Машинное обучение (Machine Learning)

Метрические методы классификации и регрессии

Уткин Л.В.

ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

Содержание

Презентация является компиляцией и заимствованием материалов из замечательных курсов и презентаций по машинному обучению:

К.В. Воронцова, А.Г. Дьяконова, Н.Ю. Золотых, С.И. Николенко, Andrew Moore, Lior Rokach, Rong Jin, Luis F. Teixeira, Alexander Statnikov и других.

Наивный байесовский классификатор

Теорема Байеса

Thomas Bayes

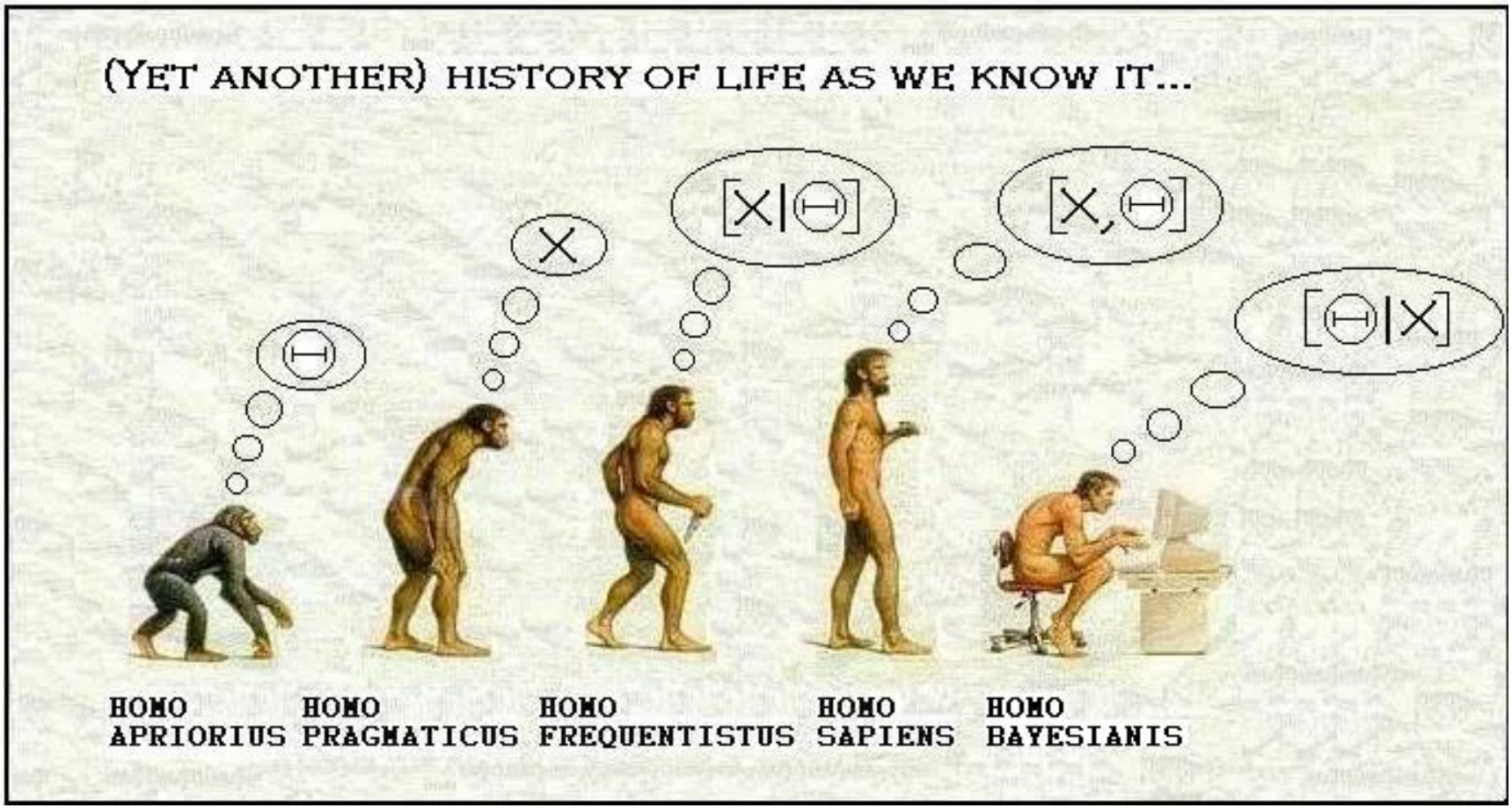
1702 - 1761

Теорема Байеса

$$P(y = c|x) = \frac{P(x|y = c)P(y = c)}{P(x)},$$

$P(y = c|x)$ - вероятность что объект x принадлежит классу c (апостериорная вероятность класса);
$P(x|y = c)$ - вероятность встретить объект x среди всех объектов класса c;
$P(y = c)$ - безусловная вероятность встретить объект класса c (априорная вероятность класса);
$P(x)$ - безусловная вероятность объекта x.

Эволюция по Байесу



Теорема Байеса и классификация

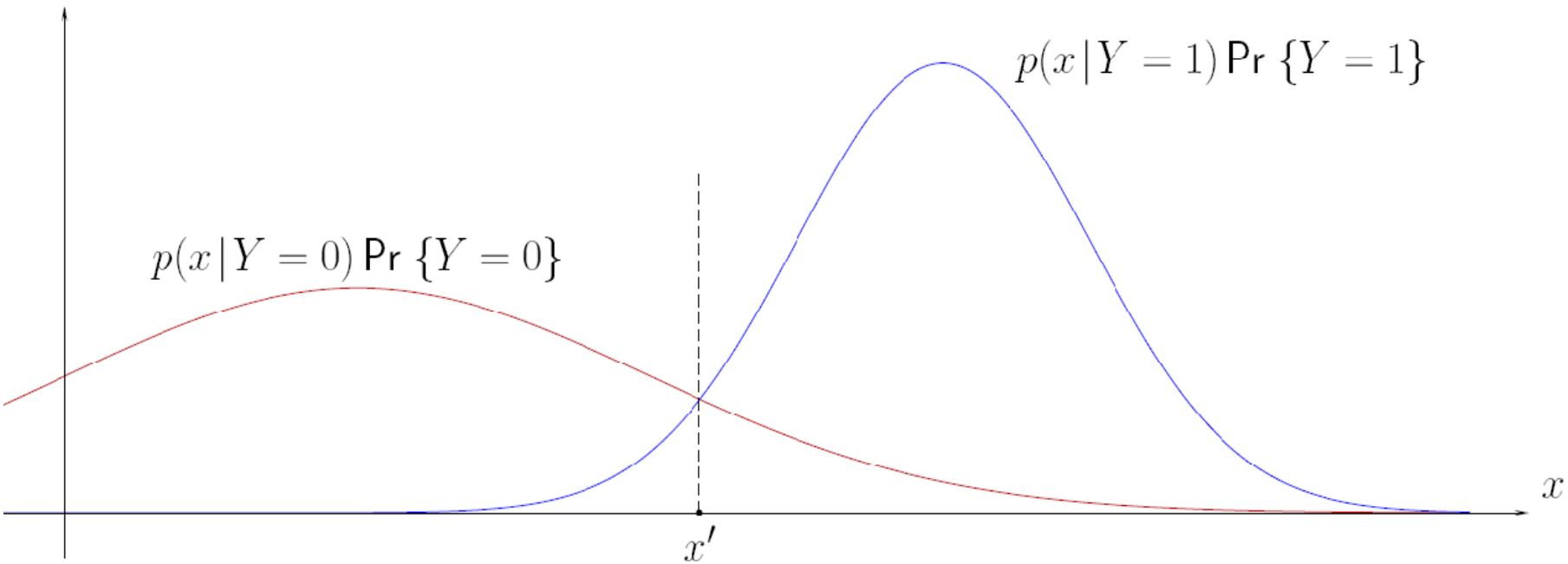
Цель классификации состоит в том чтобы понять к какому классу принадлежит объект x. Следовательно необходимо найти наиболее вероятный класс объекта x, т.е., необходимо из всех классов выбрать тот, который дает максимум вероятности $P(y = c|x)$:

$$c_{opt} = \arg\max_{c \in C} P(y = c|x) = \arg\max_{c \in C} \frac{P(x|y = c)P(y = c)}{P(x)}.$$

Для каждого класса c вычисляется $P(y = c|x)$ и выбирается класс, имеющий максимальную вероятность. Вероятность $P(x)$ не зависит от c и является константой:

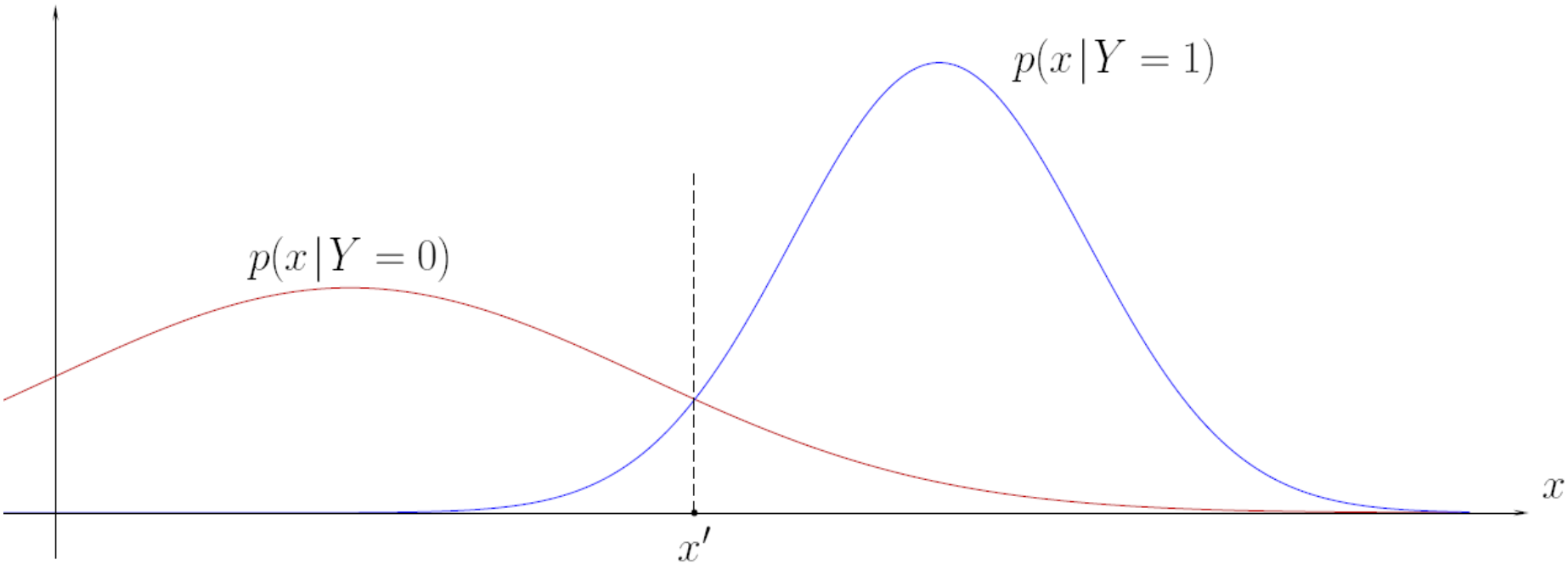
$$c_{opt} = \arg\max_{c \in C} P(x|y = c)P(y = c).$$

Принцип максимума апостериорной вероятности



При $x < x'$ считаем $c_{opt} = 0$ иначе $c_{opt} = 1$

Принцип максимального правдоподобия



При $x < x'$ считаем $c_{opt} = 0$ иначе $c_{opt} = 1$

Теорема Байеса и классификация (2 класса)

Выбор:

$$\begin{cases} \text{класс } c_1, & \text{если } P(y = c_1|x) > P(y = c_2|x) \\ \text{класс } c_2, & \text{иначе} \end{cases}$$

или

$$\begin{cases} \text{класс } c_1, & \text{если } \dfrac{P(x|y = c_1)}{P(x|y = c_2)} > \dfrac{P(y = c_2)}{P(y = c_1)} \\ \text{класс } c_2, & \text{иначе} \end{cases}$$

Байесовский классификатор минимизирует ошибку принятия решений

Наивность классификатора

Байесовский классификатор представляет объект как набор признаков (атрибутов), вероятности которых условно не зависят друг от друга:

$$P(x|y=c) = P(f_1|y=c)P(f_2|y=c)\cdots P(f_m|y=c) = \prod_{i=1}^{m} P(f_i|y=c).$$

Наивный байесовский классификатор:

$$c_{opt} = \arg\max_{c\in C} \left(P(y=c)\prod_{i=1}^{m} P(f_i|y=c)\right).$$

или

$$c_{opt} = \arg\max_{c\in C} \left(\log P(y=c) + \sum_{i=1}^{m} \log P(f_i|y=c)\right).$$

Оценка априорных вероятностей классов, если признаки категориальные

Вероятность класса $P(y = c)$ оценивается по обучающей выборке как:

$$P(y = c) = N_c/N$$

N_c – количество объектов, принадлежащих классу ,
N – общее количество объектов в обучающей выборке.

Оценка вероятностей признаков, если они категориальные

Вероятность $P(f_i|y = c)$ оценивается по обучающей выборке как:

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

$M_i(c)$ - общее количество элементов с заданным значением признака i в классе c.
$\alpha > 0$ - для избежания нулевых значений вероятности, например, $\alpha = 1$

Пример классификации спама (1)

Есть три письма для которых известны их классы (C - спам и H - не спам):

- [C] предоставляю услуги бухгалтера;
- [C] спешите купить iPhone;
- [H] надо купить молоко.

Модель классификатора будет выглядеть следующим образом:

	C	H	$P(y = \text{C})$	$P(y = \text{H})$
частота классов	2	1	2/3	1/3

Пример классификации спама (2)

	С	Н	$P(f_i \mid y = \text{С})$	$P(f_i \mid y = \text{Н})$
предоставляю	1	0	$(1+1)/(6+8)$	$(0+1)/(3+8)$
услуги	1	0	$(1+1)/(6+8)$	$(0+1)/(3+8)$
бухгалтера	1	0	$(1+1)/(6+8)$	$(0+1)/(3+8)$
спешите	1	0	$(1+1)/(6+8)$	$(0+1)/(3+8)$
купить	1	1	$(1+1)/(6+8)$	$(1+1)/(3+8)$
iPhone	1	0	$(1+1)/(6+8)$	$(0+1)/(3+8)$
надо	0	1	$(0+1)/(6+8)$	$(1+1)/(3+8)$
молоко	0	1	$(0+1)/(6+8)$	$(1+1)/(3+8)$

Пример классификации спама (3)

“Надо купить вино” - спам или нет?

- Для класса СПАМ:

$$P(y = \text{СПАМ} \mid \ldots \text{вино}) = \frac{2}{3} \cdot \frac{1}{6+8} \cdot \frac{2}{6+8} \cdot \frac{1}{6+8} = 5 \times 10^{-4}$$

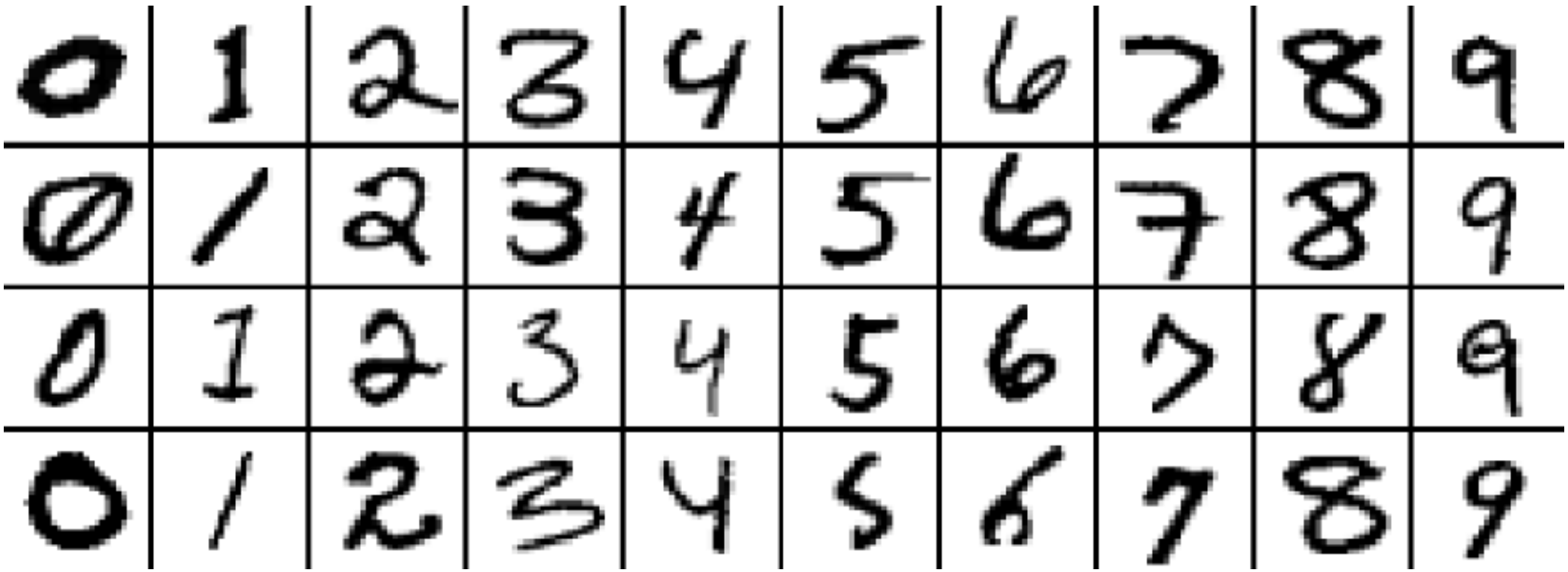
- Для класса НЕ СПАМ:

$$P(y = \text{НЕ СПАМ} \mid \ldots \text{вино}) = \frac{1}{3} \cdot \frac{2}{3+8} \cdot \frac{2}{3+8} \cdot \frac{1}{3+8} = 1 \times 10^{-3}$$

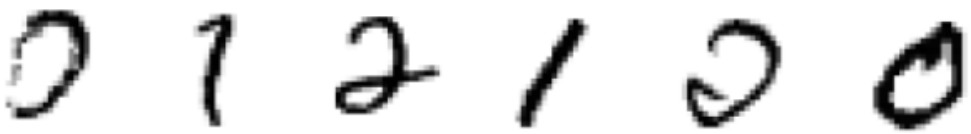
Итог: это не спам!

Пример распознавания рукописных цифр

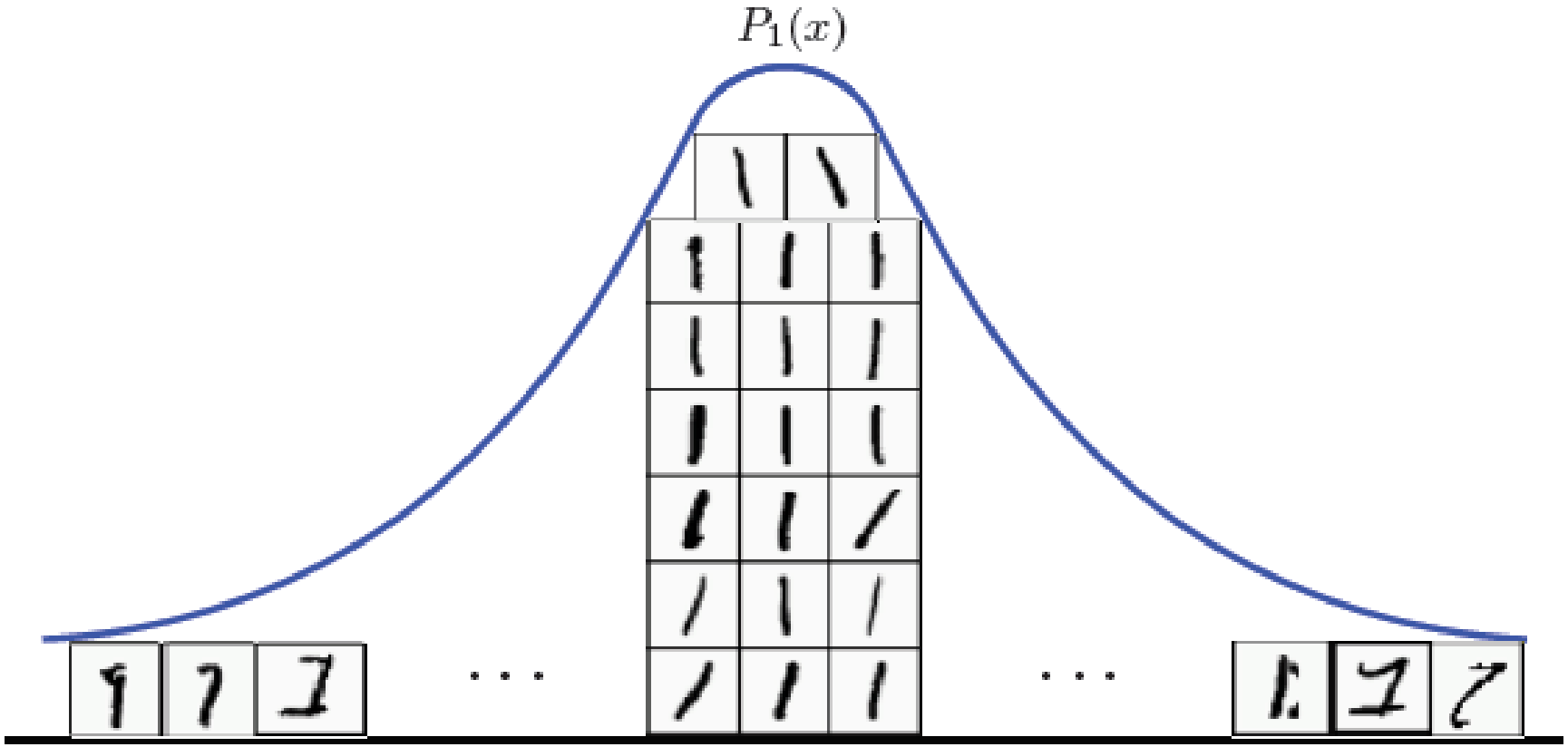
Обучающая выборка



Что это?

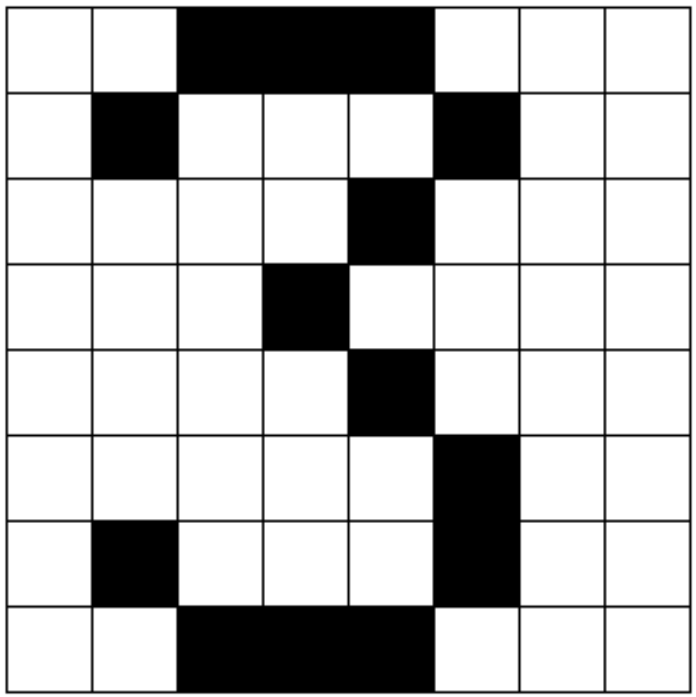


Пример распознавания рукописных цифр



Пример распознавания рукописных цифр

- 64 клетки - 64 бинарных признака (более белая/более черная-0/1): $f_{i,j}$
- Вектор признаков для 3: $(0, 0, 1, 1, ..., 1, 0, 0, 0)$

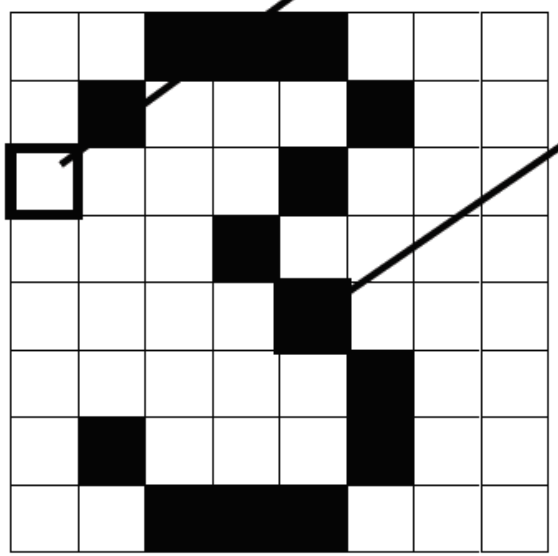


Пример распознавания рукописных цифр

$P(Y)$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1

$Y = 3$



$P(f_{3,1} = 1 \mid Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(f_{5,5} = 1 \mid Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

Пример распознавания рукописных цифр

$P(f_{i,j} = 1|y = 3)$ - доля всех троек с черной клеткой i, j
$P(f_{i,j} = 0|y = 3)$ - доля всех троек с белой клеткой i, j
Сравним

$$P(y = 3|f) = P(y = 3) \cdot P(f_{1,1} = 0|y = 3)P(f_{1,2} = 0|y = 3) \times$$
$$\times P(f_{1,3} = 1|y = 3) \cdots P(f_{8,8} = 0|y = 3)$$

с другими цифрами, например, с цифрой 8:

$$P(y = 8|f) = P(y = 8) \cdot P(f_{1,1} = 0|y = 8)P(f_{1,2} = 0|y = 8) \times$$
$$\times P(f_{1,3} = 1|y = 8) \cdots P(f_{8,8} = 0|y = 8)$$

Случай количественных признаков

Одномерный непрерывный случай: эмпирическая оценка **плотности**

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

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

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

Случай количественных признаков

Многомерный непрерывный случай (m признаков объекта): оценка плотности в точке $x = (\xi_1, ..., \xi_m)$:

$$p_h(x) = \frac{1}{2n} \sum_{i=1}^{n} \prod_{j=1}^{m} \frac{1}{h_j} K\left(\frac{\xi_j - f_j(x_i)}{h_j}\right)$$

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

Программная реализация в R

- https://cran.r-project.org/web/views/MachineLearning.html
- Package **e1071**, функция **naiveBayes**
- Package **klaR**, функция **NaiveBayes**
- Package **BayesTree**, функция **bart**

Метрические методы классификации и регрессии

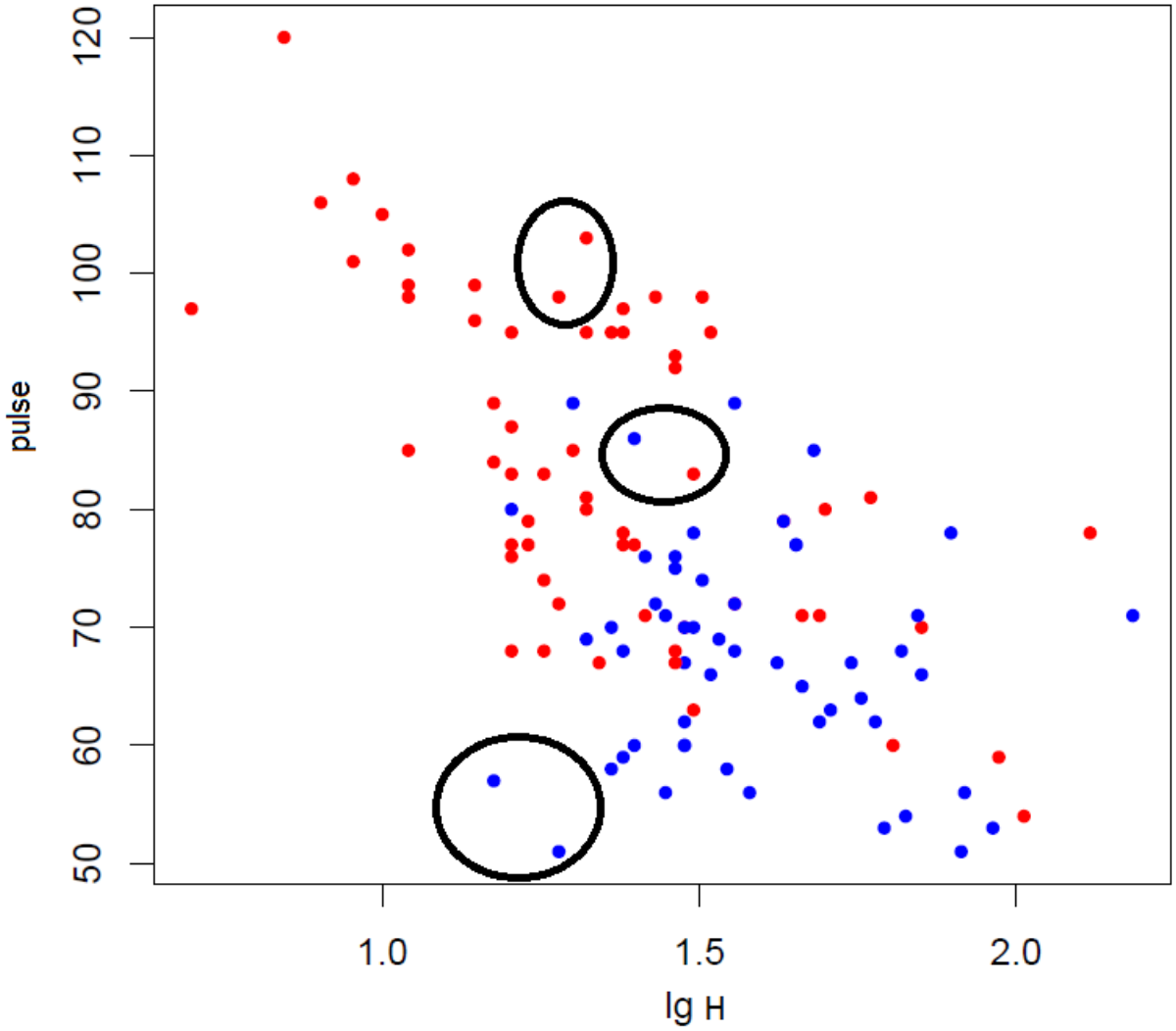
Гипотезы компактности или непрерывности

Задачи классификации и регрессии:
X - объекты, Y - ответы; $X^n = (x_i, y_i)_{i=1}^n$ - обучающая выборка.

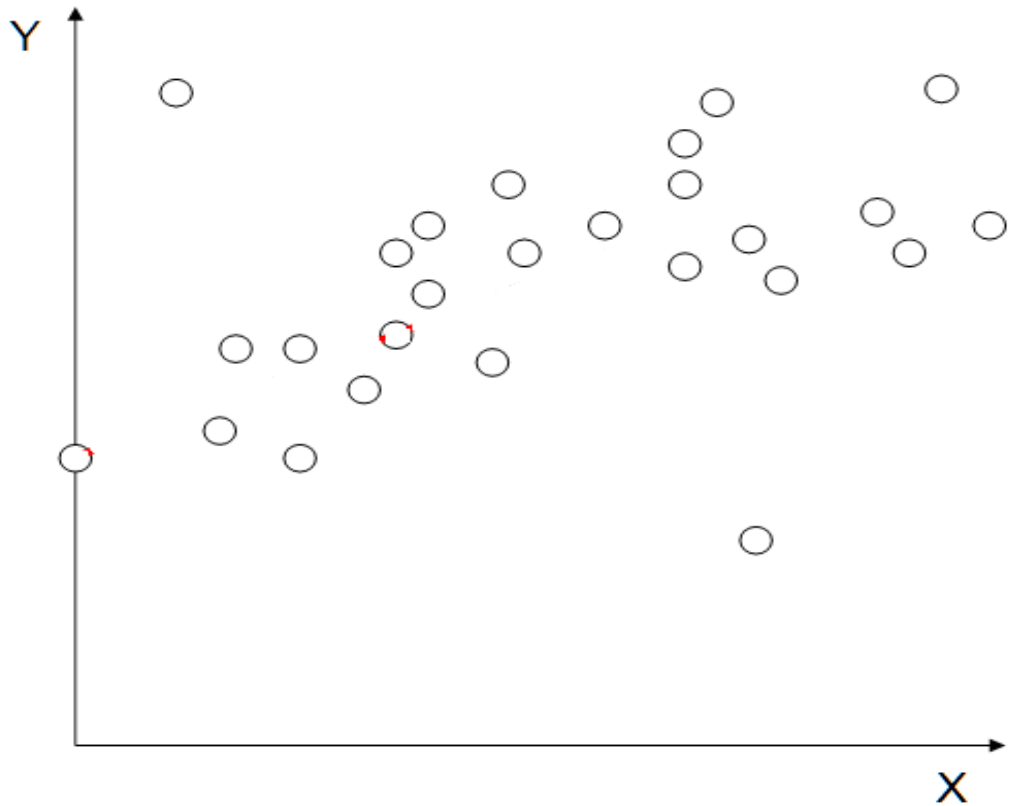
***Гипотеза компактности (для классификации)**: Близкие объекты, как правило, лежат в одном классе.*

***Гипотеза непрерывности (для регрессии):** Близким объектам соответствуют близкие ответы.*

Гипотеза компактности



Гипотеза непрерывности (нарушение)



Метод k ближайших соседей

Метод k ближайших соседей (kNN — k nearest neighbours) метрический алгоритм для классификации объектов, основанный на оценивании сходства объектов.
Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки.

Алгоритм:

1. Вычислить расстояние до каждого из объектов обучающей выборки
2. Отобрать k объектов обучающей выборки, расстояние до которых минимально
3. Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

Мера близости

Что такое близкие объекты? Задана функция расстояния $\rho : X \times X \to [0, \infty)$.
Виды функций расстояния:

- Евклидово: $\rho(x_i, x_j) = \sqrt{\sum_{k=1}^{m} w_k \left(x_i^{(k)} - x_j^{(k)}\right)^2}$
- L_p-метрика: $\rho(x_i, x_j) = \left(\sum_{k=1}^{m} w_k \left|x_i^{(k)} - x_j^{(k)}\right|^p\right)^{1/p}$
- L_∞-метрика: $\rho(x_i, x_j) = \max_{k=1,\ldots,m} \left|x_i^{(k)} - x_j^{(k)}\right|$
- L_1-метрика: $\rho(x_i, x_j) = \sum_{k=1}^{m} \left|x_i^{(k)} - x_j^{(k)}\right|$

$x_i = (x_i^{(1)}, \ldots, x_i^{(m)})$ - вектор m признаков i-го объекта;
$x_j = (x_j^{(1)}, \ldots, x_j^{(m)})$ - вектор m признаков i-го объекта.

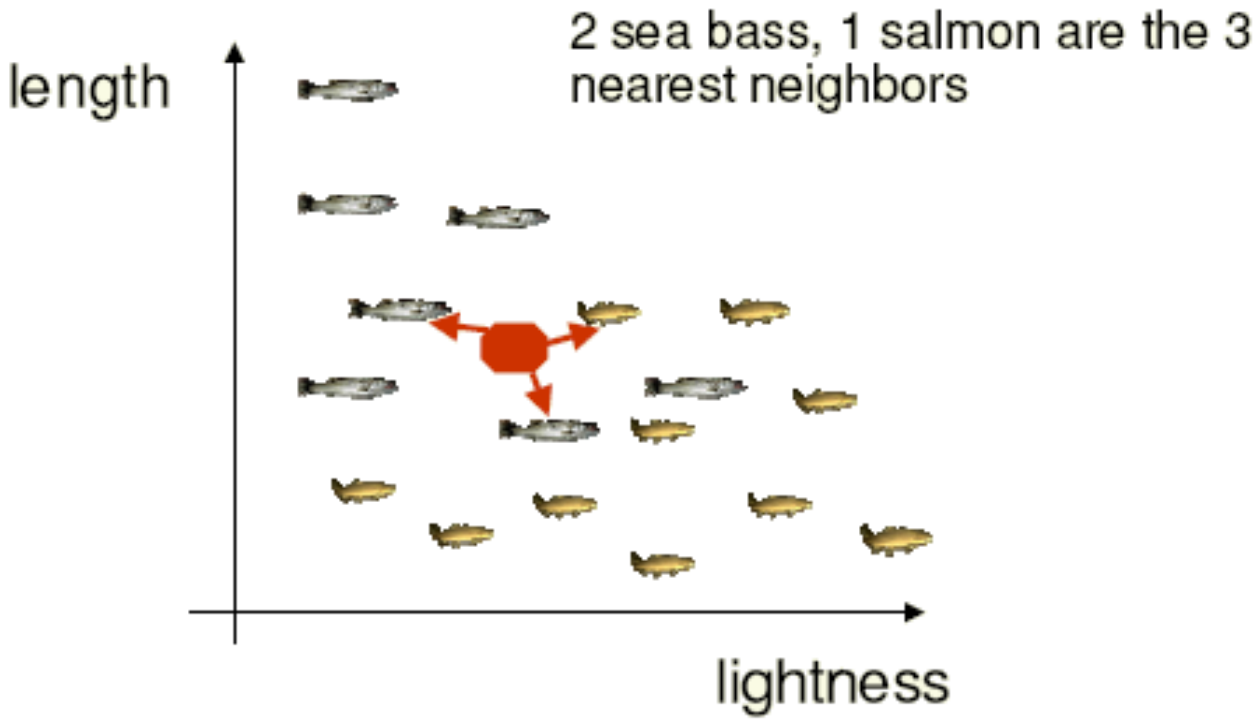
Мера близости

Что такое близкие объекты? Задана функция расстояния $\rho : X \times X \to [0, \infty)$.
Виды функций расстояния:

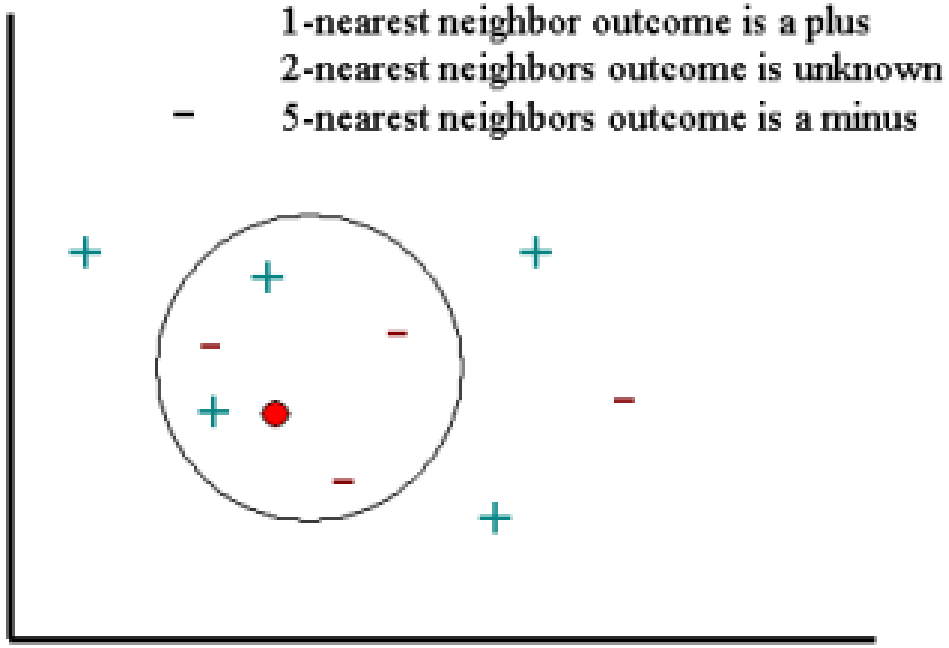
- Ланса-Уильямся: $\rho(x_i, x_j) = \frac{\sum_{k=1}^{m} \left| x_i^{(k)} - x_j^{(k)} \right|}{\sum_{k=1}^{m} \left(x_i^{(k)} + x_j^{(k)} \right)}$
- косинусная мера: $\rho(x_i, x_j) = \frac{\sum_{k=1}^{m} x_i^{(k)} x_j^{(k)}}{\sqrt{\sum_{k=1}^{m} \left(x_i^{(k)} \right)^2} \sqrt{\sum_{k=1}^{m} \left(x_j^{(k)} \right)^2}}$

$x_i = (x_i^{(1)}, ..., x_i^{(m)})$ - вектор m признаков i-го объекта;
$x_j = (x_j^{(1)}, ..., x_j^{(m)})$ - вектор m признаков i-го объекта.

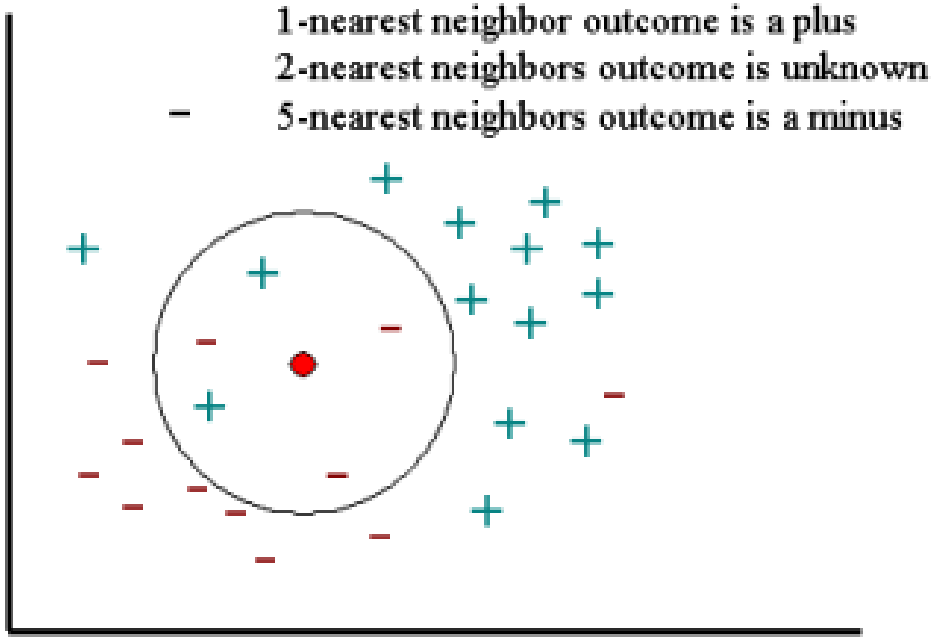
Метод k ближайших соседей (классификация)



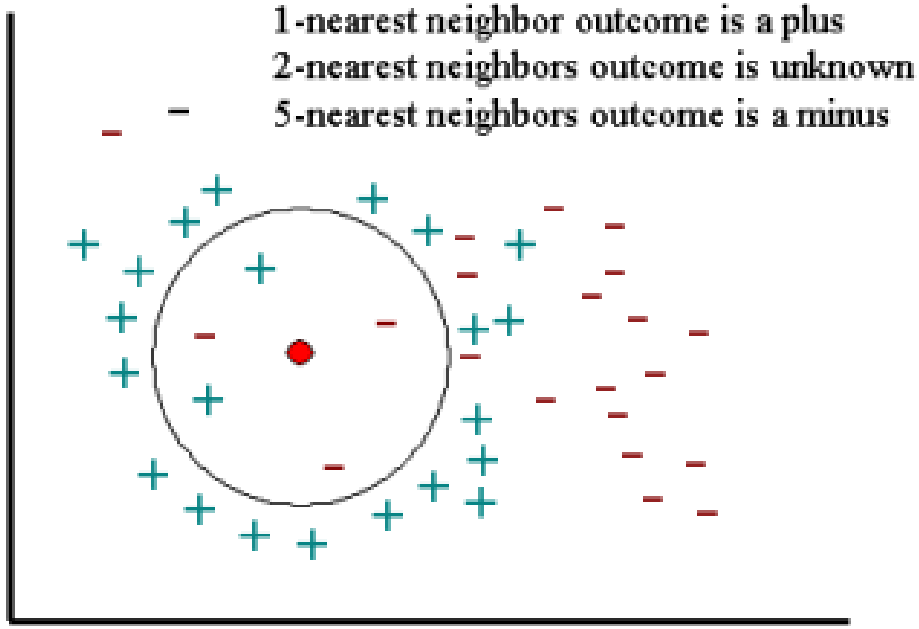
Метод k ближайших соседей (классификация)



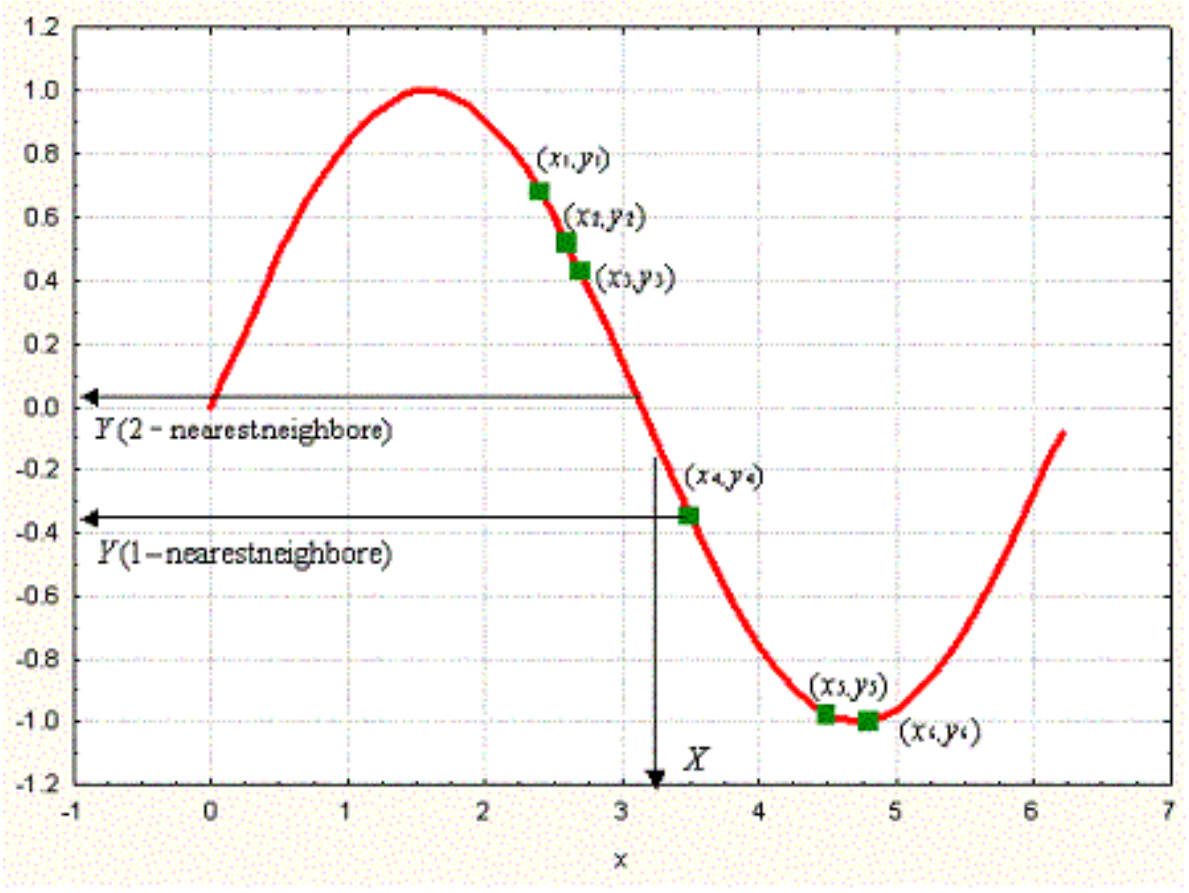
Метод k ближайших соседей: пример “правильной” классификации



Метод k ближайших соседей: пример ошибочной классификации



Метод k ближайших соседей (регрессия)



Метод k ближайших соседей

Достоинства:

- Простота реализации.
- Классификацию, проведенную алгоритмом, легко интерпретировать путем предъявления пользователю нескольких ближайших объектов.

Недостатки:

- Необходимость хранения обучающей выборки целиком.
- Поиск ближайшего соседа предполагает сравнение классифицируемого объекта со всеми объектами выборки.

Выбор k

- Малые значения k приведут к тому, что “шум” (выбросы) будет существенно влиять на результаты.
- Большие значения усложняют вычисления и искажают логику ближайших соседей, в соответствии с которой ближайшие точки могут принадлежать одному классу (гипотеза компактности).
- **Эвристика**: $k = \sqrt{n}$

Метод k ближайших соседей (пример)

Анализ брака древесины: по признакам средняя длина трещины и средний диаметр сучка

длина трещины	диаметр сучка	класс
7	7	брак
7	4	брак
3	4	не брак
1	4	не брак

Новый объект (длина трещины=3, диаметр сучка=7), $k = 3$

Метод k ближайших соседей (пример)

длина трещины	диаметр сучка	ρ
7	7	$(7-3)^2+(7-7)^2=16$
7	4	$(7-3)^2+(4-7)^2=25$
3	4	$(3-3)^2+(4-7)^2=9$
1	4	$(1-3)^2+(4-7)^2=13$

Метод k ближайших соседей (пример)

длина трещины	диаметр сучка	ρ	ранк	входит в 3 соседа?
7	7	16	3	да
7	4	25	4	нет
3	4	9	1	да
1	4	13	2	да

Метод k ближайших соседей (пример)

длина трещины	диаметр сучка	ρ	ранк	класс объекта
7	7	16	3	брак
7	4	25	4	-
3	4	9	1	не брак
1	4	13	2	не брак

*Объект (3,7) принадлежит классу "**не брак**"*

Вероятностная интерпретация метода ближайших соседей

- Метод ближайших соседей пытается аппроксимировать байесовское решающее правило на множестве обучающих данных
- Для этого необходимо вычислить условную вероятность $P(x|y)$ данных x при условии их принадлежности классу y, априорную вероятность каждого класса $P(y)$ и маргинальную вероятность данных $P(x)$.
- Эти вероятности вычисляются для некоторой малой области вокруг нового примера, размер области будет зависеть от распределения вероятностей на тестовых примерах

Вычисление вероятностей для kNN

- Пусть “шар” размерности m (m - число признаков) вокруг нового примера zсодержит k ближайших соседей для z
- Тогда

$$P(z) = \frac{k}{n}, \; P(z|y=1) = \frac{k_1}{n_1}, \; P(y=1) = \frac{n_1}{n}$$

- $P(z)$ - вероятность того, что случайная точка находится в “шаре”
- $P(z|y=1)$ - вероятность того, что случайная точка из класса 1 находится в “шаре”
- n_1, k_1 - число примеров из класса 1 и из класса 1 в k

Вычисление вероятностей для kNN

- $$P(z) = \frac{k}{n},\ P(z|y=1) = \frac{k_1}{n_1},\ P(y=1) = \frac{n_1}{n}$$
- Используем правило Байеса

$$P(y=1|z) = \frac{P(z|y=1)P(y=1)}{P(z)} =$$

$$= \frac{\frac{k_1}{n_1} \cdot \frac{n_1}{n}}{\frac{k}{n}} = \frac{k_1}{k}$$

Вычисление вероятностей для kNN

- Правило Байеса

$$P(y = 1|z) = \frac{k_1}{k}, \quad P(y = -1|z) = \frac{k_{-1}}{k}$$

Используя решающее правило Байеса, мы выбираем класс с наибольшей вероятностью, т.е. сравниваем $P(y = 1|z)$ и $P(y = -1|z)$. А это тоже самое, что сравнение k_1/k и k_{-1}/k.

Метод ближайшего соседа (общий вид)

Для произвольного $x^* \in X$ отсортируем объекты $x_1, \ldots, x_n$:

$$\rho(x^*, x_1) \leq \rho(x^*, x_2) \leq \ldots \leq \rho(x^*, x_n)$$

x_i - i-ый сосед объекта x^*; y_i - ответ на i-ом соседе объекта x^*.

Метрический алгоритм классификации:

$$a(x^*) = \arg\max_{y \in Y} \underbrace{\sum_{i=1}^{n} [y_i = y] \cdot w(i, x^*)}_{\Gamma_y(x^*)}$$

$w(i, x^*)$ - вес (степень важности) i-го соседа объекта x^*, неотрицателен, не возрастает по i.

$\Gamma_y(x^*)$ - оценка близости объекта x^* к классу y.

Метод ближайшего соседа (частный случай)

$$w(i, x^*) = [i = 1] = \begin{cases} 1, & i = 1, \\ 0, & i > 1. \end{cases},$$

т.е. решение принимается только по одному первому ближайшему соседу.

Преимущества:

- простота реализации;
- интерпретируемость решений,
- вывод на основе прецедентов (case-based reasoning)

Недостатки:

- неустойчивость к погрешностям (шуму, выбросам);
- отсутствие настраиваемых параметров;
- низкое качество классификации;
- приходится хранить всю выборку целиком.

Метод k ближайших соседей (частный случай общего вида)

$$w(i, x^*) = [i \leq k] = \begin{cases} 1, & i \leq k, \\ 0, & i > k. \end{cases}$$

т.е. решение принимается только по k ближайшим соседям.

Преимущества:

- менее чувствителен к шуму;
- появился параметр k.

Как найти оптимальное значение k в различный ситуациях

Оптимизация числа соседей k: функционал скользящего контроля leave-one-out:

$$LOO(k, X) = \sum_{i=1}^{n} [a(x_i) \neq y_i] \to \min_k$$

Метод окна Парзена

Усложнение: определить $w(i, x^*)$ как функцию от расстояния $\rho(x^*, x_i)$, а не от **ранга** соседа i:

$$w(i, x^*) = K\left(\frac{1}{h}\rho(x^*, x_i)\right)$$

Алгоритм:

$$a(x^*, X, h) = \arg\max_{y \in Y} \sum_{i=1}^{n} [y_i = y] K\left(\frac{1}{h}\rho(x^*, x_i)\right).$$

Параметр h - ширина окна (та же роль, что и число соседей k). Окно - сферическая окрестность объекта x^* радиуса h, при попадании в которую обучающий объект x_i “голосует” за отнесение объекта x^* к классу y_i.

Метод окна Парзена

Параметр h можно задавать априори или определять по скользящему контролю.

- Слишком узкие окна приводят к неустойчивой классификации,
- Слишком широкие окна приводят к вырождению алгоритма в константу.

Метод потенциальных функций

В методе парзеновского окна центр радиального ядра $K\left(\frac{1}{h}\rho(x^*, x_i)\right)$ помещается в классифицируемый объект x^*.
Двойственный взгляд: ядро помещается в каждый обучающий объект x_i и "притягивает" объект x^* к классу y_i, если он попадает в его окрестность радиуса h_i:

$$a(x^*, X, h) = \arg\max_{y \in Y} \sum_{i=1}^{n} [y_i = y]\, \gamma_i K\left(\frac{\rho(x^*, x_i)}{h_i}\right).$$

γ_i - величина "заряда" в точке x_i;
h_i - "радиус действия" потенциала с центром в точке x_i;
y_i - знак "заряда".

Метод потенциальных функций

- Идея метода потенциальных функций имеет прямую физическую аналогию с электрическим потенциалом
- При $Y = \{-1, +1\}$ обучающие объекты - это положительные и отрицательные электрические заряды
- коэффициенты γ_i - абсолютные величины этих зарядов
- ядро $K(z)$ - зависимость потенциала от расстояния до заряда
- сама задача классификации - ответ на вопрос: какой знак имеет электростатический потенциал в заданной точке пространства x^*.

Программная реализация в R

- Package **kknn**, функция **kknn**

Вопросы

?