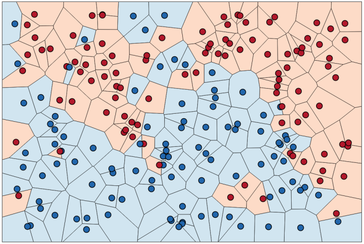
Метрические методы

Виктор Китов

v.v.kitov@yandex.ru



Содержание

Метод ближайших центроидов

- Рассмотрим обучающую выборку $(x_1, y_1), ... (x_N, y_N)$ с
    - N_1 представителями 1го класса
    - N_2 представителями 2го класса
    - и т.д.
- **Обучение:**
  Рассчитать центроиды для каждого класса $c = 1, 2, ... C$:

$$\mu_c = \frac{1}{N_1} \sum_{n=1}^{N} x_n \mathbb{I}[y_n = c]$$

- **Классификация:**
    1. Для каждого объекта x найти ближайший центроид:

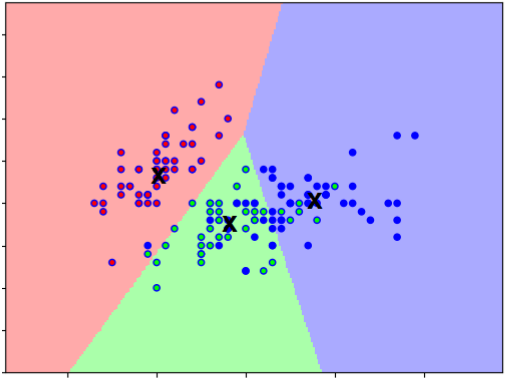
    $$c = \arg\min_i \rho(x, \mu_i)$$

    2. Назначить x самый распространенный класс в центроиде:

    $$\widehat{y}(x) = c$$

Демонстрация работы

Решающее правило для 3-х классового метода ближайших центроидов.



Вопросы

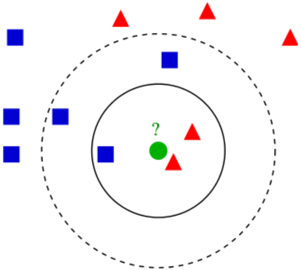
- Чему равны дискриминантные ф-ции $g_c(x)$ метода?
- Какова сложность:
  - обучения?
  - предсказания?
- Что собой представляют границы между классами?
- Применимы ли схожие принципы к регрессии? (рассмотрите кластеризацию)
- Подвержен ли метод "проклятию размерности"?

Содержание

Метод ближайших соседей

Классификация:

1. Найти K ближайших объектов в обучающей выборке к заданному x.
2. Сопоставить x самый частотный класс среди K ближайших объектов.



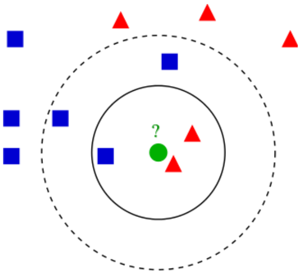
Метод ближайших соседей

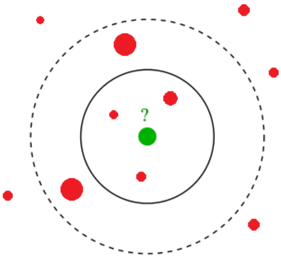
Классификация:

1. Найти K ближайших объектов в обучающей выборке к заданному x.
2. Сопоставить x самый частотный класс среди K ближайших объектов.

Регрессия:

1. Найти K ближайших объектов в обучающей выборке к заданному x.
2. Сопоставить x среднему отклику среди K ближайших объектов.

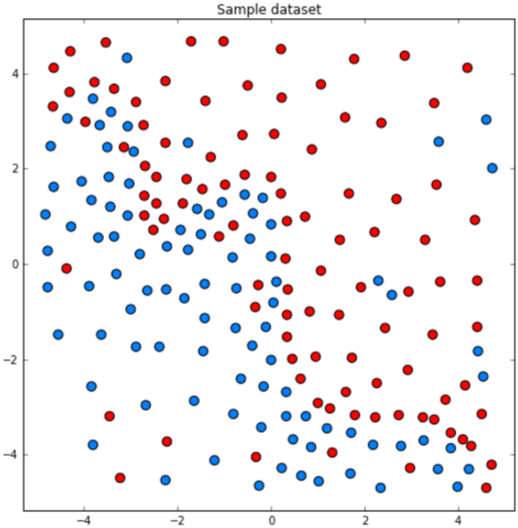




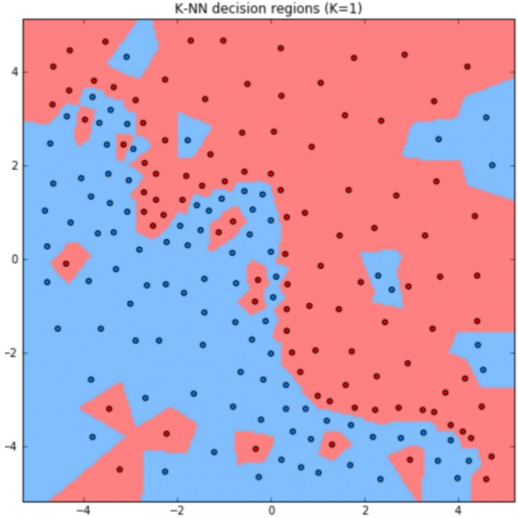
Комментарии

- Англ. *K-nearest neighbors*, сокращенно K-NN.
- $K = 1$: алгоритм ближайшего соседа.
- Базовое предположение: близким объектам соответствуют похожие отклики.
- Как будет работать алгоритм при $K = N$?
- Что вычислительно проще - обучить метод или применять его?

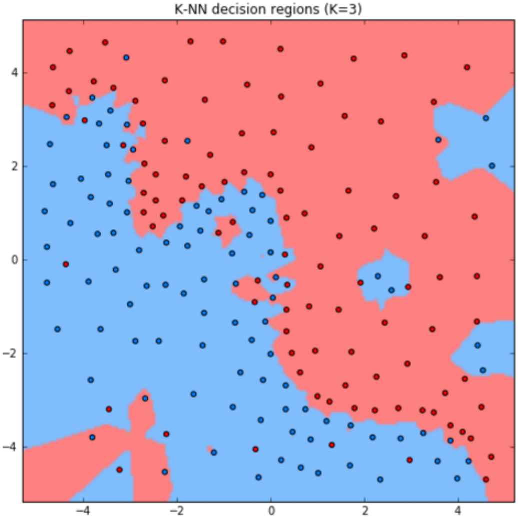
Демонстрационная выборка



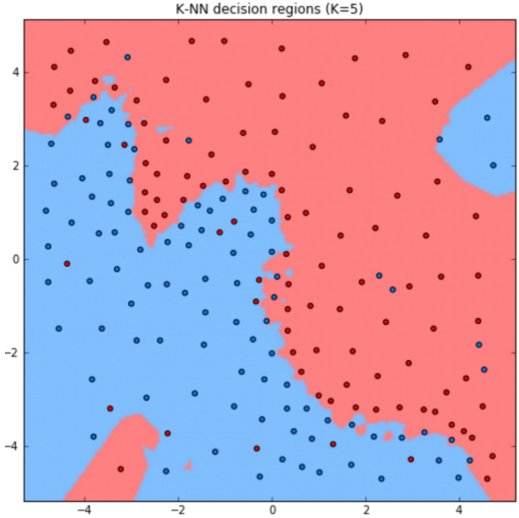
Пример: K-NN, K=1



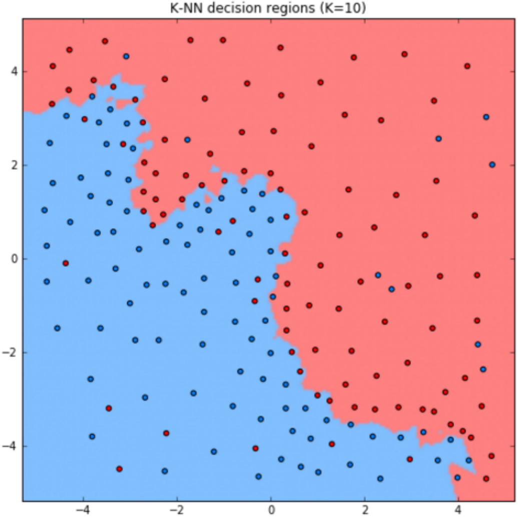
Пример: K-NN, K=3



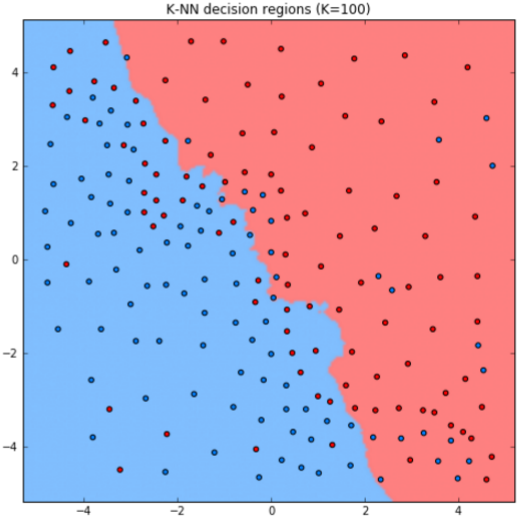
Пример: K-NN, K=5



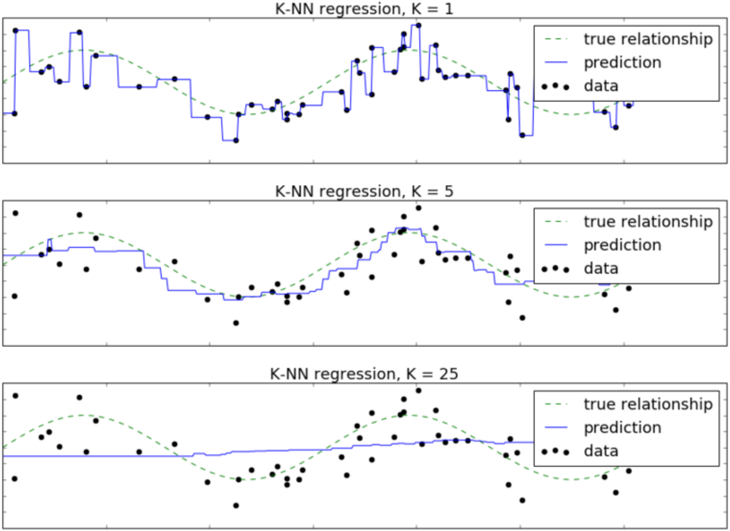
Пример: K-NN, K=10



Пример: K-NN, K=100



Пример: K-NN, регрессия $y = \sin x + \varepsilon$



Если два класса равноценно побеждают

Если два класса набирают одинаковый рейтинг, можно сопоставить класс:

Если два класса равноценно побеждают

Если два класса набирают одинаковый рейтинг, можно сопоставить класс:

- случайно
- более распространенный в обучающей выборке
- имеющий ближайшего представителя:
    - из ближайших среди K соседей
    - ближайшего среднего представителя среди K соседей
    - из самых удаленных среди K соседей (поощряем компактность распределения)

Параметры метода

- Параметры метода:
  - число ближайших соседей K
  - метрика расстояния $\rho(x, x')$
- Модификации:
  - возможный отказ, если прогноз неопределённый [1]
  - адаптивный выбор $K(x)$[2]

[1] Предложите критерий неопределенности прогноза для классификации и для регрессии.

[2] Предложите, как именно $K(x)$ мог бы меняться в зависимости от локальной плотности объектов.

Свойства K-NN

- **Достоинства:**
  - для прогноза нужна только степень близости между объектами, а не конкретные признаковые представления.
    - может применяться к объектам произвольно сложной структуры (тексты, графы, ...).
  - легко реализовать
  - интерпретируемый (прогноз по похожим известным случаям)
    - важно, например, в медицине.
  - не требует обучения (нужно только сохранить объекты)
    - может применяться в онлайн-сценариях
    - кросс-валидация может заменяться скользящим контролем (leave-one-out).
- **Недостатки:**
  - сложность прогноза $O(ND)$
  - точность снижается с $\uparrow D$ ("проклятие размерности")

Содержание

Зависимость от масштаба признаков

- Влияет ли масштабирование признаков на прогнозы K-NN?

Зависимость от масштаба признаков

- Влияет ли масштабирование признаков на прогнозы K-NN?
  - да, поэтому нужно нормализовывать их
- Единый масштаб => одинаковое влияние признаков
- Разный масштаб => различное влияние признаков.
  - Повышение масштаба увеличивает или уменьшает вклад признака в прогноз?

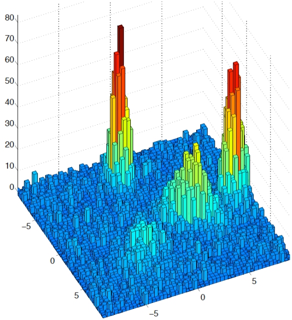
Распространенные нормализации признаков

Название	**Преобразование**	**Выходные свойства**
Стандартизация	$\frac{x_j - \mu_j}{\sigma_j}$	нулевое среднее и единичная дисперсия
Нормализация средним	$\frac{x_j - \mu_j}{\max(x_j) - \min(x_j)}$	нулевое среднее, с единичным диапазоном
Диапазонное шкалирование	$\frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}$	принадлежит интервалу $[0, 1]$

- Какой тип шкалирования более устойчив к выбросам?
- Какой тип шкалирования сохраняет свойство разреженности? (много нулевых значений)

Проклятие размерности: идея

- Проклятие размерности: $\uparrow D$ => близких точек становится мало.
    - за счет повышения объема пространства
- Пример: оценка гистограмм[3]



[3] С какой скоростью должно расти N, чтобы с точки зрения точности компенсировать увеличение D?

Проклятие размерности: обоснование

1. Предположим, точки распределены равномерно в $\mathbb{R}^D$.
2. Шар радиуса R имеет объем $V(R) = CR^D$, где $C = \frac{\pi^{D/2}}{\Gamma(D/2+1)}$ $\quad (2R,\ \pi R^2,\ \frac{4}{3}\pi R^3, ...)$.
3. Отношение объемов шаров радиуса $R - \varepsilon$ и R:

$$\frac{V(R-\varepsilon)}{V(R)} = \left(\frac{R-\varepsilon}{R}\right)^D \xrightarrow{D\to\infty} 0$$

   - объем уменьшается вокруг центра (даже с малым ε) и концентрируется на поверхности.
   - в K-NN: ближайшие соседи перестают быть близкими.
   - в гистограмме: сложнее набрать сгусток точек в заданной окрестности.

- Хорошие новости: в практических задачах объекты распределены неравномерно, на многообразиях меньшей размерности.

Содержание

Равномерный учет объектов

- Обозначим K ближайших соседей к точке x:

$$(x_1, y_1), (x_2, y_2), ...(x_K, y_K)$$

$$\rho(x, x_1) \le \rho(x, x_2) \le ... \le \rho(x, x_K)$$

- Регрессия:

$$\widehat{y}(x) = \frac{1}{K}\sum_{k=1}^{K} y_k$$

- Классификация:

$$g_c(x) = \sum_{k=1}^{K} \mathbb{I}[y_k = c], \quad c = 1, 2, ...C.$$

$$\widehat{y}(x) = \arg\max_c g_c(x)$$

Взвешенное голосование

- Взвешенная регрессия:

$$\widehat{y}(x) = \frac{\sum_{k=1}^{K} w(k,\ \rho(x, x_k))y_k}{\sum_{k=1}^{K} w(k,\ \rho(x, x_k))}$$

Взвешенное голосование

- Взвешенная регрессия:

$$\widehat{y}(x)=\frac{\sum_{k=1}^{K} w(k,\ \rho(x,x_k))y_k}{\sum_{k=1}^{K} w(k,\ \rho(x,x_k))}$$

- Взвешенная классификация:

$$g_c(x)=\sum_{k=1}^{K} w(k,\ \rho(x,x_k))\mathbb{I}[y_k=c],\quad c=1,2,...C.$$

$$\widehat{y}(x)=\arg\max_{c} g_c(x)$$

Популярные варианты весов

Веса, зависящие от ранга близости:

$$w_k = \alpha^k, \quad \alpha \in (0,1)$$

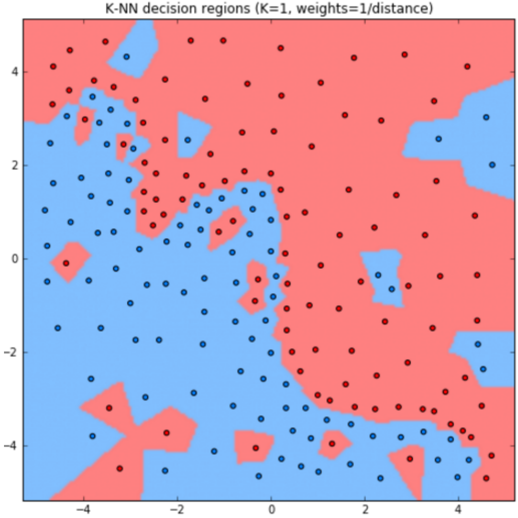
$$w_k = \frac{K+1-k}{K}$$

Веса, зависящие от расстояний до объектов:

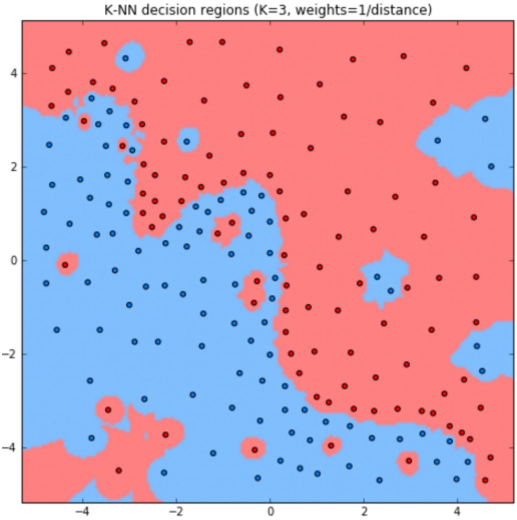
$$w_k = \begin{cases} \frac{\rho(z_K,x)-\rho(z_k,x)}{\rho(z_K,x)-\rho(z_1,x)}, & \rho(x_K,x) \neq \rho(x_1,x) \\ 1 & \rho(x_K,x) = \rho(x_1,x) \end{cases}$$

$$w_k = \frac{1}{\rho(x_k,x)}$$

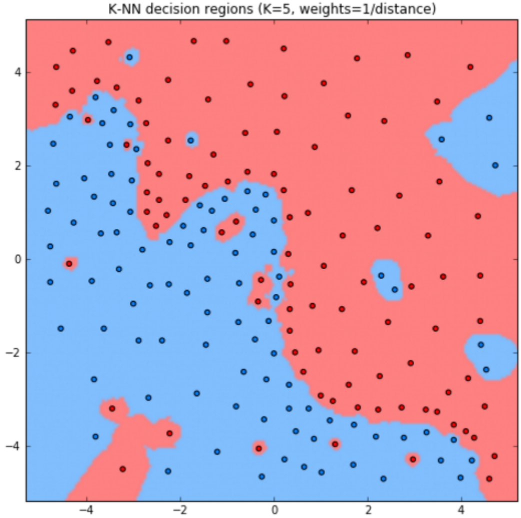
Пример: взвешенная классификация $K = 1$



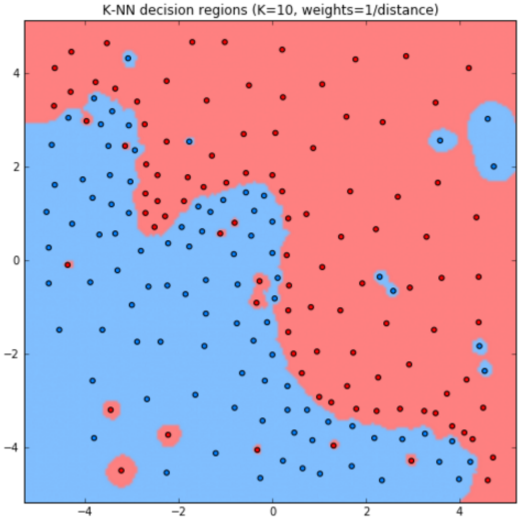
Пример: взвешенная классификация $K = 3$



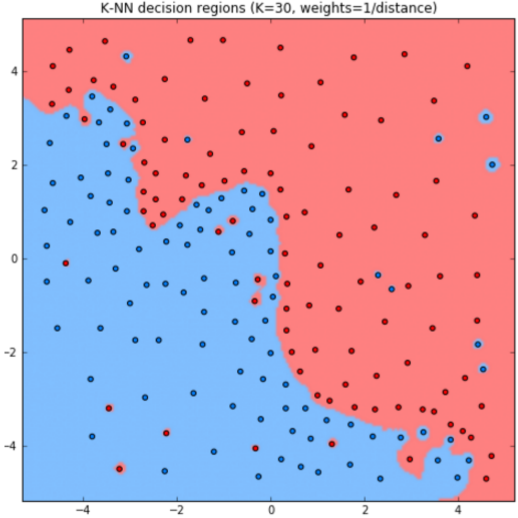
Пример: взвешенная классификация $K = 5$



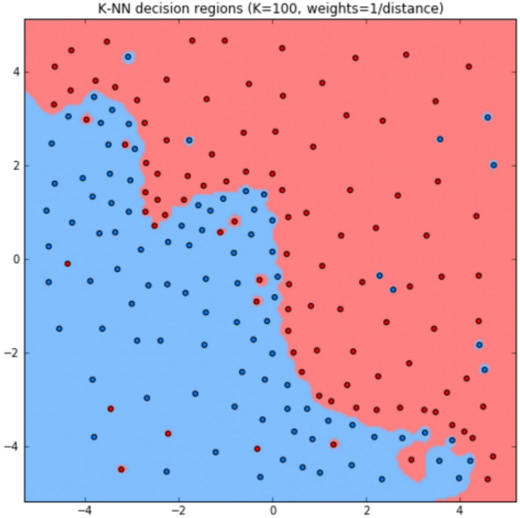
Пример: взвешенная классификация $K = 10$



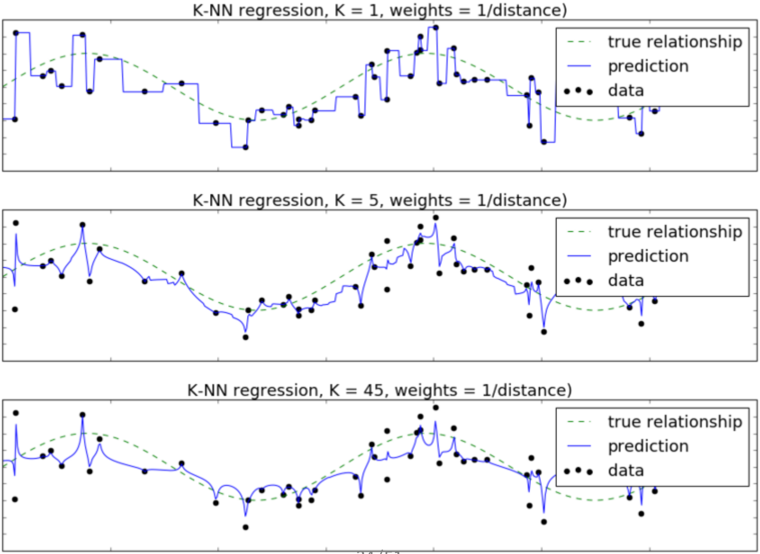
Пример: взвешенная классификация $K = 30$



Пример: взвешенная классификация $K = 100$



Пример: взвешенная регрессия K-NN



Содержание

Популярные функции расстояния[4]

Название	$\boldsymbol{\rho(x,z)}$
Евклидова	$\sqrt{\sum_{i=1}^{D}(x^i - z^i)^2}$
L_p	$\sqrt[p]{\sum_{i=1}^{D}(x^i - z^i)^p}$
L_∞	$\max_{i=1,2,\dots D} \lvert x^i - z^i \rvert$
L_1	$\sum_{i=1}^{D} \lvert x^i - z^i \rvert$
Канберра	$\frac{1}{D}\sum_{i=1}^{D}\frac{\lvert x^i - z^i\rvert}{\lvert x^i + z^i\rvert}$
Ланса-Уильямса	$\frac{\sum_{i=1}^{D}\lvert x^i - z^i\rvert}{\sum_{i=1}^{D}\lvert x^i + z^i\rvert}$

Часто определяют меру близости $S(x,z)$, тогда $\rho(x,z) = K(S(x,z)$ для $\downarrow K$, например

$$\rho(x,z) = 1 - S(x,z) \qquad \rho(x,z) = \frac{1}{S(x,z)}$$

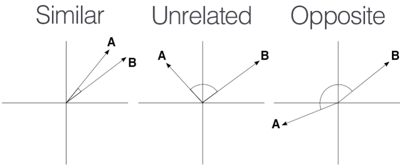
[4] Постройте единичные сферы по L_1, L_2, L_∞ метрикам.

Косинусная мера близости

- Косинусная мера близости*:* объекты близки, если угол между их векторами мал.

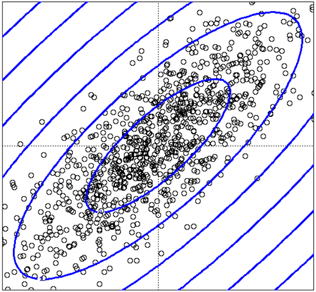
$$sim(x,z)=\frac{x^T z}{\|x\|\,\|z\|}=\frac{\sum_{i=1}^{D}x^i z^i}{\sqrt{\sum_{i=1}^{D}(x^i)^2}\sqrt{\sum_{i=1}^{D}(z^i)^2}}$$

  - $\langle x,z\rangle = x^T z = \|x\|\,\|z\|\,cos(\alpha)$, где α - угол между x и z.



  - метрика $\in [-1,1]$ и инвариантна к длинам $\|x\|,\|z\|$.
    - удобно для текстовых представлений в виде счетчиков слов.

Зависимые признаки



- Объекты вдоль оси $y = x$ более похожи, чем вдоль $y = -x$. Как это учесть?
- Посчитаем Евклидово расстояние, но для декоррелированных признаков.

Декоррелирующее преобразование

- $x \sim F(\mu, \Sigma)$, $\mu = \mathbb{E}[\mu]$, $\Sigma = cov(x, x)$, $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{DxD}$
- Декоррелирующее преобразование:

$$z = \Sigma^{-1/2}(x - \mu)$$

- Технически делается через спектральное разложение:
  - собственные вектора образуют ОНБ, т.к. $\Sigma = \Sigma^T$
  - собственные значение показывают растянутость данных вдоль осей собственных векторов.

$$\Sigma = Q \Lambda Q^T,$$

$$z = Q \Lambda^{-1/2} Q^T (x - \mu)$$

- Свойства[5]:

$$Ez = \mathbf{0},\ cov[z, z] = I.$$

[5] Докажите.

Евклидово расстояние в декоррелированном пространстве

- Расстояние между x и x' = Евклидовому расстоянию в декоррелированном пространстве между $z = \Sigma^{-1/2}(x - \mu)$ и $z' = \Sigma^{-1/2}(x' - \mu)$:

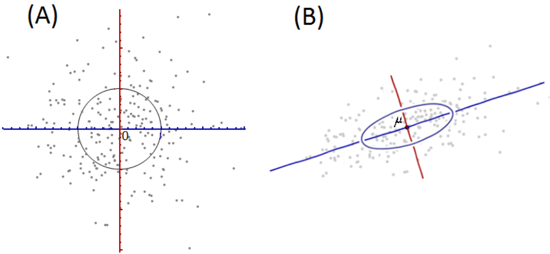
$$\begin{aligned}
\rho_M(x, x') &= \rho_E(z, z') = \sqrt{(z - z')^T (z - z')} = \\
&= \sqrt{\left(\Sigma^{-1/2}(x - x')\right)^T \Sigma^{-1/2}(x - x')} \\
&= \sqrt{(x - x')^T \Sigma^{-1/2}\Sigma^{-1/2}(x - x')} \\
&= \sqrt{(x - x')^T \Sigma^{-1}(x - x')}
\end{aligned}$$

- Это расстояние Махаланобиса[6]

[6] Как расстояние упроститься для нескоррелированных признаков разной дисперсии? Проинтерпретируйте.

Идея расстояния Махаланобиса

Множество равноудаленных объектов от (0,0)
(A): в декоррелированном пространстве $\{z : \rho_E(z,0)^2 = 1\}$.
(B): в исходном пространстве
$\{x : \rho_M(x,\mu)^2 = (x-\mu)^T\Sigma^{-1}(x-\mu) = 1\}^7$.



[7] Докажите, что это действительно будет эллипс (используйте спектральное разложение и свойства Σ).

Обобщение расстояния Махаланобиса

- Заменим Σ^{-1} на матрицу M:
$$\rho_M(x,x') = \sqrt{(x-x')^T M (x-x')}$$
- Нужна неотрицательность под корнем, поэтому ищем M в виде
$$M = LL^T$$
- Тогда $M = M^T$ и $M \succcurlyeq 0$.
- Выберем такую L, чтобы K-NN давал максимальную точность на кросс-валидации.
- Возможны и другие параметризации $\rho(x,x')$.
- Пример важности выбора расстояния:
    - классификация человека по фото
    - определение позы по фото

Более сложные типы данных

- Сравнение строк:
    - редакторское расстояние (edit distance, Levenstein distance): минимальное число правок для перевода одной строки в другую.
        - каждому типу правки можно давать свой вес

```
i n t e n t i o n
                   ← delete i
n t e n t i o n
                   ← substitute n by e
e t e n t i o n
                   ← substitute t by x
e x e n t i o n
                   ← insert u
e x e n u t i o n
                   ← substitute n by c
e x e c u t i o n
```

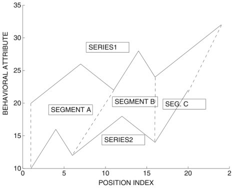
    - длина наибольшей общей подпоследовательности (longest common subsequence, используется в git):

$$MaxSubsequence(ABCDE, AXCYE) = ACE$$

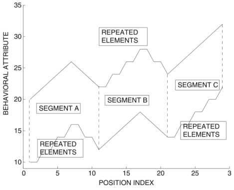
- Сравнение графов: редакторское расстояние и наибольший общий подграф.
- Реализуются эффективно методами динамического программирования.

Сравнение временных рядов

- Алгоритм динамической трансформации временной шкалы: перед сравнением ищем оптимальное локальное сжатие-растяжение рядов.
(англ. dynamic time warping)



(a) Original series



(b) Warped series

- Можно сравнивать спектры (коэффициенты разложения в ряде Фурье или др. базисе)
- Пример: распознавание речи.

Содержание

Оптимальный константный прогноз

Найдем для обучающей выборки $(x_1, y_1), \dots (x_N, y_N)$ оптимальный константный прогноз $\widehat{y} \in \mathbb{R}$:

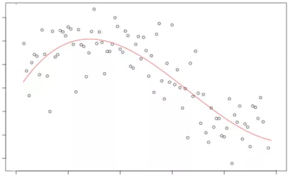
$$\widehat{y} = \arg\min_{\widehat{y} \in \mathbb{R}} \sum_{i=1}^{N} (\widehat{y} - y_i)^2 = \frac{1}{N} \sum_{n=1}^{N} y_n$$

Оптимальный константный прогноз

Найдем для обучающей выборки $(x_1, y_1), \dots (x_N, y_N)$ оптимальный константный прогноз $\widehat{y} \in \mathbb{R}$:

$$\widehat{y} = \arg\min_{\widehat{y} \in \mathbb{R}} \sum_{i=1}^{N} (\widehat{y} - y_i)^2 = \frac{1}{N} \sum_{n=1}^{N} y_n$$

Но нам нужно моделировать нелинейные закономерности:

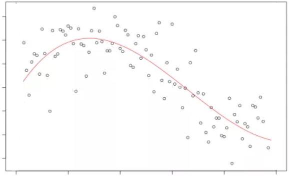


Оптимальный константный прогноз

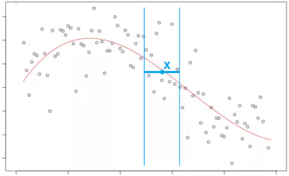
Найдем для обучающей выборки $(x_1, y_1), \ldots (x_N, y_N)$ оптимальный константный прогноз $\widehat{y} \in \mathbb{R}$:

$$\widehat{y} = \arg\min_{\widehat{y} \in \mathbb{R}} \sum_{i=1}^{N} (\widehat{y} - y_i)^2 = \frac{1}{N} \sum_{n=1}^{N} y_n$$

Но нам нужно моделировать нелинейные закономерности:



Регрессия Надарая-Ватсона - локальный константный прогноз.



Регрессия Надарая-Ватсона

- Найдем локальный оптимальный константный прогноз:

$$\widehat{y}(x)=\arg\min_{\widehat{y}\in\mathbb{R}}\sum_{i=1}^{N}w_i(x)(\widehat{y}-y_i)^2=\frac{\sum_{i=1}^{N}y_i w_i(x)}{\sum_{i=1}^{N}w_i(x)}$$

- Веса $\downarrow$ при $\uparrow \rho(x,x_i)$ за счет убывающей $K(u)$ ("ядра"):

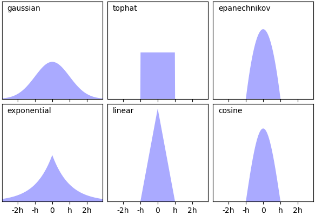
$$w_i(x)=K\left(\frac{\rho(x,x_i)}{h}\right)$$

- h - параметр "ширины окна"[8]
    - при $K(u)=\mathbb{I}[u\le 1]$ решение зависит только от окрестности радиуса h вокруг x.
- Англ. local constant regression, kernel regression.

[8] Как он влияет на сложность модели?

Функция ядра

Ядро $K(u)$	Формула
top-hat	$\mathbb{I}[\lvert u\rvert < 1]$
линейное	$\max\{0, 1 - \lvert u\rvert\}$
Епанечникова	$\max\{0, 1 - u^2\}$
экспоненциальное	$e^{-\lvert u\rvert}$
Гауссово	$e^{-\frac{1}{2}u^2}$
квартичное	$(1 - u^2)^2\mathbb{I}[\lvert u\rvert < 1]$



Комментарии

- Веса обеспечивают нелинейность прогноза, но требуют пересчета для каждого x.
- При достаточно общих условиях $\widehat{y}(x) \xrightarrow{P} E[y|x]$
- Конкретный вид $K(u)$ не так важен для точности, как выбор h.
- Выбор $K(u)$ влияет на вычислительную сложность.
- Возможен динамический выбор $h(x)$.
    - $h(x)$ ниже, если локальная плотность точек - выше, например $h(x)$ - расстояние до K ближайшего соседа x.[9]

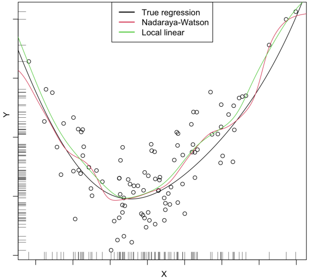
[9] При каком выборе $h(x)$ и $K(u)$ метод превращается в K-NN?

Локальная линейная регрессия

Вместо локальной константы можно оптимизировать локально линейную регрессию:

$$\sum_{i=1}^{N} w_i(x)(x^T\beta - y_i)^2 \to \min_{\beta\in\mathbb{R}}; \quad \widehat{y}(x) = x^T\beta$$

Она устойчивее, лучше аппроксимирует области низкой плотности объектов, но вычислительно сложнее.



Заключение

- Масштаб признаков влияет на прогноз.
- Проклятие размерности ухудшает качество локальных метрических методов.
- Метод ближайших центроидов - простая базовая модель.
- Метод K ближайших соседей:
    - K контролирует сложность модели
    - $\rho(x, x')$: выбирается из смысла задачи
    - быстрое обучение, медленный прогноз (возможны ускорения)
    - взвешенный учет соседей
- Регрессия Надарая-Ватсона
    - h контролирует сложность модели
    - обобщение: локальная линейная регрессия