ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Машинное обучение

Лекция 7. Метод k-ближайших соседей, наивный байесовский классификатор

Автор: Рустам Азимов

Санкт-Петербург, 2023г.

Метрические классификаторы

- **Метрический классификатор (similarity-based classifier)** — алгоритм классификации, основанный на вычислении оценок сходства между объектами
- Простейшим метрическим классификатором является **метод ближайших соседей**
- Для формализации понятия сходства вводится функция расстояния между объектами $\rho(x, x')$
- В общем случае, жёсткого требования, чтобы эта функция была метрикой не предъявляется (в частности, неравенство треугольника может нарушаться)

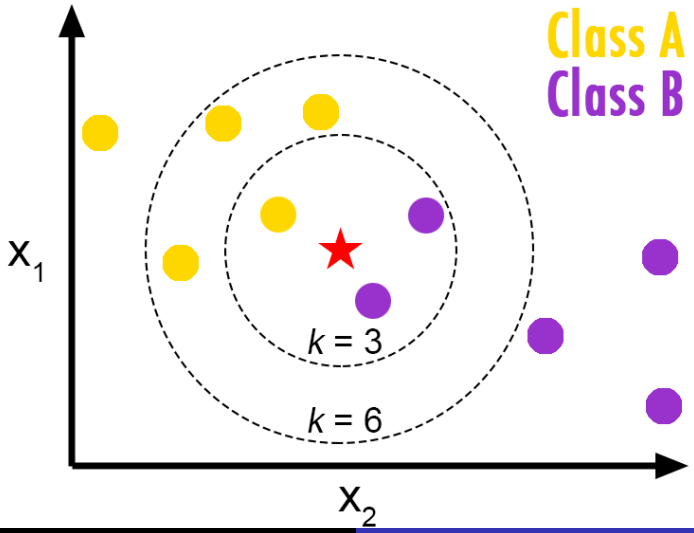
Метод ближайших соседей

- Основывается на оценивании сходства объектов обучающей выборки
- Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки
  - **метод ближайшего соседа** ($k = 1$)
  - **метод k-ближайших соседей (k-nearest neighbors, k-NN)**
  - **взвешанный метод k-ближайших соседей**
- Методы существенно опираются на **гипотезу компактности**: если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных

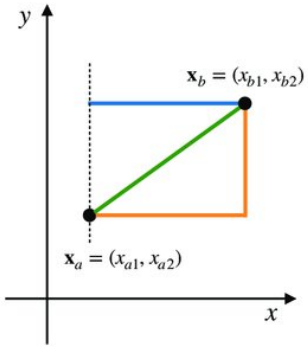
Метод k-ближайших соседей

- Во всех случаях как такового обучения нету, оно сводится к сохранению обучающей выборки
- Должна быть выбрана метрика ρ для вычисления расстояний (вычисления сходства) между объектами, а также число k
- Для предсказания целевого признака для нового объекта x производятся следующие шаги
  1. Вычисляются расстояния от x до всех объектов обучающей выборки
  2. Объекты обучающей выборки сортируются по возрастанию расстояний до x
  3. Выбираются k объектов с наименьшими расстояниями до x
  4. По этим k объектам вычисляется ответ на задачу предсказания
- Для классификации может быть выбран наиболее популярный класс (k лучше нечётное брать)
  - Можно выдавать вероятности
- Для регрессии можно взять среднее/медиану значений целевых признаков

Метод k-ближайших соседей



Метрики



$p = 2$ Euclidean distance

$$\|\mathbf{x}_a - \mathbf{x}_b\|_2 = (|x_{a1} - x_{b1}|^2 + |x_{a2} - x_{b2}|^2)^{\frac{1}{2}}$$

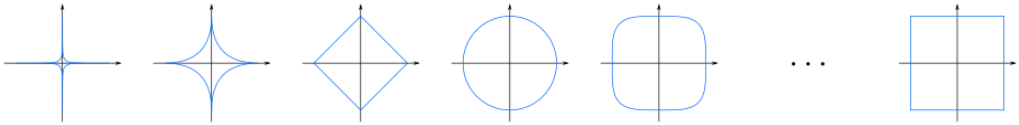
$p = 1$ Manhattan distance

$$\|\mathbf{x}_a - \mathbf{x}_b\|_M = |x_{a1} - x_{b1}| + |x_{a2} - x_{b2}|$$

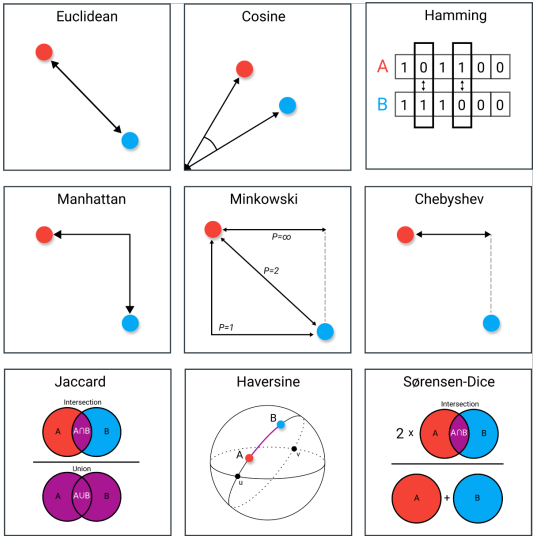
$p = \infty$ Chebyshev distance

$$\|\mathbf{x}_a - \mathbf{x}_b\|_\infty = \max\{|x_{a1} - x_{b1}|, |x_{a2} - x_{b2}|\}$$

Метрики Минковского



Ещё метрики


Euclidean
Cosine
Hamming
A 1 0 1 1 0 0
B 1 1 1 0 0 0
Manhattan
Minkowski
P=∞
P=2
P=1
Chebyshev
Jaccard
Intersection
A
A∩B
B
Union
A
A∪B
B
Haversine
B
A
v
u
Sørensen-Dice
Intersection
2 x
A
A∩B
B
A
+
B

Взвешанный k-NN

- При расчёта предсказания каждому соседу даётся определённый вес
- Например веса обратные квадрату расстояний $\frac{1}{d^2(x,x_i)}$
- Более близкие объекты обучающей выборки сильнее влияют на конечное предсказание

Применимость k-NN

- Существуют теоремы, что метод ближайших соседей идеальный метод классификаций в случае неограниченного датасета
- Но на практике применимость этого метода ограничена из-за вычислительных ограничений и **проклятия размерности**
- k-NN может служить как некоторый baseline или как конструктор **мета-признаков** (его предсказаний), которые пойдут на вход другим алгоритмам
- При проблемах с производительностью на огромных данных можно применять некоторый приближённый поиск ближайших соседей

Вероятностные классификаторы

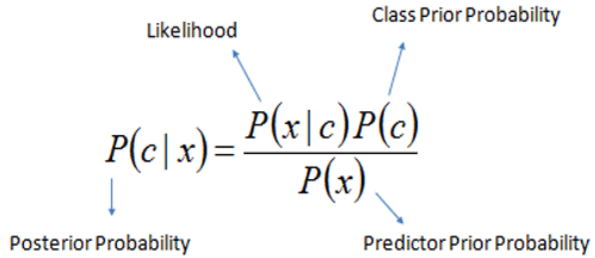
- Основываются на оценка вероятности принадлежать объекта к определённому классу
- Мы уже сталкивались с этим (логистическая регрессия)
- Ещё один пример — **наивный байесовкий классификатор (Naive Bayes classifier)**
- До недавнего времени повсеместно использовался для классификации спама в электронной почте

Naive Bayes

- **Наивный байесовский алгоритм** — это алгоритм классификации, основанный на **теореме Байеса** с допущением о независимости признаков
- Он предполагает, что наличие какого-либо признака в классе не связано с наличием какого-либо другого признака
- Например размер головы студента не зависит от его оценок
- Даже если имеется зависимость между признаками, они должны вносить **независимый** вклад в вероятность принадлежности классу
- Отсюда название — "наивный"

Naive Bayes

- Теорема Байеса позволяет рассчитать апостериорную вероятность $P(c|x)$ на основе $P(c)$, $P(x)$ и $P(x|c)$
- Наивная часть — в переходе от $P(X|c)$ к произведению $P(x_i|c)$



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Naive Bayes

- Здесь $P(c|x)$ — апостериорная вероятность данного класса c (т.е. данного значения целевой переменной) при данном значении признака x
- $P(c)$ — априорная вероятность данного класса
- $P(x|c)$ — правдоподобие, т.е. вероятность данного значения признака при данном классе
- $P(x)$ — априорная вероятность данного значения признака, не зависит от c и является константой
- Байесовский классификатор минимизирует ошибку принятия решений
- Для объекта X выбираем класс с максимальной $P(c|X)$ (убрали из вычислений все что не зависит от c)

Категориальные признаки: оценка априорных вероятностей классов

- Вероятность класса $P(c)$ оценивается по обучающей выборке как отношение количества объектов этого класса к объёму всей выборки:

$$P(c) = \frac{n_c}{n}$$

Категориальные признаки: оценка вероятностей признаков

- Вероятность признаков оценивается по обучающей выборке следующим образом:

$$P(x = i|c) = \frac{M_i(c) + \alpha}{\sum_{j=1}^{m}(M_j(c) + \alpha)}$$

- Где $M_i(c)$ — общее количество элементов класса c со значением признака $x = i$
- Для избежания нулевых значений добавляется $\alpha > 0$, например, $\alpha = 1$

Количественные признаки: одномерный непрерывный случай

- Эмпирическая оценка плотности:

$$p_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} [|x - x_i| < h]$$

- Где h — неотрицательный параметр, называемый шириной окна
- Локальная непараметрическая оценка Парзена-Розенблатта:

$$p_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

Количественные признаки: многомерный непрерывный случай

- У объектов m признаков
- Оценка плотности в точке $x = (\varepsilon_1, \ldots, \varepsilon_m)$:

$$p_h(x) = \frac{1}{2nh} \sum_{i=1}^{n} \prod_{j=1}^{m} \frac{1}{h_j} K(\frac{x - x_i}{h})$$

- В каждой точке многомерная плотность представляется в виде произведения одномерных плотностей

Применимость Naive Bayes

- Простотой в реализации и имеет низкие вычислительные затраты при обучении и классификации
- Подходит для случаев, когда в распоряжении малое количество данных
- В тех редких случаях, когда признаки действительно независимы (или почти независимы), наивный байесовский классификатор (почти) оптимален
- Основной его недостаток — относительно низкое качество классификации в большинстве реальных задач
- Может быть использован как baseline или в связке с другими алгоритмами (в композиции)

Дополнительные источники

- https://habr.com/ru/company/ods/blog/322534/
- machinelearning.ru
- scikit-learn.org
- kaggle