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: необходимо выбрать оптимальное подмножество признаков из множества всех признаков, среди которых есть мультиколлинеарные.

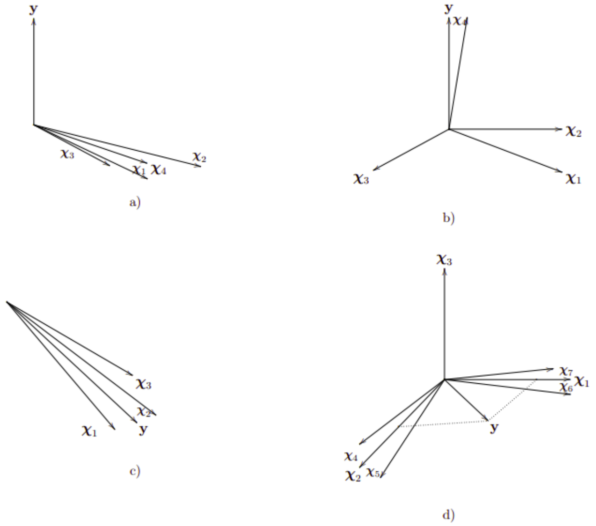
Исследуемая проблема

Требуется оценить точность метода QPFS в задаче отбора признаков и рассмотреть его связь с другими методами (например, с Lasso-регрессией). Необходимо сравнить точность отбора признаков различными методами.

Метод решения

Рассматриваются различные постановки задачи оптимизации в методе QPFS. Показана эквивалентность методов QPFS и Lasso-регрессии при определенных условиях. Рассматриваются подходы для улучшения метода QPFS на основе комбинации признаков.

Применение квадратичной оптимизации для задачи отбора признаков



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

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

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:

$$\begin{aligned} &\frac{1}{2}(1-\alpha)\mathrm{a}^{\mathsf{T}}\mathrm{Q}\mathrm{a} - \alpha\mathrm{b}^{\mathsf{T}}\mathrm{a} \to \min_{\mathrm{a}} \\ &\text{s.t. } \mathrm{a} \geq 0,\ \sum_{i=1}^{n} a_i = 1. \end{aligned} \tag{1}$$

$$\begin{aligned} &\frac{1}{2}\tilde{\mathrm{a}}^{\mathsf{T}}\mathrm{Q}\tilde{\mathrm{a}} - \mathrm{b}^{\mathsf{T}}\tilde{\mathrm{a}} \to \min_{\tilde{\mathrm{a}}} \\ &\text{s.t. } \tilde{\mathrm{a}} \geq 0 \end{aligned} \tag{2}$$

Q - матрица корреляций Пирсона между признаками

b - вектор кореляций Пирсона между признаком и целевым вектором.

Признак j активен $\Leftrightarrow a_j > 0$.

Анализ решения задачи оптимизации в методе QPFS

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

α	Равенство	Неравенстсво
$\alpha = 0$	используются все признаки, $\mathrm{Qa}^* = \eta \mathrm{e},\ \eta > 0$	исключены все признаки
$\alpha \to 0$	используются все признаки, $\mathrm{Qa}^* \to \eta \mathrm{e},\ \eta > 0$	решение задачи (2)
$\alpha \to 1$	сходится к отбору одного признака с максимальным b_j	то же, что и в «равенство»
$\alpha = 1$	отбор одного признака с максимальным b_j	то же, что и в «равенство»

Связь метода QPFS и Lasso-модели для линейной регрессии

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \tau\|\mathbf{w}\|_1 \to \min_{\mathbf{w}}$$

$$\frac{1}{2}\mathbf{w}^{\mathsf{T}}\tilde{\mathbf{Q}}\mathbf{w} - \tilde{\mathbf{b}}^{\mathsf{T}}\mathbf{w} \to \min_{\mathbf{w}}$$
$$\text{s.t. } \|\mathbf{w}\|_1 = \eta.$$

Если:

1. Признаки и целевой вектор нормированны: $\mathbf{x}_j^{\mathsf{T}}\mathbf{x}_j = 1$, $\mathbf{y}^{\mathsf{T}}\mathbf{y} = 1$

2. Попарные корреляции неотрицательны: $\mathbf{y}^{\mathsf{T}}\mathbf{x}_j \geq 0$, $\mathbf{x}_j^{\mathsf{T}}\mathbf{x}_l \geq 0$

3. Истинный вектор весов неотрицателен: $w^* \geq 0$,

то эти задачи тождественны!

Вычислительный эксперимент

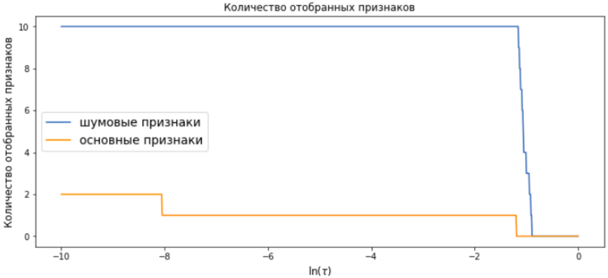
Цель

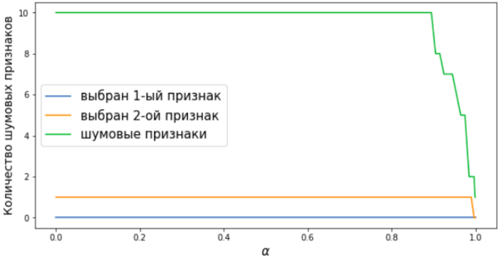
Выявить недостатки метода QPFS.

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



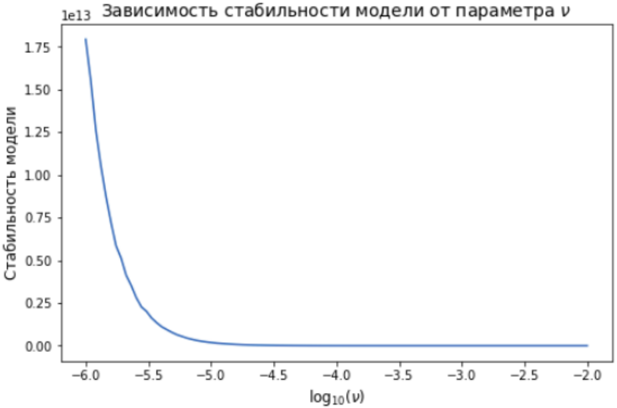


При этом, $y = \frac{x_2 - x_1}{\varepsilon}$, т.е. не выполнено одно из условий экв-ности QPFS и Lasso.

Стабильность модели

$$\|y - Xw\|_2^2 + \tau\lambda_{max}(A)/\lambda_{min}(A) \to \min_{w},$$

где $A = X(w)^T X(w)$ и $X(w)$ — сужение матрицы признаков на выбранные признаки.



Здесь $y \sim \mathcal{N}(0,1)$, $\varepsilon_i \sim \mathcal{N}(0,1)$, а $x_i = y + \mu \cdot \varepsilon_i$, где $\mu = 0.001$.

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

Заключение

Результаты

- проанализирован метод QPFS отбора признаков при различных постановках задач оптимизации
- показана эквивалентность QPFS и Lasso при определенных условиях
- теоретически и экспериментально подтверждены существенные недостатки метода QPFS
- предложены подходы для улучшения метода QPFS на основе комбинации признаков

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