Метрическая система

Метод ближайших соседей

Евгения Сумина, программист-исследователь

education

План занятия

Алгоритм K-NN

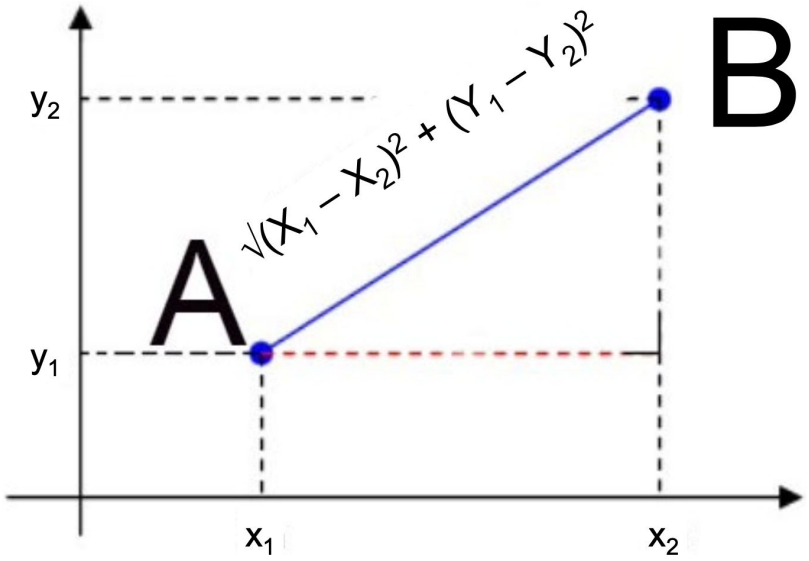
Что такое метрика. Какие бывают метрики

X - множество объектов

Y - множество ответов

d(X, X) -> R - метрика (расстояние) между объектами множества X.

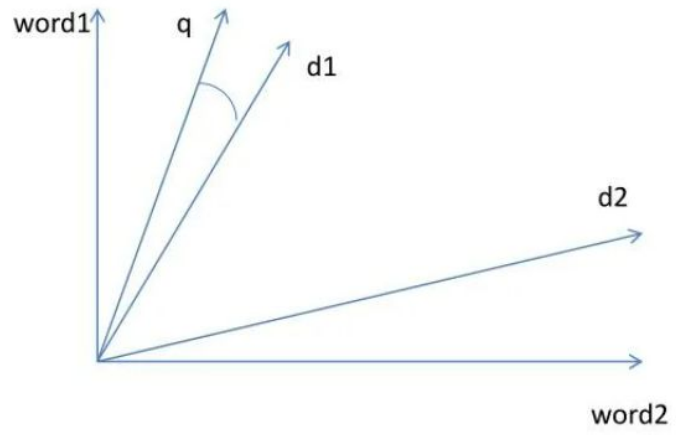
Например: евклидово расстояние — это геометрическое расстояние между двумя точками с координатами A и B.



Что такое метрика. Какие бывают метрики

Для поиска расстояния между векторами, соответствующими предложениям, часто используют косинусную метрику.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

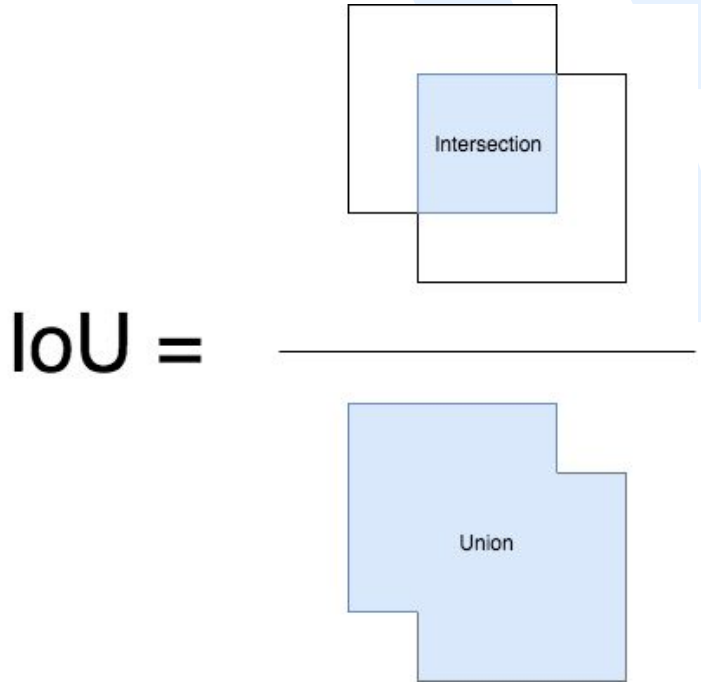


где A_i и B_i являются компонентами вектора A и B соответственно.

Что такое метрика. Какие бывают метрики

Для измерения сходства множеств (или картинок) часто используют метрику IoU - intersection over union.

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$



Как вы поняли метод ближайших соседей?

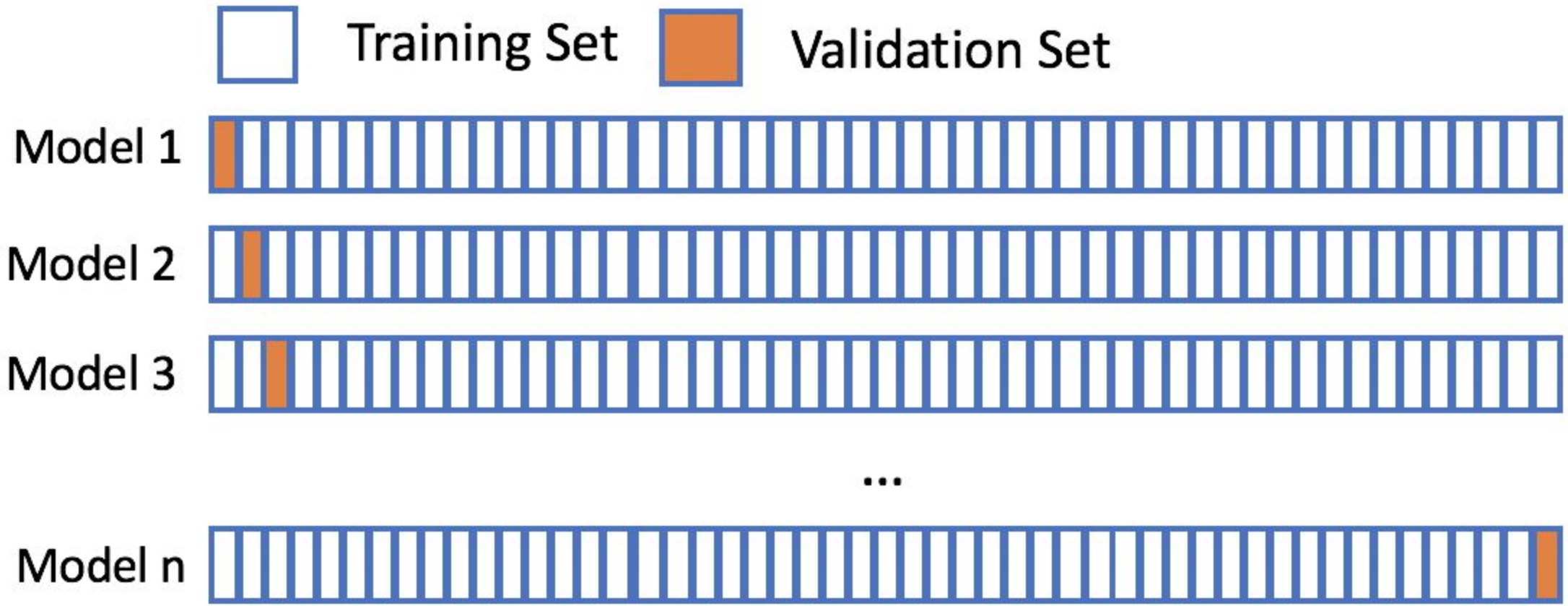
-

Что такое метод ближайших соседей?

Концептуально:

Практика

Как выбрать число соседей



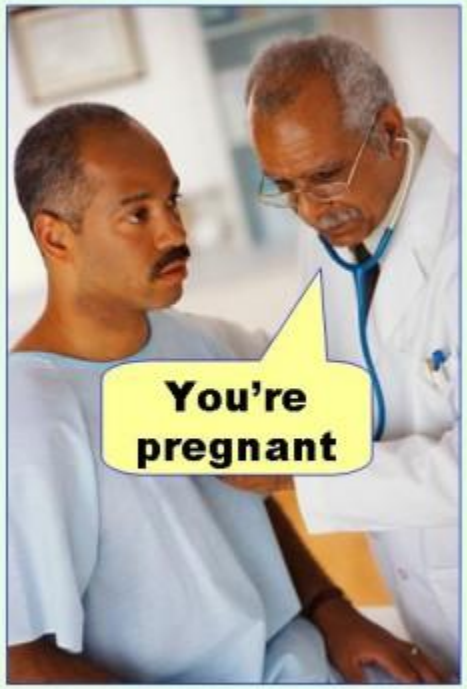
Матрица ошибок

Матрица ошибок

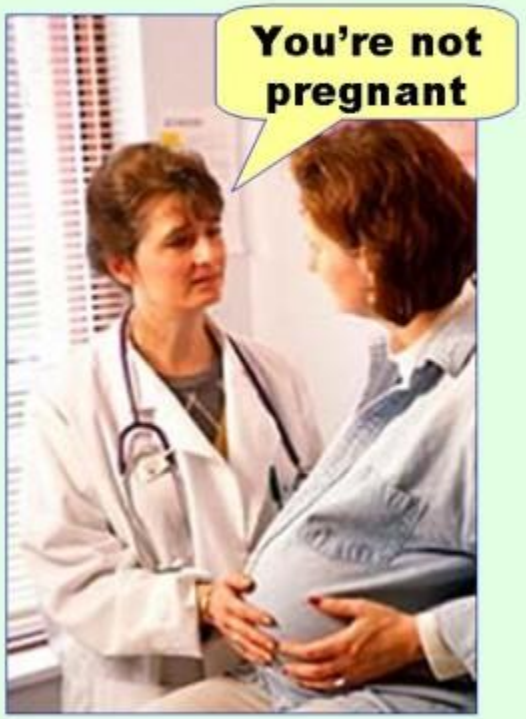
	Фактические значения	
Модель	+	–
+	TP	FP
–	FN	TN

Матрица ошибок

Type I error
(false positive)



Type II error
(false negative)



Precision. Recall

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\ \%$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\ \%$$

F-мера

$$F = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_\beta = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}$$

$0 < \beta < 1$ — приоритет точности.

$\beta > 1$ — приоритет полноты.

$\beta = 1$ — сбалансированная F-мера.

Практика

Итоги

Преимущества метода:

- интерпретируемость
- простота реализации

Недостатки метода:

- неустойчивость к погрешностям
- приходится хранить всю выборку целиком

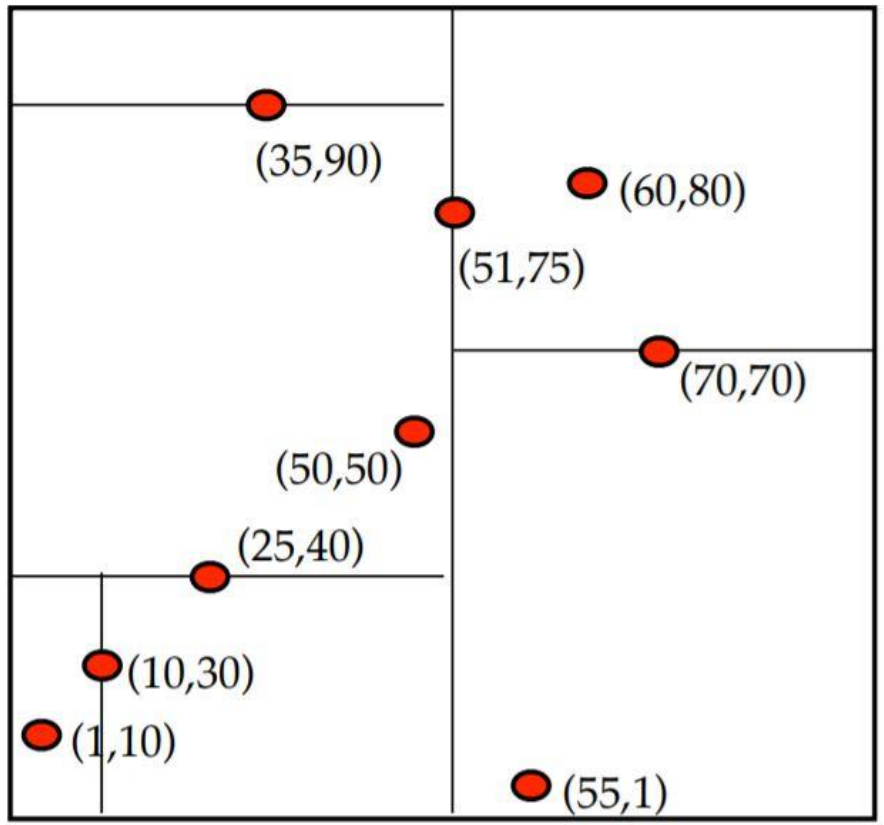
Вопросы

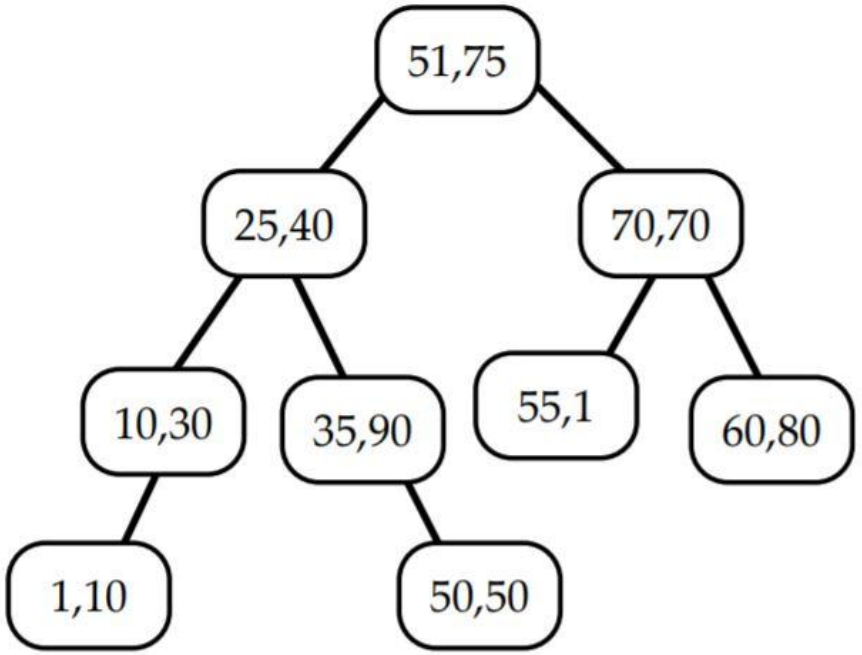
Дополнения

Что делать, если два класса набирают одинаковый рейтинг?

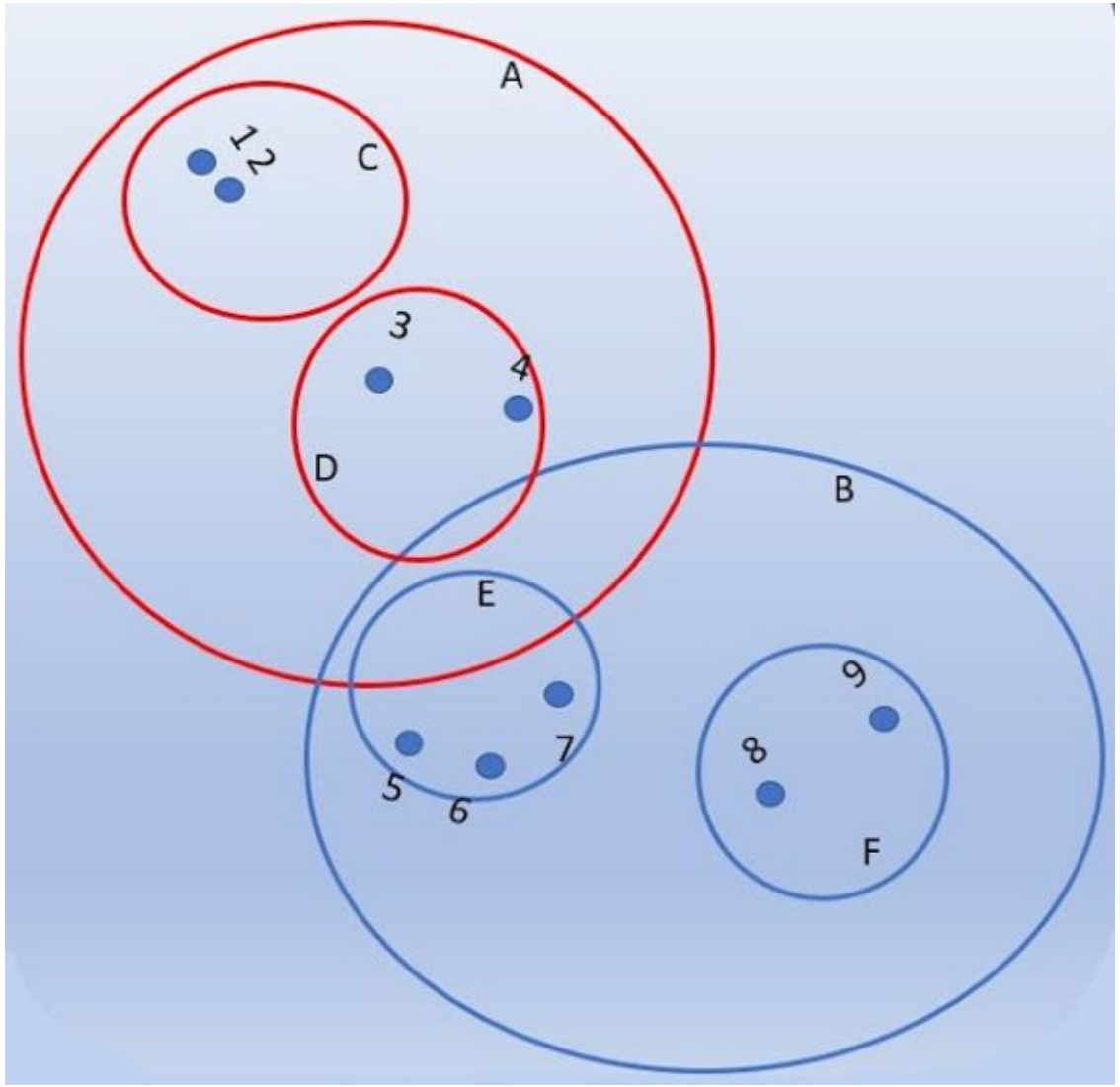
-

Вариации k-NN. KD-Tree





Вариации k-NN. Ball-Tree



Взвешенный метод ближайших соседей

“Взвешиваем” соседей, учитывая расстояние до них.

При голосовании учитываем “эталонность” объекта. (margin)

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

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

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

Влияет ли масштаб признаков на предсказание?

-

Взвешенный метод ближайших соседей

[175, 0.3], [185, 0.7]

[190, 0.1]

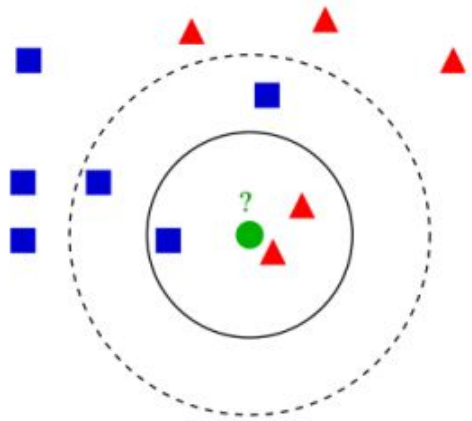
Какие методы называются метрическими?

- Основаны на анализе сходства объектов
- Не имеют фазы обучения
- Запоминают всю обучающую выборку
- На этапе предсказания просто ищут похожие на целевой объекты
- Исходят из допущения, что свойства объекта можно узнать, имея представление о его соседях

Что такое метод ближайших соседей?

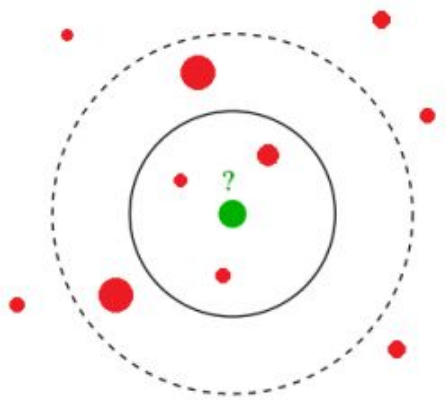
Классификация:

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.



Что такое метод ближайших соседей?

Формально:

Пусть дана обучающая выборка $X = (x_i, y_i)_{i=1}^N$, где $x_i \in \mathbb{X}$, $y_i \in \mathbb{Y} = \{1, \dots, C\}$. Пусть также задана некоторая симметричная по своим аргументам функция расстояния: $\rho : \mathbb{X} \times \mathbb{X} \to [0, +\infty)$

Предположим, что требуется классифицировать новый объект u. Для этого найдём k наиболее близких к u в смысле расстояния объектов обучающей выборки $X_k(u) = \{x_u^{(1)}, \dots, x_u^{(k)}\}$:

$$\forall x_{\text{in}} \in X_k(u)\ \forall x_{\text{out}} \in X \setminus X_k(u) \quad \rho(u, x_{\text{in}}) \leqslant \rho(u, x_{\text{out}}).$$

Метку класса объекта $x_u^{(i)}$ будем обозначать $y_u^{(i)}$. Класс нового объекта тогда естественным образом определим как наиболее часто встречающийся класс среди объектов из $X_k(u)$:

$$a(u) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} \mathbb{I}[y_u^{(i)} = y] \qquad (2)$$