Работа с категориальными признаками

Максим Урьев

НИС Машинное обучение

2016

Нумерация значений

- Каждый i-ый категориальный признак произвольно пронумеровываем натуральными числами $\{1, \ldots, n_i\}$
- **Используем bagging:** обучаем слабые алгоритмы со случайно пронумерованными признаками и в качестве финального предсказания берем усредненный ответ

Нумерация значений

Для решающего дерева, если искать разбиение

- по критерию Джини или энтропийному критерию

$$\frac{1}{N_m(1)} \sum_{x_k \in R_m(1)} [y_k = +1] \leq \ldots \leq \frac{1}{N_m(n_i)} \sum_{x_k \in R_m(n_i)} [y_k = +1]$$

- в соответствии MSE-функционалом

$$\frac{1}{N_m(1)} \sum_{x_k \in R_m(1)} y_k \leq \ldots \leq \frac{1}{N_m(n_i)} \sum_{x_k \in R_m(n_i)} y_k$$

Кодирование относительно вещественного признака

Дана симметричная функция φ (например mean, max).
Рассмотрим вещественный признак s и категориальный $f = \{f^1, \dots f^{n_f}\}$.

$$I_k = \{s_i \mid f_i = f^k\}$$

$$f^k \to \varphi(I_k)$$

One Hot Encoding

Значение категориального признака f можно заменить бинарным вектором длины n_f

$$F = \left\| f_{ij} \right\|_{l \times n_f}, f_{ij} = I[f_i = f^j]$$

$$f \to F$$

Аналогично можно кодировать конъюкции исходных признаков

Сингулярное разложение SVD

Пусть $Z-$ вещественная матрица порядка $m \times n$.

Определение

Неотрицательное вещественное число σ называется **сингулярным числом** матрицы $Z \Longleftrightarrow$
$\exists u, v : \|u\| = 1, \|v\| = 1$ и $Zv = \sigma u, Z^T u = \sigma v$

Такие векторы u и v называются, соответственно, **левым сингулярным вектором** и **правым сингулярным вектором**, соответствующим сингулярному числу σ.

Сингулярное разложение SVD

Определение

Сингулярным разложением матрицы Z является разложение следующего вида:

$$Z = U\Sigma V^T$$

где $U_{m\times m}$ и $V_{n\times n}$ — это две ортогональные матрицы, состоящие из левых и правых сингулярных векторов соответственно,
а $\Sigma_{m\times n} = diag(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$, где r — ранг матрицы Z, а λ_i — её сингулярные числа (не теряя общности $\lambda_1 \geq \ldots \geq \lambda_r$).

Сингулярное разложение SVD

$$Z_k = \sum_{i=1}^{k} \lambda_i u_i v^i$$

Теорема Эккарта-Янга

Для $\forall$ матрицы T ранга k выполнено
$\|Z - T\|_2 \geq \|Z - Z_k\|_2 = \lambda_{k+1}$

Следствие

$Z \approx U_{m \times k} \Sigma_{k \times k} V^T_{k \times n}$ — усеченное сингулярное разложение.

Сингулярное разложение SVD

$$F \approx U\Sigma V^T$$

- Линейные комбинации столбцов матрицы U достаточно точно приближают столбцы исходной матрицы.

Если изначально известны контрольные объекты

$$F \to U$$

Если же изначально контрольные объекты не известны

$$F \to FV\Sigma^{-1}$$
$$F' \to F'V\Sigma^{-1}$$

Счетчики

Значение категориального признака f можно заменить вещественным признаком g:

$$g^j = \frac{\sum_{i=1}^{l}[y_i = +1][f_i = f^j] + \Delta_f c}{\sum_{i=1}^{l}[f_i = f^j] + c}$$

c — коэффициент регуляризации

Счетчики

Обобщение: для i-ого объекта перейдем к паре признаков $(\varphi(g_i), \gamma_i)$, где $\varphi : \mathbf{R} \to \mathbf{R}$,

$$\gamma_i = \begin{cases} 1, f_i \in \{f^1, \dots f^{n_f}\} \\ 0, f_i \notin \{f^1, \dots f^{n_f}\} \end{cases}$$

При этом можно положить $\Delta_f = 0$

На практике обычно используют $\varphi(x) = x^k$

Счетчики

Что еще попробовать:

- обнулять редкие категории
- объединять редкие категории в одну
- вычислять несколько счетчиков для разных c
- счетчики для конъюкций признаков

Частоты

Можно хранить частоты значения определенного признака

$$g^j = \frac{\sum_{i=1}^{l} [f_i = f^j]}{l}$$

Также можно хранить частоты конъюкций

Методы, основанные на тензорных разложениях

Рассмотрим пару категориальных признаков f, g и матрицу $Z = \|z_{ij}\|_{n_f \times n_g}$

$$z_{f^{t_f}, g^{t_g}} = \frac{\sum_{i=1}^{l} [f_i = f^{t_f}][g_i = g^{t_g}] y_i}{\sum_{i=1}^{l} [f_i = f^{t_f}][g_i = g^{t_g}]}$$

Если знаменатель равен 0, то считаем значение неопределенным.
Хотим восстановить все неопределенные элементы.

Методы, основанные на тензорных разложениях

Будем искать разложение $Z = UV$, где $U = \|u_{ij}\|_{n_f \times k}$, $V = \|v_{ij}\|_{k \times n_g}$, минимизируя функционал

$$J = \sum_{t=1}^{l} e_t^2 + \lambda_f \sum_{s=1}^{k} \sum_{i=1}^{n_f} u_{is}^2 + \lambda_g \sum_{s=1}^{k} \sum_{j=1}^{n_g} v_{sj}^2$$

$$e_t = \sum_{s=1}^{k} u_{f^{t_f},s} \cdot v_{s,g^{t_g}} - z_{f^{t_f},g^{t_g}}$$

Методы, основанные на тензорных разложениях

Вместо Z_{ij} можно брать латентные вектора u_i и v^j.

Методы, основанные на тензорных разложениях

Аналогично, когда n категориальных признаков $f_1, \ldots, f_n$, будем восстанавливать значения многомерной матрицы $n_1 \times \ldots \times n_n$.
Будем искать матрицы $U(r) = \|u_{ij}^r\|_{n_r \times k}, r \in 1, 2, \ldots, n$

$$J = \sum_{t=1}^{l} e_t^2 + \sum_{r=1}^{n} \lambda_r \|U(r)\|_2$$

$$e_t = \sum_{s=1}^{k} \prod_{r=1}^{n} u_{f_{t_r},s}^r - z_{f^{t f_1},\ldots,f^{t f_n}}$$

Методы, основанные на близости

Можно вычеслять оценки принадлежности классам следующим образом:

$$\Gamma_y(\bar{f}) = \frac{1}{N_y} \sum_{\Omega \in \Omega^*} w_\Omega \sum_{i:y_i=y} B_\Omega(\bar{f}, \bar{f}_i)$$

Ω^* — система опорных множеств: подмножеств множества признаков
w_Ω — вес опорного множества Ω
B_Ω — функция близости, равная $\prod_{j \in \Omega}[f_j = f_{ij}]$

Методы, основанные на близости

- Ω^* — в простейшем случае конъюкции.
- существуют более сложные способы выбора опорного множества
- можно возводить функционал в некоторую степень d

Stacking

Кодируя категориальные признаки различными способами, получаем различные модели.
Применяем Стекинг над нашим набором моделей.

Литература

- *www.alexanderdyakonov.narod.ru/sw-factors-dyakonov.pdf*
- *www.machinelearning.ru/wiki/images/7/73/MOTP14_5.pdf*