Машинное обучение, ФКН ВШЭ
Семинар №20

1 Обучение метрик

В методе k ближайших соседей не так много параметров — число соседей, функция расстояния, ядро и его ширина. Можно выбирать метрику из числа известных — например, из евклидовой, манхэттенской и косинусной. Эти метрики фиксированы и никак не могут быть подстроены под особенности данных. Кажется, что обучение метрики под выборку могло бы увеличить число степеней свободы у метрических методов и позволить добиваться более высокого качества. Например, масштаб признаков может существенно влиять на их важность при вычислении расстояний.

Рассмотрим простой пример. Допустим, решается задача определения пола человека по двум признакам: росту (в сантиметрах, принимает значения примерно от 150 до 200) и уровню экспрессии гена SRY (безразмерная величина от нуля до единицы; у мужчин ближе к единице, у женщин ближе к нулю). Обучающая выборка состоит из двух объектов: $x_1 = (180, 0.2)$, девочка и $x_2 = (173, 0.9)$, мальчик. Требуется классифицировать новый объект $u = (178, 0.85)$. Воспользуемся классификатором одного ближайшего соседа. Евклидовы расстояния от u до объектов обучения равны $\rho(u, x_1) \approx 2.1$ и $\rho(u, x_2) \approx 5$. Мы признаем новый объект девочкой, хотя это не так — высокий уровень экспрессии гена SRY позволяет с уверенностью сказать, что это мальчик. Из-за сильных различий в масштабе признаков уровень экспрессии практически не учитывается при классификации, что совершенно неправильно.

Удобнее всего обучать метрику через линейные преобразования признаков:

$$\rho(x, z) = \|Ax - Az\|^2 = (x - z)^T A^T A(x - z),$$

где $A \in \mathbb{R}^{n \times d}$ — матрица, которую можно подбирать. По сути, обучение линейного преобразования равносильно настройке параметра Σ в метрике Махалонобиса:

$$\rho(x, z) = (x - z)^T \Sigma^{-1}(x - z),$$

если положить $A = \Sigma^{-1/2}$. Нелинейные методы часто сводятся к обучению линейных в новом признаковом пространстве (т.е. $\rho(x, z) = \|A\varphi(x) - A\varphi(z)\|^2$) либо путём ядрового перехода в линейном методе [1].

Мы разберём два подхода к обучению расстояния Махалонобиса.

§1.1 Neighbourhood Components Analysis

Метод NCA [2] выбирает метрику так, чтобы для каждого объекта ближайшими оказывались объекты его же класса. Рассмотрим объект x_i и рассмотрим следующий

эксперимент: мы выбираем из оставшейся выборки случайный объект x_j и относим x_i к классу y_j. Зададим вероятности через расстояния между объектами:

$$p_{ij} = \begin{cases} \frac{\exp(-\|Ax_i - Ax_j\|^2)}{\sum_{k \neq i} \exp(-\|Ax_i - Ax_k\|^2)}, & i \neq j \\ 0, & i = j \end{cases}$$

Можно вычислить вероятность того, что объект x_i будет отнесён к правильному классу. Если обозначить через $C_i = \{j \mid y_i = y_j\}$ множество индексов объектов того же класса, то данная вероятность равна

$$p_i = \sum_{j \in C_i} p_{ij}.$$

Будем максимизировать матожидание количества верно классифицированных объектов:

$$Q(A) = \sum_{i=1}^{\ell} p_i \to \max_A$$

Этот функционал можно продифференцировать по A:

$$\frac{\partial Q}{\partial A} = 2A \sum_i \left(p_i \sum_k p_{ik}(x_i - x_k)(x_i - x_k)^T - \sum_{j \in C_i} p_{ij}(x_i - x_j)(x_i - x_j)^T \right).$$

Далее матрицу A можно обучать любым градиентным методом.

Отметим, что метод NCA можно использовать и для ускорения поиска ближайших соседей. Если взять матрицу $A \in \mathbb{R}^{n \times d}$ с небольшой первой размерностью n, то она будет переводить объекты в компактные представления, евклидова метрика на которых позволяет хорошо отделять классы друг от друга.

§1.2 Large margin nearest neighbor

Метод LMNN [3] пытается обучить метрику так, чтобы k ближайших соседей каждого объекта относились к нужному классу, а объекты из других классов отделялись с большим отступом. Попытаемся ввести соответствующий функционал.

Определим для каждого объекта x_i набор из k целевых соседей — объектов, расстояние до которых должно оказаться минимальным. В простейшем варианте это могут быть ближайшие k объектов из этого же класса, но можно выбирать их и иначе. Введём индикатор $\eta_{ij} \in \{0, 1\}$, который равен единице, если объект x_j является целевым соседом для x_i.

Выше мы поставили перед собой две цели: минимизировать расстояние до целевых соседей и максимизировать расстояние до объектов других классов. Суммарное расстояние до целевых соседей можно вычислить как

$$\sum_{i \neq j} \eta_{ij} \|Ax_i - Ax_j\|^2.$$

Для объектов других классов будем требовать, чтобы расстояние до них хотя бы на единицу превосходило расстояния до целевых соседей:

$$\sum_{i=1}^{\ell}\sum_{j\neq i}\sum_{\substack{m\neq i\\ m\neq j}}\eta_{ij}[y_m \neq y_i]\max(0, 1 + \|Ax_i - Ax_j\|^2 - \|Ax_i - Ax_m\|^2).$$

Суммируя эти два выражения, получим итоговый функционал:

$$\sum_{i\neq j}\eta_{ij}\|Ax_i - Ax_j\|^2 + \\ + C\sum_{i=1}^{\ell}\sum_{j\neq i}\sum_{\substack{m\neq i\\ m\neq j}}\eta_{ij}[y_m \neq y_i]\max(0, 1 + \|Ax_i - Ax_j\|^2 - \|Ax_i - Ax_m\|^2) \to \min_A$$

Данную задачу можно свести к стандартной задаче с линейным функционалом и ограничениями на неотрицательную определённость матрицы и решена стандартными солверами.

Список литературы

[1] *Kulis, B.* (2012). Metric Learning: A Survey. // Foundations and Trends in Machine Learning.

[2] *Goldberger J., Hinton G., Roweis S., Salakhutdinov R.* (2005). Neighbourhood Components Analysis. // Advances in Neural Information Processing Systems.

[3] *Weinberger, K. Q.; Blitzer J. C.; Saul L. K.* (2006). Distance Metric Learning for Large Margin Nearest Neighbor Classification. // Advances in Neural Information Processing Systems.