Алгоритмы выбора модели регрессии в больших массивах данных

Морозов Алексей Олегович

Выпускная квалификационная работа на соискание степени магистра.

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

Московский Физико-Технический Институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

21 июня 2018 г.

Проблема линейного регрессионного анализа при большом числе регрессоров

Конечное множество объектов: $t \in \{1, \ldots, T\}$
Скрытые значения целевой характеристики объектов: $y_t \in \mathbb{R}$
Совокупность наблюдаемых признаков объектов (регрессоров):
$\mathbf{x}_t = (x_{t,i}, i = 1, \ldots, n) \in \mathbb{R}^n$

Основное предположение:

$$y_t \cong \sum_{i=1}^{n} \beta_i x_{t,i} = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{x}_t$$

$$\boldsymbol{\beta} = (\beta_i, i = 1, \ldots, n) \in \mathbb{R}^n$$

Обучающая совокупность: $\{(\mathbf{x}_t, y_t), t = 1, \ldots, T\}$
Классическая задача регрессионного анализа для большого числа регрессоров $n \gg T$: найти значения коэффициентов регрессии, позволяющие наилучшим образом предсказывать значение целевой характеристики для новых объектов

Проблема линейного регрессионного анализа при большом числе регрессоров

Для этого, как правило, надо сократить число активных регрессоров $\hat{\mathbb{I}} = \{i : \beta_i > 0\} \subset \mathbb{I} = \{1, \ldots, n\}$, чтобы повысить обобщающую способность модели.

Специфика нашей задачи Factor Search: предполагается, что скрытый механизм формирования модели задал небольшое подмножество активных регрессоров $\hat{n} = |\hat{\mathbb{I}}| \ll n = |\mathbb{I}|$ с характерной совокупностью регрессоров при них.

Необходимость учета априорной информации о механизме формирования структуры модели регрессии

Если $n \gg T$, то регрессоры $\boldsymbol{x}_i = (x_{t,i}, t = 1, \ldots, T) \in \mathbb{R}^{\mathsf{T}}$ неизбежно сильно коррелированы.
Factor Search возможен только при наличии априорной информации о специфике активных регрессоров.
Такая задача известна в литературе в предположении упорядоченных признаков: $\mathbb{I} = \{1, \ldots, n\}$

Моттль В.В., Двоенко С.Д., Середин О.С., Долгова О.В. Обучение распознаванию сигналов с учетом критерия гладкости решающего правила. Доклады IX Всероссийской конференции «ММРО», Москва, 22-26 ноября 1999 г., с. 86-88

Поиск подмножества активных регрессоров по принципу Beta Parity

$$\begin{cases} \sum\limits_{t=1}^{T} \left(y_t - \sum\limits_{i=1}^{n} \beta_i x_{t,i} \right)^2 \longrightarrow \min(\beta_1, \ldots, \beta_n) \\ \sum\limits_{i=1}^{n} \beta_i = 1,\ \beta_i \geqslant 0,\ i = 1, \ldots, n \end{cases}$$

Идея наделения такой задачи свойством селективности Beta Parity – введение регуляризатора Modulus Quadratic:

Татарчук А.И. Байесовские методы опорных векторов для обучения распознаванию образов с управляемой селективностью отбора признаков. Диссертация к.ф.-м.н. ВЦ РАН, 2014

Регрессионный критерий Beta Parity:

$$\begin{cases} \sum\limits_{i=1}^{n} \begin{pmatrix} 2\mu\beta_i, & \beta_i \leqslant \mu \\ \mu^2 + \beta_i^2, & \beta_i > \mu \end{pmatrix} + c \sum\limits_{t=1}^{T} \left(y_t - \sum\limits_{i=1}^{n} \beta_i x_{t,i} \right)^2 \longrightarrow \min(\beta_1, \ldots, \beta_n) \\ \sum\limits_{i=1}^{n} \beta_i = 1,\ \beta_i \geqslant 0,\ i = 1, \ldots, n \end{cases}$$

Финансовая интерпретация — априорное предположение о диверсифицированной структуре инвестиционного портфеля

Известные доходности портфеля: $y_t, t = 1, \ldots, T$
Доходности очень большого числа биржевых активов:
$\mathbf{x}_t = (x_{t,i}, i = 1, \ldots, n) \in \mathbb{R}^n$
Найти относительно небольшое подмножество активов, из которых фактически состоит портфель: $\hat{\mathbb{I}} \subset \mathbb{I}$, $\hat{n} = |\hat{\mathbb{I}}| \ll n = |\mathbb{I}|$
Регуляризация Beta Parity — капитал вложен в выбранные активы в равных долях: $\beta_i \cong \text{const}$, $i \in \hat{\mathbb{I}} \subset \mathbb{I}$

Более сложное понимание диверсифицированности портфеля — Risk Parity

Дисперсия $D(\beta)$ доходности портфеля β

$$D(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^{\mathsf{T}}x_j)\beta_j\beta_i = \sum_{i=1}^{n}\underbrace{\left(\sum_{j=1}^{n}(x_i^{\mathsf{T}}x_j)\beta_j\right)}_{\text{доля } i\text{-го актива}}\beta_i$$

$$D(\beta) = \sum_{i=1}^{n} D_i(\beta),\ D_i(\beta) = \left(\sum_{j=1}^{n}(x_i^{\mathsf{T}}x_j)\beta_j\right)\beta_i$$

Risk Parity — равенство долей риска от всех выбранных активов:
$D_i(\beta) \cong \text{const},\ i \in \hat{\mathbb{I}} \subset \mathbb{I}$

Предположение формировании инвестиционного портфеля по принципу Risk Parity

$$\begin{cases} \underbrace{\sum_{i=1}^{n}\left(\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j)\beta_j\right)\beta_i}_{\text{общий риск}} \longrightarrow \min \\ \sum_{i=1}^{n}\beta_i = 1,\ \beta_i \geqslant 0,\ i = 1,\ldots,n \end{cases}$$

Задача выпуклого программирования
Оптимальный портфель: минимум риска
Очень селективный критерий
Противоречит принципу диверсификации

Risk Parity — сочетание требований малого риска и диверсификации:

$$\begin{cases} \sum_{i=1}^{n}\left[nD_i - \sum_{j=1}^{n}D_j\right]^2 \longrightarrow \min(\beta_1,\ldots,\beta_n) \text{ Все активы вошли в портфель} \\ D_i = \left(\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j)\beta_j\right)\beta_i,\ \beta_i \geqslant 0,\ i = 1,\ldots,n,\ \sum_{i=1}^{n}\beta_i = 1 \end{cases}$$

Идея управления размером портфеля Risk Parity – параметр размера
$d_{\min} \leqslant d \leqslant d_{\max}$, $d_{\max} = \max\limits_i(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i)$

Параметрическое семейство портфелей Risk Parity

$$\begin{cases} \sum\limits_{i=1}^{n} D_i^2 \longrightarrow \min(\beta_1, \ldots, \beta_n) \\ \sum\limits_{i=1}^{n} D_i = d, \;\; D_i = \left(\sum\limits_{j=1}^{n} (\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j) \beta_j \right) \beta_i \\ \sum\limits_{i=1}^{n} \beta_i = 1, \;\; \beta_i \geqslant 0, \;\; i = 1, \ldots, n \end{cases} \quad \Longrightarrow \quad \begin{array}{c} \underbrace{\mathbb{I} = \{i = 1, \ldots, n\}} \\ \text{все активы} \\ \underbrace{\mathbb{I}_d = \{i : \hat{\beta}_i > 0\} \subseteq \mathbb{I}} \\ \text{портфель с параметром } d \end{array}$$

Меньше d — меньше риск, больше портфель

Больше d — больше риск, меньше портфель

Параметр d — психологическая характеристика инвестора

Оценивание состава наблюдаемого портфеля по принципу Risk Parity

$\mathbf{y} = (y_t, t = 1, \ldots, T)$ — наблюдаемый портфель,представленный временным рядом его доходностей
Однопараметрическое семейство моделей Risk Parity:

$$\mathbb{J} = \{\hat{\mathbb{I}}_d : \ d_{\min} \leqslant d \leqslant d_{\max}\}$$

Какая модель $d_{\min} \leqslant d \leqslant d_{\max}$ из семейства $\mathbb{J}$ наилучшим образом соответствует временному ряду $\mathbf{y} \in \mathbb{R}^T$?
Критерий Leave-One-Out:

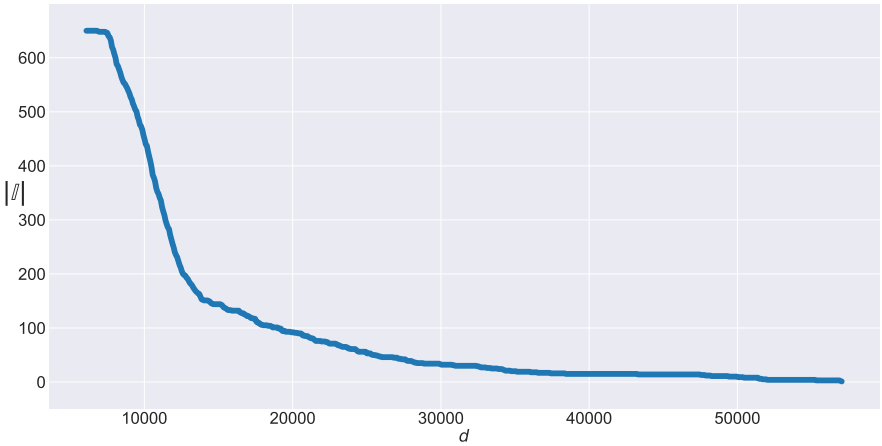
$$LOO(d) = \frac{1}{T} \sum_{t \in \mathbb{T}} \left(y_t - \sum_{i=1}^{n} \beta_{d,i}^{(t)} x_{t,i} \right)^2$$

Эксперименты

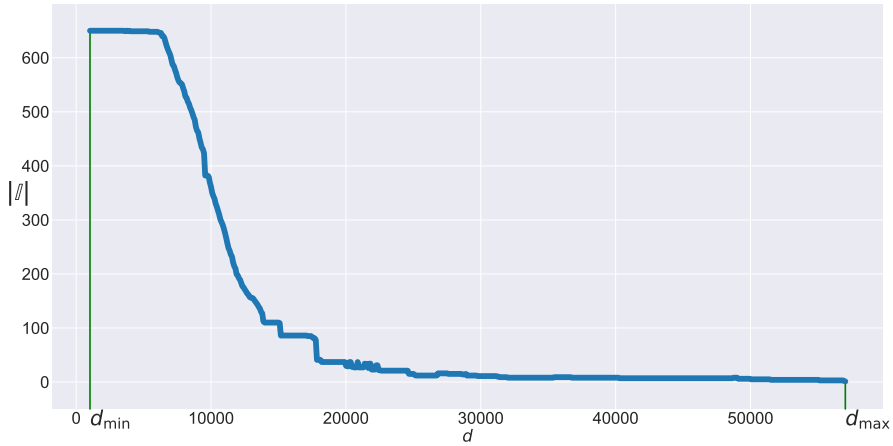
Таблица: Состав портфеля Risk Parity

Размер портфеля	Номера индексов
1	**7**
2	**3**, 7
3	**2**, 3, 7
4	**1**, 2, 3, 7
5	1, 2, 3, 7, **35**
6	1, 2, 3, 7, **34**, 35
7	1, 2, 3, 7, **12**, **13**, 35
8	1, 2, 3, 7, 12, 13, **16**, 35
8	1, 2, 3, **4**, 7, **20**, **21**, **50**
9	1, 2, 3, 4, 7, 20, 21, **31**, 50
10	1, 2, 3, 4, 7, 20, 21, 31, **41**, 50
11	1, 2, 3, 4, 7, 20, 21, 31, 41, 50,**92**
12	1, 2, 3, 4, 7, 20, 21, 31, 41, 50,**62**,92

Эксперименты



Размер портфеля Beta Parity при увеличении параметра d



Размер портфеля Risk Parity при увеличении параметра d

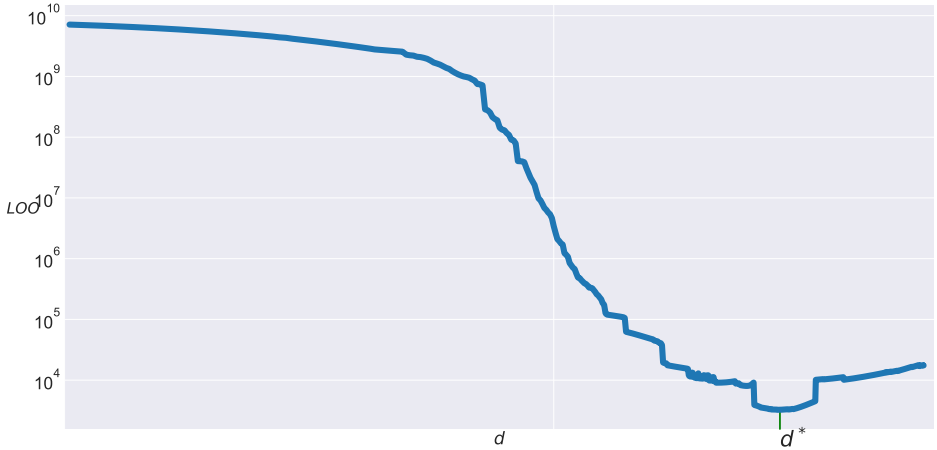


Рис.: $LOO(d)$

Публикация по материалам диссертации

. Krasotkina, M. Markov, V. Mottl, D. Babichev, I. Pugach, A. Morozov. Constrained Regularized Regression Model Search in Large Sets of Regressors. *Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science*, Springer, 2018 (to appear)

Спасибо за внимание!