2$$

Наш критерий *DiffLOO* построен как среднее скоростей роста потерь для каждого отдельного объекта

Общий критерий дифференциального LOO

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

Для линейной регрессии получаем

Критерий дифференциального LOO

$$\mathrm{DifLOO}(\gamma,\mu) = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - \hat{z}_{\gamma,\mu,j}\right)^2\left(\mathrm{v}^T_{\mathbb{I}_{\gamma,\mu,j}}\mathrm{D}_{\mathbb{I}_{\gamma,\mu,j}}\right)$$

Для логистической регрессии получаем

Полученный критерий дифференциального LOO

$$\mathrm{DifLOO}(\gamma,\mu) = \frac{1}{N}\sum_{j=1}^{N}\left(\tilde{y}_j - \mathsf{v}_{\mathbb{I}_{\gamma,\mu},j}\hat{\boldsymbol{\lambda}}_{\gamma,\mu}\right)^2\left(\mathsf{v}^T_{\mathbb{I}_{\gamma,\mu},j}\tilde{\mathsf{D}}_{\mathbb{I}_{\gamma,\mu},j}\right)$$

Эксперименты

Линейная регрессия

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

$$X_t = (x_{t,i}, i = 1, \ldots, N) \in \mathbb{R}^n, \ n = 650$$

Наблюдаемый сигнал, состоящий из $N = 251$.

$$y_t, \ t = 1, \ldots, N = 251$$

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

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

где a_i - это доли вложения капитала.

Эксперименты

Результаты экспериментов для линейной регрессии

Найденное подмножество $\hat{\mathbb{I}}_{\gamma,\mu}$ содержит $\hat{n}_{\gamma,\mu} = 16$ биржевых активов, содержащее $n* = 13$ истинных.

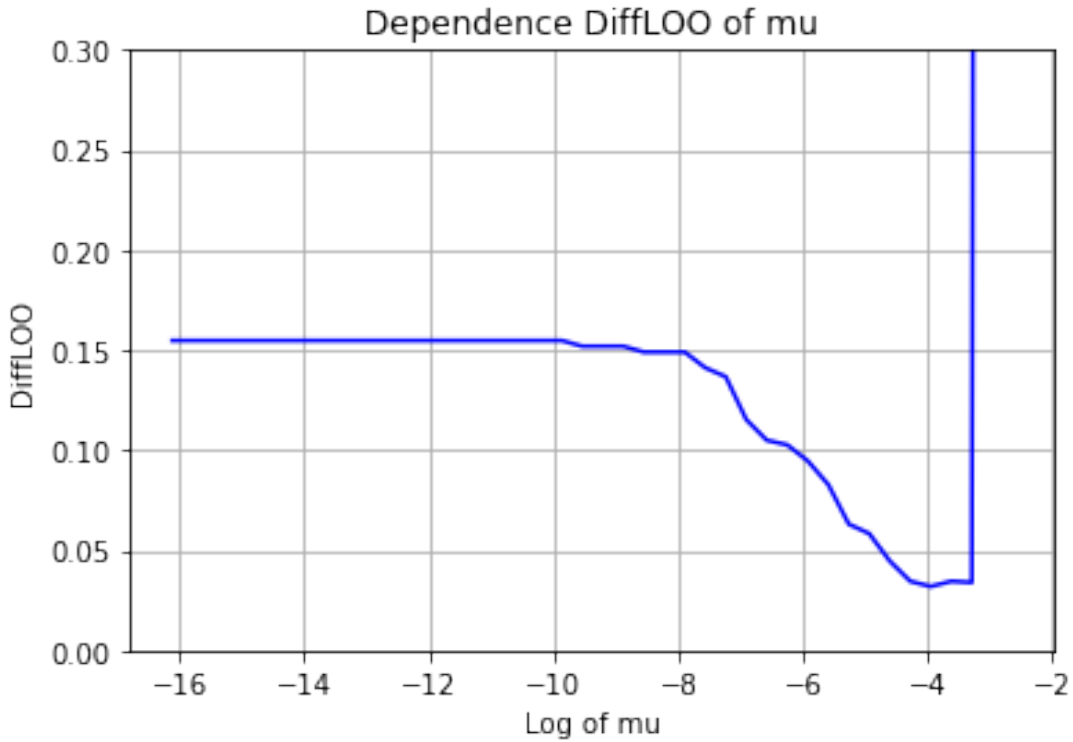

Figure: Зависимость дифференциального LOO от параметра μ

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.

$$X_t = (x_{t,i=1,\ldots,N}) \in \mathbb{R}^n, \ n = 400$$

Целевая переменная, состоящая из $N = 200$.

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

Оптимальные признаки заданы из нормального распределения с $\mathbb{E} = 0$ и $\mathbb{D} = 1$.

Эксперименты

Результаты экспериментов для логистической регрессии

Найденное подмножество $\hat{\mathbb{I}}_{\gamma,\mu}$ в точности совпадает с оптимальным подмножеством.

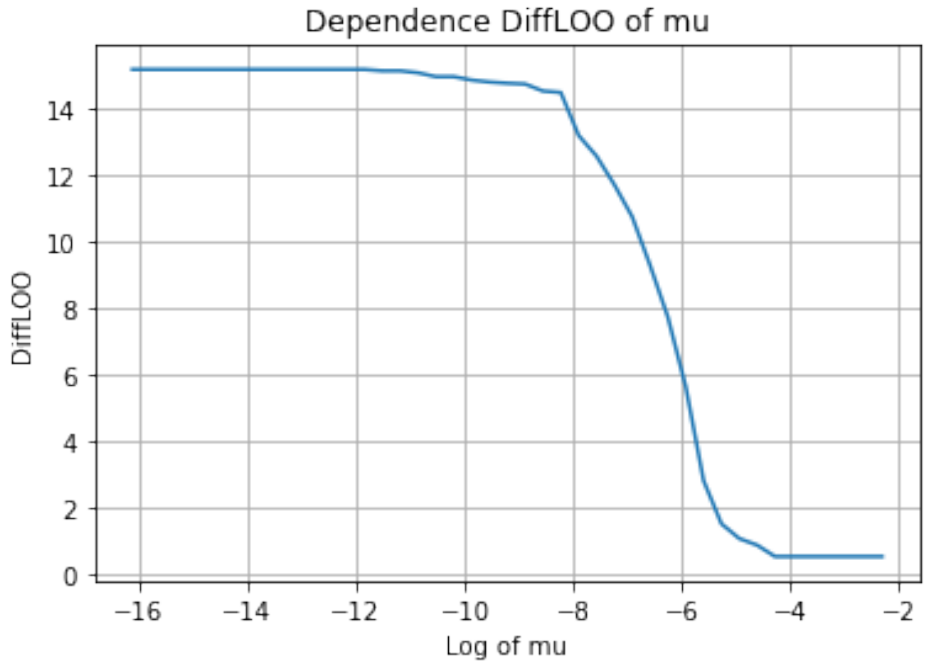

Figure: Зависимость дифференциального LOO от параметра μ

Публикация результатов

1. Ангуло Б., Морозов А.О., Моттль В.В. Метод дифференциальной кросс-валидации для выбора уровня сложности обобщенных линейных моделей зависимостей. Математические методы распознавания образов: Тезисы докладов 19-ой Всероссийской конференции с международным участием, г. Москва, 2019 г.
2. B. Angulo, A. Morozov, V. Mottl, A. Tatarchuk, O. Krasotkina. Differential Leave-One-Out Cross-Validation for Hyperparameter Optimization in Generalized Linear Dependence Models. 25-th International Conference on Pattern Recognition ICPR- 2021. Milan, Italy, January 10-15, 2021. (To appear)

Заключение

- Сформулирован и подобран алгоритм численного решения для выбора активных признаков;
- Разработан метод дифференциальной кросс-валидации для выбора оптимальных гиперпараметров модели;
- Получен критерий качества решения;
- Продемонстрирована правильность и корректность работы метода на примере прикланых задач.

планы

- Изменить метод на общий случай когда каждый объект будет учитываться с разным весом
- Изменить метод на общий случай когда вместо скалярных произведений можно было бы взять ядро
- Усовершенствовать имеющее численное решение на эти случаи