Семинары по метрическим методам

Евгений Соколов

3 октября 2015 г.

1 Метод k ближайших соседей

§1.1 Описание алгоритма

Пусть дана обучающая выборка $X = (x_i, y_i)_{i=1}^{\ell} \subset \mathbb{X}$ и функция расстояния $\rho : \mathbb{X} \times \mathbb{X} \to [0, \infty)$, 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:

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

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:

$$a(u, X^\ell, k) = \frac{\sum_{i=1}^{k} w_{(i)} y_{(i)}}{\sum_{i=1}^{k} w_{(i)}}.$$

§1.2 Случай евклидовой метрики

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

$$\rho(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^2\right)^{1/2}.$$

1.2.1 Границы классов

Диаграмма Вороного, соответствующая выборке X^ℓ — это такое разбиение пространства на области, что каждая область состоит из точек, для которых одна и та же точка из выборки является ближайшей. Более формально, диаграмма Вороного для выборки X^ℓ состоит из ℓ областей $R_1, \dots, R_\ell$, определяемых как

$$R_i = \{x \in \mathbb{R}^d \mid \rho(x, x_i) < \rho(x, x_j), j \neq i\}.$$

Очевидно, что при использовании классификатора ближайшего соседа ($k = 1$) граница между классами является подмножеством границ между такими областями.

Опр. 1.1. *Область $R \subset \mathbb{R}^d$ называется выпуклым многогранником, если она является пересечением конечного числа полупространств:*

$$R = \bigcap_{i=1}^{n} \{x \in \mathbb{R}^d \mid \langle w_i, x \rangle < 0\}.$$

Задача 1.1. *Показать, что множество точек, для которых ближайшим соседом из выборки является заданный объект x_i, представляет собой выпуклый многогранник.*

Решение. Условие того, что x_i является ближайшей точкой выборки к u, записывается как

$$\sum_{p=1}^{d} (x_{ip} - u_p)^2 < \sum_{p=1}^{d} (x_{jp} - u_p)^2, \quad j \neq i.$$

Распишем его:

$$\sum_{p=1}^{d} (x_{ip}^2 - 2x_{ip}u_p + u_p^2) < \sum_{p=1}^{d} (x_{jp}^2 - 2x_{jp}u_p + u_p^2), \quad j \neq i;$$

$$\sum_{p=1}^{d} (x_{ip}^2 - 2x_{ip}u_p) < \sum_{p=1}^{d} (x_{jp}^2 - 2x_{jp}u_p), \quad j \neq i;$$

$$\sum_{p=1}^{d} (x_{ip}^2 - x_{jp}^2 + 2(x_{jp} - x_{ip})u_p) < 0, \quad j \neq i;$$

$$2\sum_{p=1}^{d} (x_{jp} - x_{ip})u_p + \sum_{p=1}^{d} (x_{ip}^2 - x_{jp}^2) < 0, \quad j \neq i.$$

Мы получили набор линейных относительно u неравенств, каждое из которых задает полупространство. Их пересечение является множеством точек, для которых x_i является ближайшим соседом, и является выпуклым многогранником по определению.

■

Классификатор одного ближайшего соседа является крайне чувствительным к шумовым объектам и выбросам, и граница между классами может оказаться очень сложной. По мере увеличения k граница сглаживается за счет «усреднения» по нескольким объектам.

1.2.2 Нормализация признаков

Умножим один из признаков (например, первый) на константу C. Евклидово расстояние примет следующий вид:

$$\rho_2(x, y) = \sqrt{C(x_1 - y_1)^2 + \sum_{i=2}^{d} (x_i - y_i)^2}.$$

Таким образом, различие по первому признаку будет считаться в C раз более значимым, чем различия по всем остальным признакам. При этом расположение объектов относительно друг друга не изменилось — изменился лишь масштаб!

Рассмотрим простой пример чувствительности метода ближайшего соседа к масштабу признаков. 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. 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.

Чтобы избежать подобных проблем, признаки следует нормировать. Это можно делать, например, следующими способами:

- Нормировка на единичную дисперсию:

$$\tilde{x}^j = \frac{x^j - \bar{x}^j}{\sigma(x^j)}.$$

- Нормировка на отрезок $[0, 1]$:

$$\tilde{x}^j = \frac{x^j - \min(x^j)}{\max(x^j) - \min(x^j)}.$$

Здесь x^j — это вектор, составленный из j-х признаков всех объектов. Иными словами, это j-й столбец матрицы «объекты-признаки».

1.2.3 Шумовые признаки

Задача 1.2. *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:

$$\mathbb{P}(0.5^2 + \xi_2^2 \leqslant 0.1^2 + \xi_1^2) = \mathbb{P}(\xi_1^2 \geqslant 0.24 + \xi_2^2) = \\ = \int_0^{\sqrt{0.76}} \int_{\sqrt{x_2^2+0.24}}^1 dx_1 dx_2 = \int_0^{\sqrt{0.76}} \left(1 - \sqrt{x_2^2 + 0.24}\right) dx_2 \approx 0.275.$$

■

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. Более сложные методы отбора информативных признаков будут разобраны позже на лекциях.

1.2.4 «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 единичного куба, включающий в себя начало координат и имеющий объем δ, и найдем такое значение δ, при котором в этот подкуб попадет как минимум пять объектов выборки с вероятностью 0.95.

Задача 1.3. *Запишите выражение для δ.*

Решение.

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

■

Минимальное значение δ, удовлетворяющее этому уравнению, приблизительно равно приблизительно 0.0018. Отсюда находим, что 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. Этот вопрос будет разобран более подробно на следующем занятии.

§1.3 Примеры функций расстояния

1.3.1 Метрика Минковского

Метрика Минковского определяется как:

$$\rho_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

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

$$\rho_\infty(x, y) = \max_{i=1,\dots,d} |x_i - y_i|.$$

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

$$\rho_0(x, y) = \sum_{i=1}^{d} [x_i \neq y_i].$$

Отметим, что считающее расстояние не является нормой.

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:

$$\rho_p(x, y; w) = \left(\sum_{i=1}^{d} w_i |x_i - y_i|^p \right)^{1/p}, \quad w_i \geqslant 0.$$

Задача 1.4. *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):

$$f^2(x) = \sum_{i=1}^{d} w_i x_i^2.$$

Сделаем замену $x_i = \frac{x'_i}{\sqrt{w_i}}$:

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

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 $\sqrt{w_i}$.

■

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:

$$w_i = \left| \frac{\sum_{j=1}^{\ell} x_{ji} y_j}{\left(\sum_{j=1}^{\ell} x_{ji}^2\right)^{1/2} \left(\sum_{j=1}^{\ell} y_j^2\right)^{1/2}} \right|.$$

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:

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

где S — симметричная положительно определенная матрица.

Напомним, что собственным вектором матрицы S называется такой вектор x, что $Sx = \lambda x$ для некоторого λ. Если матрица S симметричная, то из ее собственных векторов можно составить ортонормированный базис. Сформировав матрицу Q из таких собственных векторов (один столбец — один вектор), получим следующие два соотношения:

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

где Λ — диагональная матрица, в которой записаны собственные значения матрицы S. При этом матрица Q является ортогональной: $Q^T Q = I$, $Q^T = Q^{-1}$.

Изучим, как ведет себя расстояние Махалонобиса, если с его помощью сравнивать начало координат с произвольной точкой x. Сделаем для этого замену $x' = Q^T x$. Она соответствует такому повороту осей координат, что координатные оси совпадают со столбцами матрицы Q (то есть с собственными векторами).

Выясним теперь, как выглядят линии уровня метрики в новых координатах:

$$\rho^2(x, 0) = x^T S^{-1} x = {x'}^T Q^T S^{-1} Q x' = {x'}^T (Q^{-1} S Q)^{-1} x' =$$

$$= {x'}^T \Lambda^{-1} x' = \sum_{i=1}^{d} \frac{{x'_i}^2}{\lambda_i}.$$

Получаем, что линии уровня представляют собой эллипсы с осями, параллельными осям координат, причем длины полуосей равны корням из собственных значений $\sqrt{\lambda_i}$. 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: $\hat{S} = \frac{1}{n-1} X^T X$.

Задача 1.5. *Покажите, что выборочная ковариационная матрица является неотрицательно определенной.*

Решение. Напомним, что матрица A называется неотрицательно определенной, если $\langle Az, z \rangle \geqslant 0$ для всех z.

Покажем неотрицательную определенность выборочной ковариационной матрицы:

$$\langle X^T X z, z \rangle = \left(X^T X z \right)^T z = z^T X^T X z = (Xz)^T (Xz) = \|Xz\|^2 \geqslant 0.$$

■

1.3.3 Косинусная мера

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

$$\langle x, y \rangle = \|x\| \|y\| \cos(\theta).$$

Соответственно, косинусное расстояние определяется как

$$\rho_{\cos}(x, y) = \arccos\left(\frac{\langle x, y \rangle}{\|x\| \|y\|} \right) = \arccos\left(\frac{\sum_{i=1}^{d} x_i y_i}{\left(\sum_{i=1}^{d} x_i^2 \right)^{1/2} \left(\sum_{i=1}^{d} y_i^2 \right)^{1/2}} \right).$$

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 Джаккарда

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 Джаккарда*:

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

Задача 1.6. *Пусть все множества являются подмножествами некоторого конечного упорядоченного множества $U = \{u_1, \dots, u_N\}$. 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 Джаккарда, исходя из таких обозначений, и сравните ее с формулой для косинусной меры.*

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

$$|X \cap Y| = \sum_{i=1}^N x_i y_i = \langle X, Y \rangle,$$

а мощность их объединения как

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

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. Обозначим через $f_j(x)$ количество раз, которое j-й признак принимает значение x на обучающей выборке; через $p_j(x) = f_j(x)/\ell$ — частоту категории x; через $p_j^2(x) = f_j(x)(f_j(x) - 1)/\ell(\ell - 1)$ — оценку вероятности того, что у двух случайно выбранных объектов значения признака будут равны x.

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):

$$\rho_j(x_j, z_j) = [x_j \neq z_j] \log f_j(x_j) \log f_j(z_j)$$

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

Существует и другой подход к измерению расстояния между категориальными признаками, основанный на преобразовании их в вещественные. Так, в задачах бинарной классификации крайне эффективной является замена категорий на их условные вероятности получения одного из классов [1]:

$$x_j \to p(y(x) = 1 \mid f_j(x) = x_j) = \frac{\sum_{i=1}^{\ell} [f_j(x_i) = x_j][y(x_i) = 1]}{\sum_{i=1}^{\ell} [f_j(x_i) = x_j]}$$

[1] http://blogs.technet.com/b/machinelearning/archive/2015/02/17/big-learning-made-easy-with-counts.aspx

§1.4 Заключение

Основной проблемой метода ближайших соседей является то, что его обучение заключается лишь в запоминании выборки (и, возможно, построении структуры данных для эффективного поиска в ней). При этом не происходит никакой настройки параметров с целью максимизации качества, из-за чего метод не может приспособиться к ненормированным или шумовым признакам. В то же время метод работает с объектами лишь через функцию расстояния, что позволяет использовать его для работы с самыми разнообразными данными (векторами, множествами, строками, распределениями и т.д.).

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).