Алгоритмы выбора модели регрессии в больших массивах данных

А.О. Морозов
МФТИ, Москва
Markov Processes International, USA

Научный руководитель

Д.т.н, профессор В.В. Моттль

Проблема линейного регрессионного анализа при большом числе регрессоров

$t \in \{1,...,T\}$	Конечное множество объектов
$y_t \in \mathbb{R}$	Скрытые значения целевой характеристики объектов
$\mathbf{x}_t = (x_{t,i}, i = 1,...,n) \in \mathbb{R}^n$	Совокупность наблюдаемых признаков объектов (регрессоров)
$y_t \cong \sum_{i=1}^{n} \beta_i x_{t,i} = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_t$ $\boldsymbol{\beta} = (\beta_i, i = 1,...,n) \in \mathbb{R}^n$	Основное предположение: Целевая характеристика примерно равна линейной неизвестной комбинацией регрессоров
$\{(\mathbf{x}_t, y_t), t = 1,...,T\}$	Обучающая совокупность с известными значениями целевой характеристики

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

Специфика нашей задачи Factor Search

Предполагается, что скрытый механизм формирования модели задал небольшое подмножество активных регрессоров $n^* = |\mathbb{I}^*| \ll n = |\mathbb{I}|$ с характерной совокупностью регрессоров при них. Требуется найти это реально существующее подмножество.

Необходимость учета априорной информации о механизме формирования структуры модели регрессии

Если $n \gg T$, то регрессоры $\boldsymbol{x}_i = (x_{t,i}, t = 1,...,T) \in \mathbb{R}^T$ неизбежно сильно коррелированы.

Поиск небольшого подмножества активных регрессоров – поиск иголки в стоге сена. Factor Search возможен только при наличии априорной информации о специфике активных регрессоров.

Например, существует класс прикладных задач, в которых априори известно, что активные коэффициенты регрессии близки друг к другу.

Такая задача известна в литературе в предположении упорядоченных признаков:

$$\mathbb{I} = \{1,...,n\} \text{ – номера индексов абсолютны}$$

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 «Математические методы распознавания образов», Москва, 22-26 ноября 1999 г., с. 86-88
— без отбора активных признаков.

O. Seredin, A. Kopylov, V. Mottl. Selection of subsets of ordered features in Machine Learning. Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science, Vol. 5632, Springer-Verlag, Berlin / Heidelberg, 2009, pp. 16-28
— с отбором активных признаков.

В данной работе предлагается методология отбора неупорядоченных регрессоров с примерно равными коэффициентами, которую мы будем называть принципом Beta Parity.

Поиск подмножества активных регрессоров по принципу Beta Parity

Многим приложениям адекватна задача регрессионного анализа с ограничениями-неравенствами на знаки коэффициентов $\beta_i \geq 0, i = 1,...,n$, и с ограничением-равенством на их сумму $\sum_{i=1}^{n} \beta_i = 1$	$\begin{cases} \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{n} \beta_i x_{t,i} \right)^2 \to \min(\beta_1,...,\beta_n) \\ \sum_{i=1}^{n} \beta_i = 1, \beta_i \geq 0, i = 1,...,n \end{cases}$ задача квадратичного программирования

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.

Mottl V., Seredin O., Krasotkina O. Compactness Hypothesis, Potential Functions, and Rectifying Linear Space in Machine Learning. Key Ideas in Learning Theory from Inception to Current State: Emmanuel Braverman's Legacy, Springer, 2018 (to appear).

Регрессионный критерий Beta Parity. Параметр селективности $\mu \geq 0$	$\begin{cases} \sum_{i=1}^{n} \begin{pmatrix} 2\mu\beta_i, & \beta_i \leq \mu \\ \mu^2 + \beta_i^2, & \beta_i > \mu \end{pmatrix} + c \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{n} \beta_i x_{t,i} \right)^2 \to \min(\beta_1,...,\beta_n) \\ \sum_{i=1}^{n} \beta_i = 1, \beta_i \geq 0, i = 1,...,n \end{cases}$ задача выпуклого программирования

Финансовая интерпретация – априорное предположение о диверсифицированной структуре инвестиционного портфеля

Задача идентификации инвестиционного портфеля

$y_t,\ t=1,\dots,T$	Известные доходности портфеля
$\mathbf{x}_t=(x_{t,i},\ i=1,\dots,n)\in\mathbb{R}^n$	Доходности очень большого числа биржевых активов, $n\gg T$
$\hat{\mathbb{I}}\subset\mathbb{I},\quad \hat{n}=\lvert\hat{\mathbb{I}}\rvert\ll n=\lvert\mathbb{I}\rvert$	Найти относительно небольшое подмножество активов, из которых фактически состоит портфель …
Регуляризация Beta Parity $\beta_i\cong const$ для $i\in\hat{\mathbb{I}}\subset\mathbb{I}$	Априорное предположение о диверсифицированной структуре инвестиционного портфеля – Капитал вложен в выбранные активы в равных долях.

Более сложное понимание диверсифицированности портфеля – Risk Parity

Дисперсия $D(\boldsymbol{\beta})$ доходности портфеля $\boldsymbol{\beta}$. Биржевой жаргон – риск разорения	$D(\boldsymbol{\beta})=\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)\beta_i\beta_j=\sum_{i=1}^{n}\underbrace{\left(\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)\beta_j\right)}_{\text{доля } i\text{-го актива}}\beta_i$
Risk Parity – равность долей риска от всех выбранных активов $D_i(\boldsymbol{\beta})\cong const$ для $i\in\hat{\mathbb{I}}\subset\mathbb{I}$	$D(\boldsymbol{\beta})=\sum_{i=1}^{n}D_i(\boldsymbol{\beta}),\quad D_i(\boldsymbol{\beta})=\left(\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)\beta_j\right)\beta_i$

Восстановление скрытого состава инвестиционного портфеля по принципу Risk Parity

(пока без селекции)

$\begin{cases} \sum_{i=1}^{n}\left[\left(\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)\beta_j\right)\beta_i\right]^2 + c\sum_{t=1}^{T}\left(y_t - \sum_{i=1}^{n}\beta_i x_{t,i}\right)^2 \to \min(\beta_1,\ldots,\beta_n) \\ \sum_{i=1}^{n}\left(\sum_{j=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)\beta_j\right)\beta_i = D,\ \sum_{i=1}^{n}\beta_i = 1,\ \beta_i \geq 0,\ i = 1,\ldots,n \end{cases}$	Задача выпуклой оптимизации

Поиск подмножества активных регрессоров по принципу Beta Parity

Критерий и алгоритм в стадии отладки.

Публикация по материалам диссертации

O. 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).