

Методы анализа данных
Александр Широков ПМ-1701

Преподаватель:

ИВАХНЕНКО ДАРЬЯ АЛЕКСАНДРОВНА

Санкт-Петербург
2020 г., 7 семестр

Список литературы

[1]

Содержание

1	01.09.2020	2
1.1	Задача обучения по предедентам	2
1.2	Типы задач	2
1.3	08.09.2020	3
1.3.1	Задача бинарной классификации	3
1.3.2	Задача многоклассовой классификации	4

1 01.09.2020

1.1 Задача обучения по прецедентам

Пусть X - множество объектов, а Y - множество ответов. $y : X \rightarrow Y$ - неизвестная зависимость.

Дано: $\{x_1, \dots, x_l\} \subset X$ - обучающая выборка, а $y_i = y(x_i), i = 1, \dots, l$ - известные ответы.

Требуется найти $a : X \rightarrow Y$ - алгоритм, решающую функцию, приближающую y на всем множестве X .

1.2 Типы задач

Задачи восстановления регрессии:

- $Y = \mathbb{R}$ - вся числовая ось:
 - определение температуры воздуха метеорологического поля
 - оценка влияния факторов потребления
- $Y \in [0; +\infty)$:
 - задачи медицинской диагностики: прогнозирование ожидаемого время действия препарата
 - задачи кредитного скоринга: определение величины кредитного лимита
 - определение расхода топлива по техническим характеристикам
- $Y \in [0, 1, \dots, +\infty)$ - счетная целевая переменная

Задача классификации:

- $Y = \{-1, +1\}$ - классификация на два класса:
 - задачи кредитного скоринга: решение о выдаче кредита
 - предсказание оттока клиентов
- $Y = \{1, \dots, K\}$ - классификация на K непересекающихся классов:
 - задачи медицинской диагностики: определение диагноза
 - распознавание символов
 - определение жанра
- $Y = \{0, 1\}^K$ - на K классов, которые могут пересекаться:
 - определение ключевых слов для оптимизации поиска
 - определение присутствующих на фото объектов

Типы признаков

- $D_j = \{0, 1\}$ - бинарный признак f_j :
 - пол
 - является ли..?
- $|D_j| < \infty$ - номинальный признак f_j :
 - город
 - цвет
- $|D_j| < \infty$, D_j - упорядочено - порядковый признак f_j :
 - уровень холестерина (ниже нормы, норма, выше нормы)
- $D_j = \mathbf{R}$ - количественный признак f_j :
 - длина и ширина объекта

1.3 08.09.2020

1.3.1 Задача бинарной классификации

Выведем функцию связи через биномиальное распределение через задачу классификации:

$$y \in \{0, 1\}$$

$$g^{-1}(p) = \bar{\theta}^T \bar{X}$$

$$f(y) = e^{\frac{y\alpha - c(\alpha)}{\varphi} + h(y, \varphi)}$$

Плотность биномиального распределения:

$$f(y) = p^y (1-p)^{1-y} = \exp(y \log p + (1-y) \log(1-p)) = \exp(y \log \frac{p}{1-p} + \log(1-p)) \equiv c(\alpha)$$

$$\alpha = \log \frac{p}{1-p} = g(p) = \bar{\theta}^T \bar{X}$$

$$\frac{p}{1-p} = e^{\bar{\theta}^T \bar{X}}$$

$$p = \sigma(\bar{\theta}^T \bar{X}) = \frac{1}{1 + e^{-\bar{\theta}^T \bar{X}}}$$

Осталось получить функционал качества для задачи классификации. Функция называется ЛОГИСТИЧЕСКОЙ СИГМОИДОЙ. Данная функция преобразует линейную комбинацию в интервал $[0, 1]$. Дальнейшее значение целевой переменной мы будем предсказывать в качестве:

$$\hat{y} = \sigma(\bar{\theta}^T \bar{X})$$

Функционал качества найдём через метод максимального правдоподобия:

$$p(x, y, p) = \prod_{i=1}^l p_i^{y_i} (1 - p_i)^{1-y_i} \rightarrow \max_{\theta}$$

$$\sum_{i=1}^l y_i \log p_i + (1 - y_i) \log(1 - p_i) \rightarrow \max_{\theta}$$

Функция потерь называется LOGLOSS:

$$\text{LogLoss} = L(x_i) = y_i \log p_i - (1 - y_i) \log p_i \rightarrow \min, p_i = \sigma(x_i)$$

Если мы правильно предсказываем отношение принадлежности класса к 1, то функция потерь будет равна нулю, если правильно предсказываем 0 правильно, то тоже 0, а если 1 - правильный ответ, а 0 - нет, то ошибка будет $+\infty$, и ошибку ограничивают значениями 100, чтобы ошибка не уходила далеко.

$$p = \sigma(\bar{\theta}^T \bar{X}) \in [0, 1]$$

$$L = -y \log p - (1 - y) \log(1 - p)$$

1.3.2 Задача многоклассовой классификации

Определение 1.3.1. CATEGORICAL DISTRIBUTION: $y \in \{1, 2, \dots, K\}$

$$f(y) = \prod_{i=1}^K p_i^{y_i}$$

Выведем функцию связи через CATEGORICAL DISTRIBUTION:

$$f(y) = \prod_{i=1}^K p_i^{y_i} = \exp(\log \prod_{i=1}^K p_i^{y_i}) = \exp(\sum_{i=1}^K y_i \log p_i)$$

Рассмотрим определенный y_i :

$$f(y) = \exp(y_1 \log p_1 + y_2 \log p_2 + \dots + y_K \log p_K)$$

$$\sum y_i = 1$$

$$\sum p_i = 1$$

Так как y_i имеет принадлежность определенному классу и может быть равен только одной единице в векторе:

$$\{0, 0, 1, 0\}$$

а сумма вероятностей по определению. Тогда:

$$\begin{aligned} f(y) &= \exp(y_1 \log p_1 + \dots + \left(1 - \sum_i^{K-1} y_i\right) \log p_K) = \\ &= \exp\left(y_1 \log \frac{p_1}{p_K} + y_2 \log \frac{p_2}{p_K} + \dots + \log p_K\right) \end{aligned}$$

X для всех одинаковый, а y разные. Для определения y_i нам необхо-

димому определить α_i :

$$\alpha_i = \log \frac{p_i}{p_k} = \bar{\theta}^T \bar{X}$$

$$\frac{p_i}{p_k} = e^{\bar{\theta}^T \bar{X}}$$

$$\sum_i^K e^{\bar{\theta}_i^T \bar{X}} = \frac{\sum_i p_i}{p_k} = \frac{1}{p_k}$$

$$p_k = \frac{1}{\sum_i^K e^{\bar{\theta}_i^T \bar{X}}}$$

$$\frac{p_i}{p_k} = e^{\bar{\theta}_i^T \bar{X}}$$

$$p_i = \frac{e^{\bar{\theta}_i^T \bar{X}}}{\sum_{i=1}^K e^{\bar{\theta}_i^T \bar{X}}}$$

Данная функция называется функцией SOFTMAX.

Функционал качества мы можем легко получить:

$$L(x, y, p) = \prod_{i=1}^l \prod_{j=1}^K p_{ij}^{y_{ij}} \rightarrow \max$$

где i - i -ый объект, а j - принадлежность j -му классу.

Тогда функция потерь для каждого класса:

$$-\sum_{j=1}^K y_j \log p_j \rightarrow \min$$

и данная функция называется КРОСС-ЭНТРОПИЕЙ.