Предсказание качества для процедуры выбора признаков

Аминов Тимур

Московский физико-технический институт

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

Цель исследования

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

Постановка задачи

Пусть $\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n] \in \mathbb{R}^{m \times n}$ — заданная матрица, где $\boldsymbol{\chi}_j \in \mathbb{R}^m$ — j-ый признак, $\mathbf{y} \in \mathbb{R}^m$ — значеня функии целевой функции

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

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

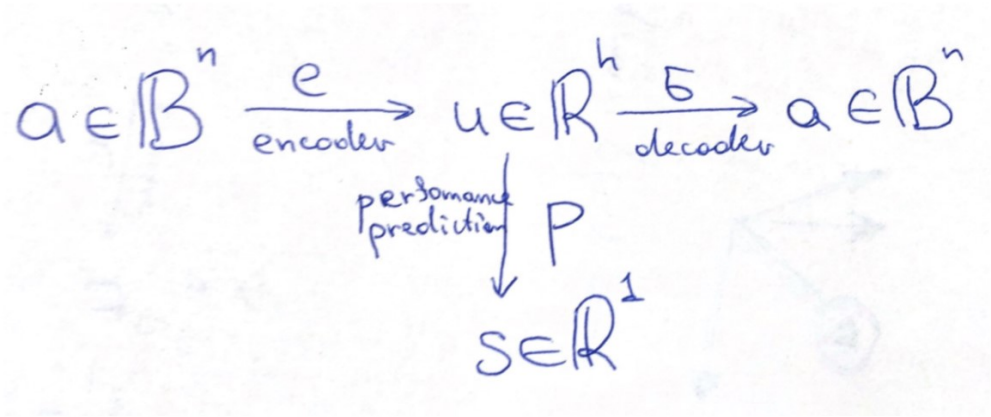
$$\mathcal{A} = \{j : a_j = 1\}.$$

Определение структуры выборки

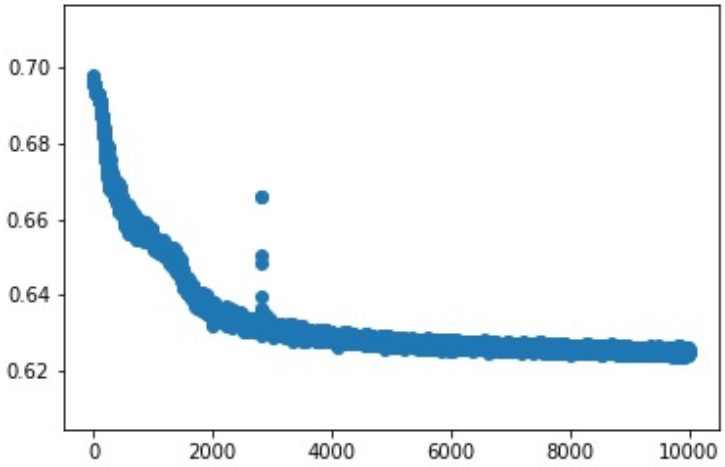
Типы признаков

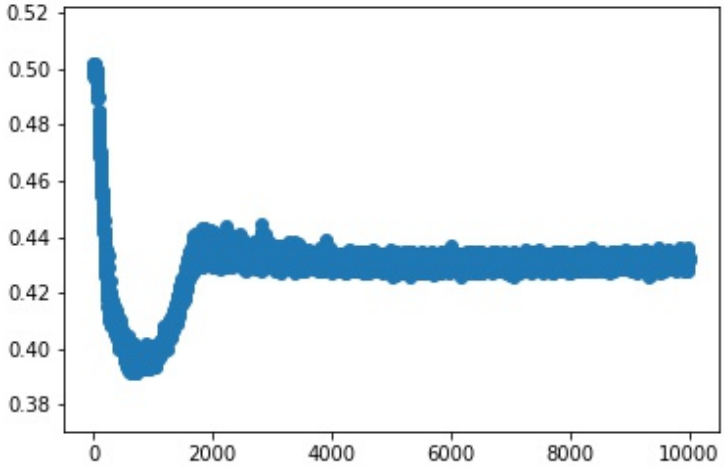
- информативные —— существенно влияют на точность приближения целевого вектора
- шумовые — не влияют на точность приближения целевого вектора
- мультиколлинеарные — существует линейная зависимость между признаками, снижают устойчивость модели

Решение


$a \in \mathbb{B}^n$
e
encoder
$u \in \mathbb{R}^h$
b
decoder
$a \in \mathbb{B}^n$
performance prediction
p
$s \in \mathbb{R}^1$

Вычислительный эксперимент





Заключение

Проблема выбора признаков сведена к проблеме непрерывной оптимизации

Пока что сравнивать результаты невозможно