Прогнозирование функции ошибки алгоритма выбора признаков

Аминов Тимур

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов) Группа 674, весна 2019

Цель работы

Предложить алгоритм выбора оптимального подмножества признаков

Проблема

В случае избыточного признакового пространства предсказатеьная модель машинного обучения является неустойчивой

Метод решения

Использование методов выпуклой оптимизации для получения оптимального подмножества признаков

Пусть $\mathbf{X} = [\chi_1, ..., \chi_n] \in \mathbb{R}^{m \times n}$ — заданная матрица, где $\chi_j \in \mathbb{R}^m$ — j-ый признак, $\mathbf{y} \in \mathbb{R}^m$ — значеня функии целевой функции

$$\mathbf{w}^* = arg \min_{\mathbf{w}} s(\mathbf{w}, \mathbf{X}_{\mathbf{tr}}, \mathbf{y}_{\mathbf{tr}}, \mathcal{A}),$$

s — функция ошибки Множество $\mathcal{A} \subseteq \{1, \dots, n\}$ — индикатор $\{0, 1\}$. Существует соответствие между множеством $\mathcal{A}$ и двоичным вектором $\mathbf{a} \in^n$:

$$\mathcal{A} = \{j : a_j = 1\}.$$

Типы признаков

- информативные —— существенно влияют на точность приближения целевого вектора
- шумовые — не влияют на точность приближения целевого вектора
- мультиколлинеарные — существует линейная зависимость между признаками, снижают устойчивость модели

$$(1-\alpha)\cdot\underbrace{\mathbf{z}^{\mathsf{T}}\mathbf{Q}\mathbf{z}}_{\mathrm{Sim}(\mathbf{X})}-\alpha\cdot\underbrace{\mathbf{b}^{\mathsf{T}}\mathbf{z}}_{\mathrm{Rel}(\mathbf{X},\boldsymbol{\nu})}\to\min_{\substack{\mathbf{z}\geq\mathbf{0}_n\\ \mathbf{1}_n^{\mathsf{T}}\mathbf{z}=1}}$$

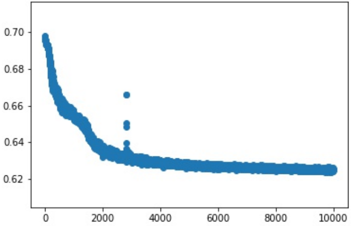
$$\alpha=\frac{\overline{\mathbf{Q}}}{\overline{\mathbf{Q}}+\overline{\mathbf{b}}},\quad \overline{\mathbf{Q}}=\mathrm{mean}(\mathbf{Q}),\ \overline{\mathbf{b}}=\mathrm{mean}(\mathbf{b})$$

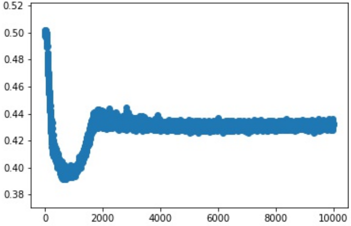
$$\mathbf{Q}=\left[|\mathrm{corr}(\chi_i,\chi_j)|\right]_{i,j=1}^{n},\quad \mathbf{b}=\left[|\mathrm{corr}(\chi_i,\boldsymbol{\nu})|\right]_{i=1}^{n}$$

$$\mathrm{corr}(\chi,\boldsymbol{\nu})=\frac{\sum_{i=1}^{m}(\chi_i-\overline{\chi})(\boldsymbol{\nu}_i-\overline{\boldsymbol{\nu}})}{\sqrt{\sum_{i=1}^{m}(\chi_i-\overline{\chi})^2\sum_{i=1}^{m}(\boldsymbol{\nu}_i-\overline{\boldsymbol{\nu}})^2}}$$

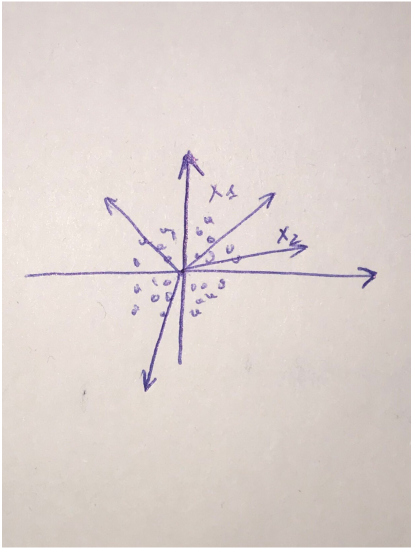
Решение

$$
\begin{tikzcd}
a \in \mathbb{B}^n \arrow[r, "e"', "\text{encoder}"] & u \in \mathbb{R}^h \arrow[r, "б"', "\text{decoder}"] \arrow[d, "P"', "\substack{\text{performance}\\ \text{prediction}}"] & a \in \mathbb{B}^n \\
& s \in \mathbb{R}^1 &
\end{tikzcd}
$$


0.70
0.68
0.66
0.64
0.62
0
2000
4000
6000
8000
10000


0.52
0.50
0.48
0.46
0.44
0.42
0.40
0.38
0
2000
4000
6000
8000
10000

Проблемы



Заключение

Проблема выбора признаков сведена к проблеме непрерывной оптимизации

Пока что сравнивать результаты невозможно