

Анализ метода отбора признаков QPFS для обобщенно-линейных моделей

Александр Дмитриевич Толмачев

Московский физико-технический институт

*Курс: Автоматизация научных исследований
(практика, В. В. Стрижов)/Группа 821*

Эксперт: В. В. Стрижов

Консультант: А. А. Адуенко

2021

Задача отбора признаков в обобщенно-линейной модели

Цель

Исследовать проблему отбора признаков в обобщенно-линейной модели.

Исследуемая проблема

Требуется оценить точность метода QPFS в задаче отбора признаков и сравнить его с другими методами.

Метод решения

Рассматриваются задачи оптимизации в методе QPFS и других методах отбора признаков.

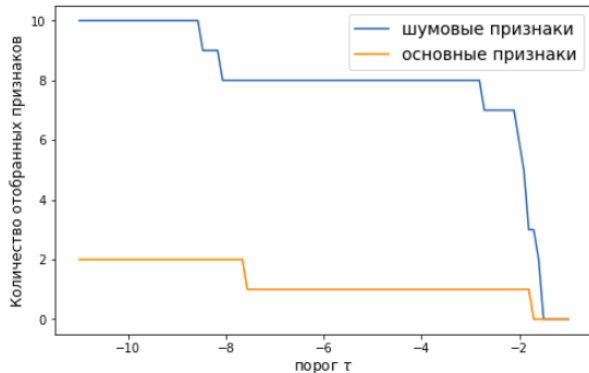
Применение квадратичной оптимизации для задачи отбора признаков

$$\begin{cases} z^* = \arg \min_{z \in [0,1]^n} z^T Q z - b^T z \\ \|z\|_1 \leq 1 \end{cases}$$

Q – матрица схожести между признаками




b – вектор схожести между признаком и целевым вектором.

τ – порог, т.ч. $z_i^* > \tau \Leftrightarrow j$ -ый признак отобран моделью



Замечаем, что очень важно правильно подбирать значение порога τ , но даже при таком подходе мы можем отобрать много шумовых признаков...

Основная литература

-  Irene Rodriguez-Lujan и др. “Quadratic Programming Feature Selection”. в: *Journal of Machine Learning Research* 11 (апр. 2010), с. 1491—1516.
-  Aleksandr Katrutsa и Vadim Strijov. “Stresstest procedure for feature selection algorithms”. в: *Chemometrics and Intelligent Laboratory Systems* 142 (февр. 2015). DOI: 10.1016/j.chemolab.2015.01.018.
-  Alexandr Katrutsa и Vadim V. Strijov. “Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria”. в: *Expert Syst. Appl* 76 (2017), с. 1—11.

Постановка задачи метода QPFS

Добавим нормировку в оптимизируемый функционал:

$$\begin{aligned} \frac{1}{2}(1 - \alpha)a^T Q a - \alpha b^T a \rightarrow \min_a \\ \text{s.t. } a \geq 0, \sum_{i=1}^n a_i = 1. \end{aligned} \quad (1)$$

$$\begin{aligned} \frac{1}{2}\tilde{a}^T Q \tilde{a} - b^T \tilde{a} \rightarrow \min_{\tilde{a}} \\ \text{s.t. } \tilde{a} \geq 0, \sum_{i=1}^n \tilde{a}_i \leq \beta \end{aligned} \quad (2)$$

Q - матрица корреляций Пирсона между признаками

b - вектор корр-ий Пирсона между признаком и целевым вектором.

Признак j активен $\Leftrightarrow a_j > 0$.

Анализ решения задачи оптимизации

Таблица: Свойства решения в методе QPFS в зависимости от параметра α

α	Равенство	Неравенство
$\alpha = 0$	Исп-тся все признаки, $Qa^* = \eta e, \eta > 0$	исключены все признаки
$\alpha \rightarrow 0$	Исп-тся все признаки, $Qa^* \rightarrow \eta e, \eta > 0$	Решение задачи (2)
$\alpha \rightarrow 1$	Сх-ся к отбору 1го признака с максимальным b_j	То же, что и в «рав-во»
$\alpha = 1$	Отбор 1го признака с макс. b_j	То же, что и в «рав-во»

Связь с Lasso-моделью для линейной регрессии

$$\frac{1}{2}\|y - Xw\|_2^2 + \tau\|w\|_1 \rightarrow \min_w. \quad (3)$$

$$\begin{aligned} \frac{1}{2}w^T \tilde{Q}w - \tilde{b}^T w &\rightarrow \min_w \\ \text{s.t. } \|w\|_1 &= \eta. \end{aligned} \quad (4)$$

Если $x_j^T x_j = 1$, $y^T y = 1$, $y^T x_j \geq 0$, $x_j^T x_l \geq 0$ и $w^* \geq 0$, то эти задачи тождественны!

Вычислительный эксперимент

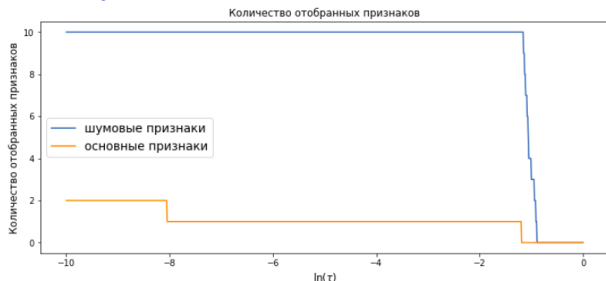
Цель

Выявить недостатки метода QPFS.

Выборка

Пусть $x_1 \sim \mathcal{N}(0, 1)$, $y \sim \mathcal{N}(0, 1)$, а $x_2 = x_1 + \varepsilon \cdot y$, где $\varepsilon = 0.001$.

Результаты

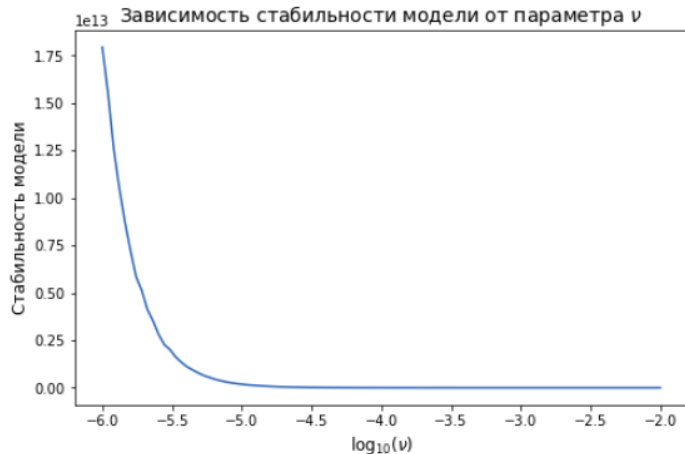


В методе QPFS может отбираться много шумовых признаков...

При этом, $y = \frac{x_2 - x_1}{\varepsilon}$, т.е. не выполнено одно из условий эквивалентности QPFS и Lasso.

Стабильность модели

$$\|y - Xw\|_2^2 + \tau \lambda_{\max}(X(w)^T X(w)) / \lambda_{\min}(X(w)^T X(w)) \rightarrow \min_w,$$



В качестве апостериорного знания можно рассматривать индекс обусловленности, однако такая задача очень тяжела для оптимизации...

Результаты

- ▶ проанализирован метод QPFS отбора признаков при различных постановках задач оптимизации,
- ▶ показана эквивалентность QPFS и Lasso при определенных условиях
- ▶ рассмотрены способы достижения стабильности модели

Направления будущей работы

- ▶ исследовать возможности применения байесовского подхода к методу QPFS при различных априорных распределениях,
- ▶ рассмотреть новые методы отбора признаков и сравнить их с QPFS