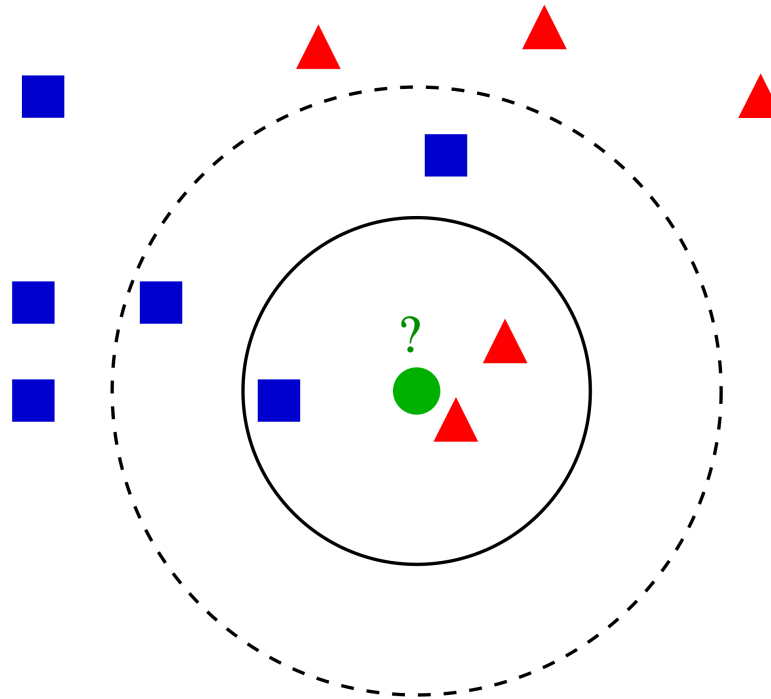


Машинное обучение • Основные алгоритмы

- разберем основные алгоритмы
- ответим на вопрос "какой признак самый важный?"
- визуализируем построенные алгоритмы

Метод ближайших соседей



- классификация:
 - берем k соседей и смотрим, какой класс встречается чаще
- регрессия:
 - берем и вычисляем среднее (можно средневзвешенное) значение для нового объекта

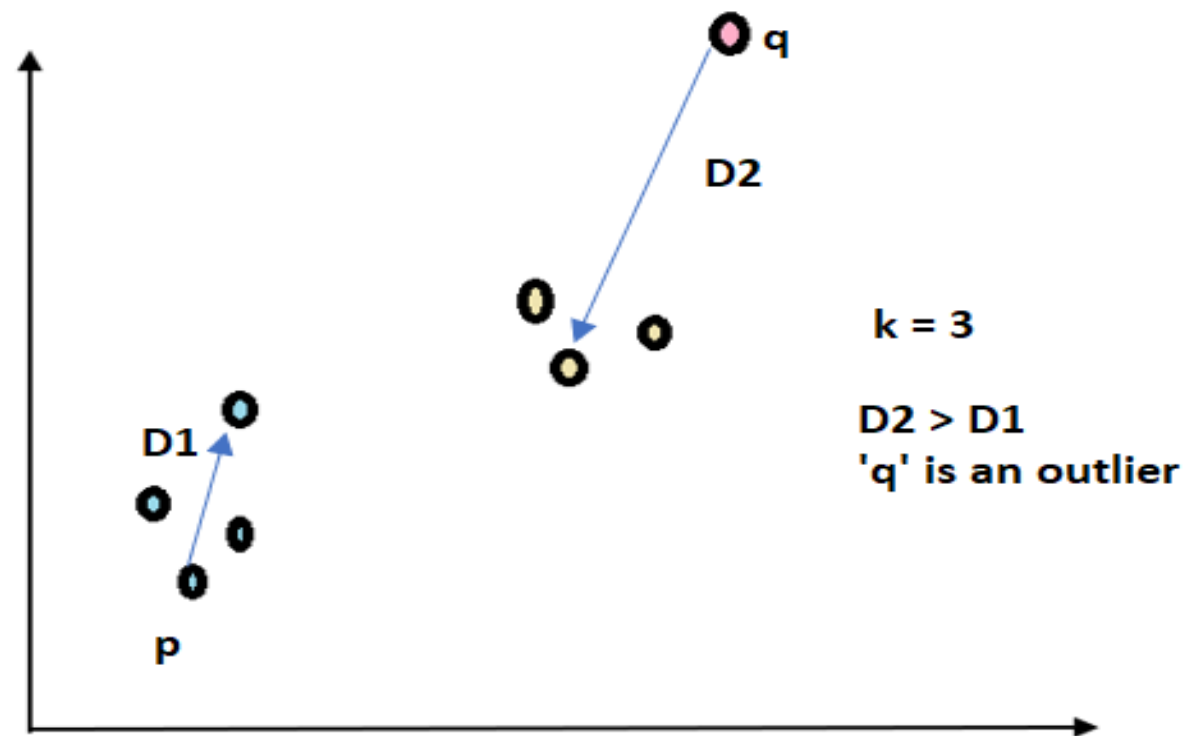
- Что значит ближайшие?

Близкие по метрике Минковского:

$$\rho(x, y) = \left(\sum_{i=0}^d |x_i - y_i|^p \right)^{1/p}$$

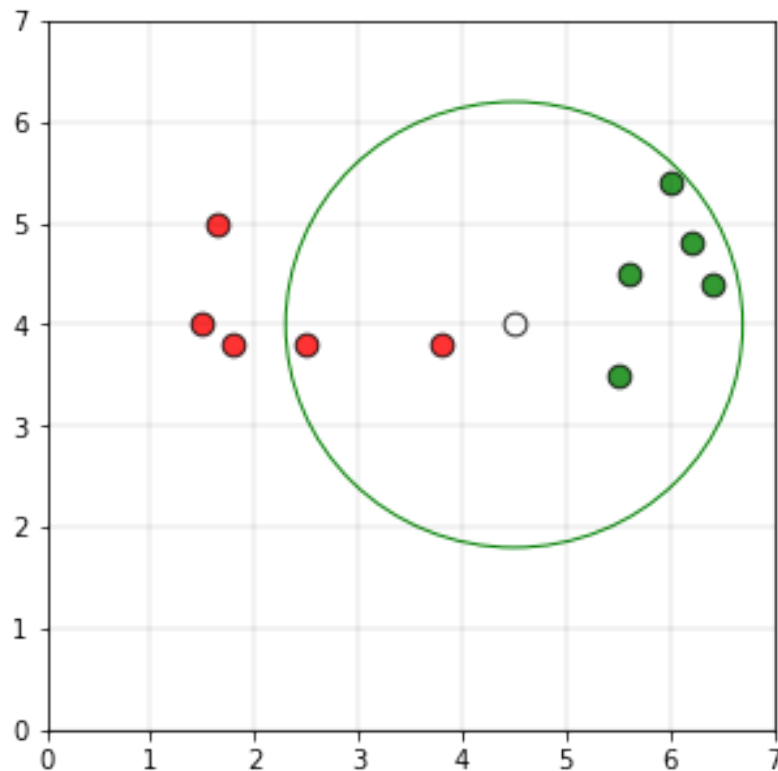
- при $p = 2$ это евклидово расстояние
- при $p = 1$ Манхэттенская метрика
- при $p = \infty$ - метрика Чебышева (наибольшее из всех расстояний)

Метод ближайших соседей: проблема



Если целевой объект расположен далеко, алгоритм все равно классифицирует объект. Можно использовать радиальный вариант.

Радиальный NN / RadiusNN



Но такой радиус сложно подобрать.

- число соседей / радиус
- метрика
- способ вычисления весов объектов

Decision Tree

- Классификация: игра состоится?

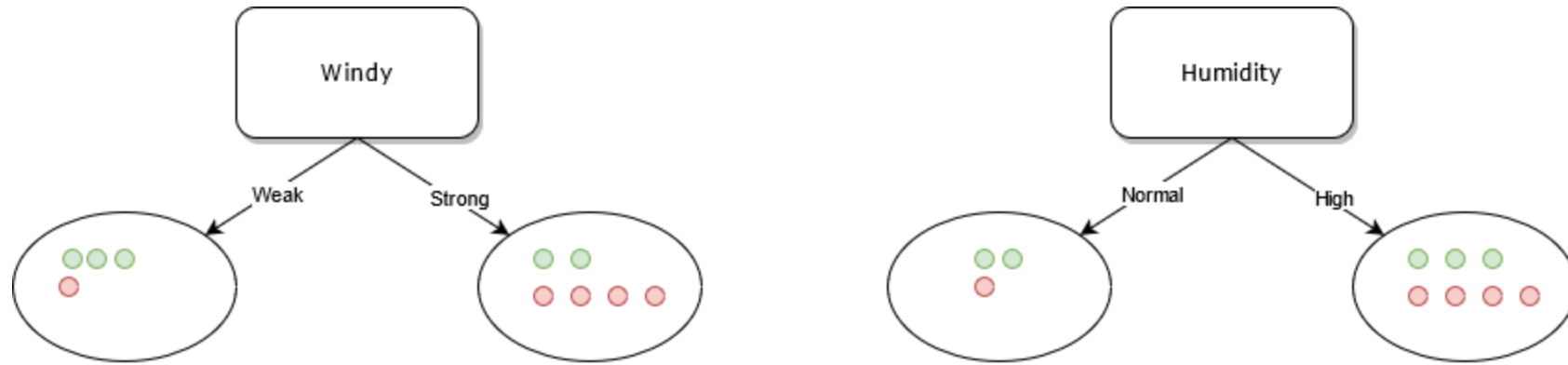
Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Какой признак самый информативный?

Тот, при использовании которого для классификации, мы получаем наиболее "чистые" подмножества объектов выборки:

- Признак `Windy` имеет два значения: `Weak` и `Strong`
 - Для `Windy=Weak`: `Play=Yes` – 3 объекта, `Play=No` – 1 объект
 - Для `Windy=Strong`: `Play=Yes` – 2 объекта, `Play=No` – 4 объекта
- Признак `Humidity` имеет два значения: `High` и `Normal`:
 - Для `Humidity=Normal`: `Play=Yes` – 2 объекта, `Play=No` – 1 объект
 - Для `Humidity=High`: `Play=Yes` – 3 объекта, `Play=No` – 4 объект

Оценки гомогенности для разбиений



Как измерить гомогенность выборки в полученных разбиениях?




- Коэффициент Джини / Gini impurity:

$$G = 1 - \sum_i p_i^2$$

- Энтропия разбиения:

$$H = -\sum p_i \log_2 p_i$$

p_i – частота объектов класса i в разбиении

Разбиение	Entropy
Исходное 	$-(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$
Windy=Weak 	$-(0.25 \log_2 0.25 + 0.75 \log_2 0.75) = 0.81$
Windy=Strong 	$-(0.33 \log_2 0.33 + 0.66 \log_2 0.66) = 0.92$

$$IG = S_0 - \sum_{i=1}^q \frac{N_i}{N} S_i,$$

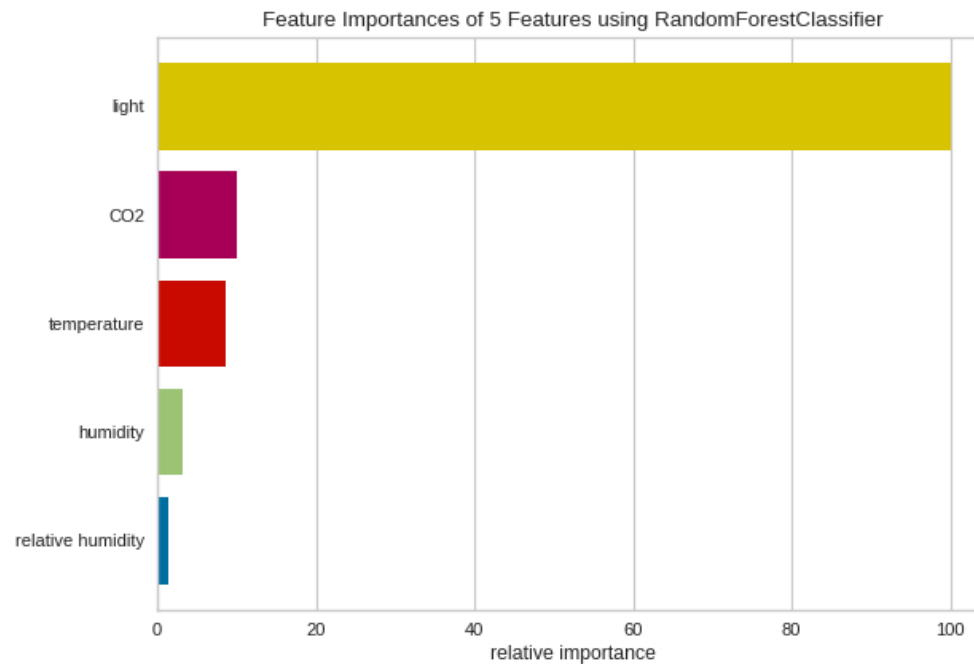
- q – число листьев (обычно 2),
- N_i – число объектов, попавших в i -ое разбиение,
- N – общее число в родительской вершине объектов,
- S_0 – *impurity metric* (*gini* или *entropy*) для исходного разбиения,
- S_i – *impurity metric* для i -го разбиения.

Для разбиения, построенного по признаку Windy :

$$IG = 1 - 0.4 \times 0.81 - 0.6 \times 0.92 = 0.12$$




Важность признаков • Feature importances

Суммарный (а в `sklearn` и нормированный) показатель уменьшения гетерогенности выборки используется для оценки важности признаков: если признак выбирался часто и сильно уменьшал энтропию или коэффициент Джини, то он является **информативным**.



- Критерий ветвления: `criterion: gini, entropy`
- Максимальная глубина: `max_depth`
- Минимальное число объектов в листе: `min_samples_leaf`
- Минимальное значение уменьшения гетерогенности для осуществления деления: `min_impurity_decrease`

`sklearn.tree` : <https://scikit-learn.org/stable/modules/tree.html>

- деревья - мощный алгоритм
- но склонен к переобучению
- может выступать в качестве составного блока для более сложных концепций - об этом завтра
- у обученного дерева есть атрибут: `model.feature_importances_`
- может использоваться для задач регрессии:
 -  **CART**: classification and regression trees
 -   **A Step By Step Regression Tree Example**