Машинное обучение, ФКН ВШЭ
Семинар №21

1 Метод k ближайших соседей

§1.1 Описание алгоритма

Пусть дана обучающая выборка $X = \{(x_i, y_i)\}_{i=1}^{\ell}$ и функция расстояния $\rho : \mathbb{X} \times \times \mathbb{X} \to [0, \infty)$, и требуется классифицировать новый объект $u \in \mathbb{X}$. Расположим объекты обучающей выборки X в порядке возрастания расстояний до u:

$$\rho(u, x_u^{(1)}) \leqslant \rho(u, x_u^{(2)}) \leqslant \ldots \leqslant \rho(u, x_u^{(\ell)}),$$

где через $x_u^{(i)}$ обозначается i-й сосед объекта u. Алгоритм k *ближайших соседей* относит объект u к тому классу, представителей которого окажется больше всего среди k его ближайших соседей:

$$a(u; X, k) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} w_i [y_u^{(i)} = y].$$

Параметр k обычно настраивается с помощью кросс-валидации.

В классическом методе k ближайших соседей все объекты имеют единичные веса: $w_i = 1$. Такой подход, однако, не является самым разумным. Допустим, что $k = 3$, $\rho(u, x_u^{(1)}) = 1$, $\rho(u, x_u^{(2)}) = 1.5$, $\rho(u, x_u^{(3)}) = 100$. Ясно, что третий сосед находится слишком далеко и не должен оказывать сильное влияние на ответ. Эта идея реализуется с помощью весов, обратно пропорциональных расстоянию:

$$w_i = K\left(\rho(u, x_u^{(i)})\right),$$

где $K(x)$ — любая монотонно убывающая функция.

С помощью метода k ближайших соседей можно решать и задачи регрессии. Для этого нужно усреднить значения целевой функции на соседях с весами:

$$a(u; X, k) = \frac{\sum_{i=1}^{k} w_i y_u^{(i)}}{\sum_{i=1}^{k} w_i},$$

где $y_u^{(i)}$ — значение целевой переменной на объекте $x_u^{(i)}$.

§1.2 Особенности и проблемы метода

Разберем особенности и проблемы метода k ближайших соседей, возникающие при использовании евклидовой метрики в качестве функции расстояния:

$$\rho(x,z) = \left(\sum_{j=1}^{d} |x_j - z_j|^2\right)^{1/2}.$$

1.2.1 Шумовые признаки

Задача 1.1. *Рассмотрим задачу с одним признаком и двумя объектами обучающей выборки: $x_1 = 0.1$, $x_2 = 0.5$. Первый объект относится к первому классу, второй — ко второму. Добавим к объектам шумовой признак, распределенный равномерно на отрезке $[0,1]$. Пусть требуется классифицировать новый объект $u = (0,0)$. Какова вероятность, что после добавления шума второй объект окажется к нему ближе, чем первый?*

Решение. Задача сводится к вычислению вероятности $\mathbb{P}(0.5^2 + \xi_2^2 \leqslant 0.1^2 + \xi_1^2)$, где ξ_1 и ξ_2 — независимые случайные величины, распределенные равномерно на $[0,1]$. Вычислим ее:

$$\mathbb{P}(0.5^2 + \xi_2^2 \leqslant 0.1^2 + \xi_1^2) = \mathbb{P}(\xi_1^2 \geqslant 0.24 + \xi_2^2) =$$
$$= \int_0^{\sqrt{0.76}} \int_{\sqrt{z_2^2+0.24}}^{1} dz_1 dz_2 = \int_0^{\sqrt{0.76}} \left(1 - \sqrt{z_2^2 + 0.24}\right) dz_2 \approx 0.275.$$

■

Таким образом, шумовые признаки могут оказать сильное влияние на метрику. Обнаружить шумовые признаки можно, удаляя поочередно все признаки и смотря на ошибку на тестовой выборке или ошибку кросс-валидации.

1.2.2 «Проклятие размерности»

Пусть объекты выборки — это точки, равномерно распределенные в d-мерном кубе $[0,1]^d$. Рассмотрим выборку, состоящую из 5000 объектов, и применим алгоритм пяти ближайших соседей для классификации объекта u, находящегося в начале координат. Выясним, на сколько нужно отступить от этого объекта, чтобы с большой вероятностью встретить пять объектов выборки. Для этого построим подкуб $[0;\varepsilon]^d \subset [0;1]^d$, $\varepsilon \in (0,1)$, и положим его объём равным $\delta = \varepsilon^d$. Найдем такое значение δ, при котором в этот подкуб попадет как минимум пять объектов выборки с вероятностью 0.95.

Задача 1.2. *Запишите выражение для δ.*

Решение.

$$\min\left\{\delta \,|\, \sum_{k=5}^{5000} \binom{5000}{k} \delta^k (1-\delta)^{5000-k} \geqslant 0.95\right\} \approx 0.0018.$$

■

Таким образом, для того, чтобы найти пять соседей объекта u, нужно по каждой координате отступить на $0.0018^{1/d}$. Уже при $d = 10$ получаем, что нужно отступить на 0.53, при $d = 100$ — на 0.94. Таким образом, при больших размерностях объекты становятся сильно удалены друг от друга, из-за чего классификация на основе сходства объектов может потерять смысл. В то же время отметим, что в рассмотренном примере признаки объектов представляли собой равномерный шум, тогда как в реальных задачах объекты могут иметь осмысленные распределения, позволяющие построение модели классификации даже при больших размерностях.

Настоящая же проблема, связанная с «проклятием размерности», заключается в невозможности эффективного поиска ближайших соседей для заданной точки. Было показано, что сложность всех популярных методов решения этой задачи становится линейной по размеру выборки по мере роста размерности [1]. В то же время можно добиться эффективного поиска, если решать задачу поиска ближайших соседей приближенно (locality-sensitive hashing).

§1.3 Примеры функций расстояния

1.3.1 Метрика Минковского

Метрика Минковского определяется как:

$$\rho_p(x, z) = \left(\sum_{j=1}^{d} |x_j - z_j|^p \right)^{1/p}$$

для $p \geqslant 1$. При $p \in (0, 1)$ данная функция метрикой не является, но все равно может использоваться как мера расстояния.

Частными случаями данной метрики являются:

- Евклидова метрика ($p = 2$). Задает расстояние как длину отрезка прямой, соединяющей заданные точки.
- Манхэттенское расстояние ($p = 1$). Минимальная длина пути из x в z при условии, что можно двигаться только параллельно осям координат.
- Метрика Чебышева ($p = \infty$), выбирающая наибольшее из расстояний между векторами по каждой координате:

$$\rho_\infty(x, z) = \max_{j=1,\dots,d} |x_j - z_j|.$$

- «Считающее» расстояние ($p = 0$), равное числу координат, по которым векторы x и z различаются:

$$\rho_0(x, z) = \sum_{j=1}^{d} [x_j \neq z_j].$$

Отметим, что по мере увеличения параметра p метрика слабее штрафует небольшие различия между векторами и сильнее штрафует значительные различия.

В случае, если признаки неравнозначны, используют взвешенное расстояние:

$$\rho_p(x, z; w) = \left(\sum_{j=1}^{d} w_j |x_j - z_j|^p \right)^{1/p}, \quad w_j \geqslant 0.$$

Задача 1.3. *Рассмотрим функцию $f(x) = \rho_2(x, 0; w)$. Что представляют из себя линии уровня такой функции?*

Решение. Распишем квадрат функции $f(x)$ (форма линий уровня от этого не изменится):

$$f^2(x) = \sum_{j=1}^{d} w_j x_j^2.$$

Сделаем замену $x_j = \frac{x'_j}{\sqrt{w_j}}$:

$$f^2(x') = \sum_{j=1}^{d} {x'_j}^2.$$

В новых координатах линии уровня функции расстояния представляют собой окружности с центром в нуле. Сама же замена представляет собой растяжение вдоль каждой из координат, поэтому в исходных координатах линия уровня являются эллипсами, длины полуосей которых пропорциональны $\sqrt{w_j}$.

■

Вывод: благодаря весам линии уровня можно сделать эллипсами с осями, параллельными осям координат. Это может быть полезно, если признаки имеют разные масштабы — благодаря весам автоматически будет сделана нормировка.

Веса можно брать, например, равными корреляции между признаком и целевым вектором:

$$w_j = \left| \frac{\sum_{i=1}^{\ell} x_{ij} y_i}{\left(\sum_{i=1}^{\ell} x_{ij}^2 \right)^{1/2} \left(\sum_{i=1}^{\ell} y_i^2 \right)^{1/2}} \right|.$$

Однако, лучше всего настраивать веса под обучающую выборку с помощью покоординатного спуска или другого метода оптимизации.

1.3.2 Расстояние Махалонобиса

Расстояние Махалонобиса определяется следующим образом:

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

где S — симметричная положительно определенная матрица.

Задача 1.4. *Что представляют из себя линии уровня функции $f(x) = \rho(x, 0)$?*

Решение. Поскольку матрица S — симметричная, то из её собственных векторов можно составить ортонормированный базис. Рассмотрим матрицу Q, столбцами которой являются элементы данного базиса. Заметим, что она является ортогональной (в силу ортонормированности базиса), т.е. $Q^TQ = I$, $Q^{-1} = Q^T$, а потому выполняются следующие соотношения:

$$SQ = Q\Lambda \Rightarrow \Lambda = Q^{-1}SQ,$$

где Λ — диагональная матрица, в которой записаны соответствующие собственные значения матрицы S.

Распишем квадрат функции $f(x)$, сделав замену $x' = Q^Tx$:

$$f^2(x) = \rho^2(x,0) = x^TS^{-1}x = x'^TQ^TS^{-1}Qx' = x'^T\left(Q^{-1}SQ\right)^{-1}x' = \\ = x'^T\Lambda^{-1}x' = \sum_{j=1}^{d}\frac{x'^2}{\lambda_j}.$$

Получаем, что линии уровня в новых координатах представляют собой эллипсы с осями, параллельными осям координат, причем длины полуосей равны корням из собственных значений матрицы S. При этом замена $x' = Q^Tx$ соответствует такому повороту осей координат, что координатные оси совпадают со столбцами матрицы Q. Таким образом, расстояние Махалонобиса позволяет получить линии уровня в виде произвольно ориентированных эллипсов.

■

Матрицу S можно настраивать либо по кросс-валидации, либо брать равной выборочной ковариационной матрице: $\hat{S} = \frac{1}{n-1}X^TX$.

1.3.3 Косинусная мера

Пусть заданы векторы x и z. Известно, что их скалярное произведение и косинус угла θ между ними связаны следующим соотношением:

$$\langle x, z\rangle = \|x\|\|z\|\cos\theta.$$

Соответственно, косинусное расстояние определяется как

$$\rho_{\cos}(x,y) = \arccos\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right) = \arccos\left(\frac{\sum_{j=1}^{d}x_jy_j}{\left(\sum_{j=1}^{d}x_j^2\right)^{1/2}\left(\sum_{j=1}^{d}y_j^2\right)^{1/2}}\right).$$

Косинусная мера часто используется для измерения схожести между текстами. Каждый документ описывается вектором, каждая компонента которого соответствует слову из словаря. Компонента равна единице, если соответствующее слово встречается в тексте, и нулю в противном случае. Тогда косинус между двумя векторами будет тем больше, чем больше слов встречаются в этих двух документах одновременно.

Один из плюсов косинусной меры состоит в том, что в ней производится нормировка на длины векторов. Благодаря этому она не зависит, например, от размеров сравниваемых текстов, измеряя лишь объем их схожести.

1.3.4 Расстояние Жаккара

Выше мы рассматривали различные функции расстояния для случая, когда объекты обучающей выборки являются вещественными векторами. Если же объектами являются множества (например, каждый объект — это текст, представленный множеством слов), то их сходство можно измерять с помощью *расстояния Жаккара*:

$$\rho_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

Задача 1.5. *Пусть все множества являются подмножествами некоторого конечного упорядоченного множества $U = \{u_1, \ldots, u_N\}$. Тогда любое множество A можно представить в виде бинарного вектора длины N, в котором единица в i-й позиции стоит тогда и только тогда, когда $u_i \in A$. Запишите формулу для расстояния Жаккара, исходя из таких обозначений, и сравните ее с формулой для косинусной меры.*

Решение. Пусть X и Y — два множества, $(x_j)_{j=1}^N$ и $(y_j)_{j=1}^N$ — их векторные представления. Тогда мощность их пересечения можно записать следующим образом:

$$|X \cap Y| = \sum_{j=1}^{N} x_j y_j = \langle X, Y \rangle,$$

а мощность их объединения как

$$\begin{aligned} |X \cup Y| &= \sum_{j=1}^{N} x_j + \sum_{j=1}^{N} y_j - \sum_{j=1}^{N} x_j y_j = \\ &= \sum_{j=1}^{N} x_j^2 + \sum_{j=1}^{N} y_j^2 - \sum_{j=1}^{N} x_j y_j = \\ &= \|X\|^2 + \|Y\|^2 - \langle X, Y \rangle. \end{aligned}$$

Тогда:

$$\rho_J(X, Y) = 1 - \frac{\langle X, Y \rangle}{\|X\|^2 + \|Y\|^2 - \langle X, Y \rangle}.$$

■

1.3.5 Редакторское расстояние

Для измерения сходства между двумя строками (например, последовательностями ДНК) можно использовать *редакторское расстояние*, которое равно минимальному числу вставок и удалений символов, с помощью которых можно преобразовать первую строку ко второй. В зависимости от специфики задачи можно также разрешать замены, перестановки соседних символов и прочие операции.

1.3.6 Функции расстояния на категориальных признаках

Категориальные признаки не имеют никакой явной структуры, и поэтому достаточно сложно ввести на них разумное расстояние. Как правило, ограничиваются сравнением их значений: если у двух объектов одинаковые значения категориального признака, то расстояние равно нулю, если разные — единице. Тем не менее, существуют определенные соображения по поводу того, как измерять сходство для таких признаков.

Будем считать, что метрика записывается как взвешенная сумма расстояний по отдельным признакам с некоторыми весами:

$$\rho(x, z) = \sum_{j=1}^{d} w_j \rho_j(x_j, z_j).$$

Способы измерения расстояния для вещественных признаков обсуждались выше. Обсудим некоторые варианты для категориальных признаков. Введем следующие обозначения:

1. $f_j(m) = \sum_{i=1}^{\ell} [x_{ij} = m]$ — количество раз, которое j-й признак принимает значение m на обучающей выборке;
2. $p_j(m) = \frac{f_j(m)}{\ell}$ — частота категории m на обучающей выборке;
3. $p_j^{(2)}(m) = \frac{f_j(m)(f_j(m)-1)}{\ell(\ell-1)}$ — оценка вероятности того, что у двух случайно выбранных различных объектов из обучающей выборки значения признака будут равны m.

Тогда расстояние на категориальных признаках можно задавать, например, одним из следующих способов:

1. Индикатор совпадения:
$$\rho_j(x_j, z_j) = [x_j \neq z_j]$$
2. Сглаженный индикатор совпадения. Чем выше частота у значения признака, тем больше расстояние (если оба человека живут в Москве, то эта информация не очень важна, поскольку вероятность такого совпадения высока; если оба человека живут в Снежинске, то это важная информация, так событие является достаточно редким):
$$\rho_j(x_j, z_j) = [x_j \neq z_j] + [x_j = z_j] \sum_{q: p_j(q) \leqslant p_j(x_j)} p_j^{(2)}(q)$$
3. Чем более частые значения оказались при несовпадении, тем больше расстояние (если оба человека из разных, но очень маленьких городов, то можно считать их похожими; если один человек из Москвы, а второй — из Питера, то они сильно отличаются):
$$\rho_j(x_j, z_j) = [x_j \neq z_j] \log f_j(x_j) \log f_j(z_j)$$
(обратите внимание, что для борьбы с численными проблемами имеет смысл добавлять единицу под логарифмом: $\log(f_j(x_j) + 1)$, $\log(f_j(z_j) + 1)$)

Более подробный обзор функций расстояния на категориальных признаках можно найти в работе [2].

Список литературы

[1] *Weber, R., Schek, H. J., Blott, S.* (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. // Proceedings of the 24th VLDB Conference, New York C, 194–205.

[2] *Boriah, S., Chandola, V., Kumar, V.* (2008). Similarity measures for categorical data: A comparative evaluation. // In Proceedings of the 2008 SIAM International Conference on Data Mining (pp. 243–254).