

# Машинное обучение и анализ данных. Заметки

Подвойский А.О.

## Содержание

<b>1 Калибровка классификаторов</b>	<b>1</b>
1.1 Непараметрический метод гистограммной калибровки (Histogram Binning)	1
1.2 Непараметрический метод изотонической регрессии (Isotonic Regression)	2
1.3 Параметрическая калибровка Платта (Platt calibration)	2
1.4 Логистическая регрессия в пространстве логитов	2
1.5 Деревья калибровки	2
1.6 Температурное шкалирование (Temperature Scaling)	2
<b>Список литературы</b>	<b>2</b>

## 1. Калибровка классификаторов

Подробности в статье А. Дьяконова [«Проблема калибровки уверенности»](#).

Ниже описываются способы оценить качество калибровки алгоритма. Надо сравнить *уверенность* (confidence) и *долю верных ответов* (ассигасу) на тестовой выборке.

Если классификатор «хорошо откалиброван» и для большой группы объектов этот классификатор возвращает вероятность принадлежности к положительному классу 0.8, то среди этих объектов будет приблизительно 80% объектов, которые в действительности принадлежат положительному классу. То есть, если для группы точек данных общим числом 100 классификатор возвращает вероятность положительного класса 0.8, то приблизительно 80 точек на самом деле будут принадлежать положительному классу и доля верных ответов тогда составит 0.8.

### 1.1. Непараметрический метод гистограммной калибровки (Histogram Binning)

Изначально в методе использовались бины одинаковой ширины, но можно использовать и равномоощные бины.

Недостатки подхода:

- число бинов задается наперед,
- функция деформации не непрерывна,
- в «равноширинном варианте» в некоторых бинах может содержаться недостаточное число точек.

Метод был предложен Zadrozny B. и Elkan C. [Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers](#).

## 1.2. Непараметрический метод изотонической регрессии (Isotonic Regression)

Строится монотонно неубывающая функция деформации оценок алгоритма.

Метод был предложен Zadrozny B. и Elkan C. [Transforming classifier scores into accurate multiclass probability estimates](#).

Функция деформации по-прежнему не является непрерывной.

## 1.3. Параметрическая калибровка Платта (Platt calibration)

Изначально этот метод калибровки разрабатывался только для метода опорных векторов, оценки которого лежат на вещественной оси (по сути, это расстояния до оптимальной разделяющей классы прямой, взятые с нужным знаком). Считается, что этот метод не очень подходит для других моделей.

Предложен Platt J. [Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods](#).

## 1.4. Логистическая регрессия в пространстве логитов

## 1.5. Деревья калибровки

Стандартный алгоритм строит суперпозицию дерева решений на исходных признаках и логистических регрессий (каждая в своем листе) над оценками алгоритма:

- Построить на исходных признаках решающее дерево (не очень глубокое),
- В каждом листе – обучить логистическую регрессию на одном признаке,
- Подрезать дерево, минимизируя ошибку.

## 1.6. Температурное шкалирование (Temperature Scaling)

Этот метод относится к классу DL-методов калибровки, так как он был разработан именно для калибровки нейронных сетей. Метод представляет собой простое многомерное обобщение шкалирования Платта.

## Список литературы

1. Чакон С., Штрауб Б. Git для профессионального программиста. – СПб.: Питер, 2020. – 496 с.