



# Введение в Машинное обучение

## Метрики качества/ матрица ошибок

## Простейший KNN алгоритм классификации

Преподаватель: Герард Костин

# План

- Введение
- Постановка задачи машинного обучения
- Задача классификации
- Метрики качества классификации
- KNN



# Типы ML алгоритмов

---

## **Основные категории (по типу обучения)**

- с учителем (supervised)
- без учителя (unsupervised)
- с частичным привлечением учителя (semisupervised)
- с подкреплением (reinforcement)

## **● Организация техники обучения**

- пакетная техника обучения, batch (offline) learning
- онлайновое обучение, online learning

## **● Типы алгоритмов**

- машинное обучение на примерах, instance-based
- машинное обучение на основе моделей, model-based

# Постановка Задачи

- Пусть

- $X$  — множество описаний объектов,
- $Y$  — множество допустимых ответов.

- Существует неизвестная целевая зависимость  $y^*$

- отображение  $y^* : X \rightarrow Y$
- значения  $y^*$  известны только на объектах конечной обучающей выборки  $X^n$
- $X^n = \{(x_1, y_1), \dots, (x_n, y_n)\}$

- Требуется построить алгоритм  $a$ , который приближал бы неизвестную целевую зависимость как на элементах выборки  $X^n$ , так и на всём множестве  $X$ .



# Функционал качества и функция потерь

- Вводится функция потерь  $L(y, y')$ , характеризующая величину отклонения ответа  $y = a(x)$  от правильного ответа  $y'$  на произвольном объекте  $x \in X$ .
- Типичный выбор функции потерь:
  - В задачах классификации  $L(y, y') = [y \neq y']$  (т.е. число ошибок классификации);
  - В задачах регрессии  $L(y, y') = (y - y')^2$ .
- Функционал качества
  - характеризует среднюю ошибку (эмпирический риск) алгоритма  $a$  на произвольной выборке  $X^n$
  - $$Q(a, X^n) = \frac{1}{n} \sum_{i=1}^n L(a(x_i), y^*(x_i)).$$
- Метод минимизации эмпирического риска
  - Требуется найти алгоритм  $a^*$ , минимизирующий среднюю ошибку на обучающей выборке:
  - $$a^* = \arg \min_{a \in A} Q(a, X^n).$$

# Признаки

## Типы признаков

"бинарный" признак: ;

"номинальный" признак: — конечное множество;

"порядковый" признак: — конечное упорядоченное множество;

"количественный" признак: — множество действительных чисел.

# Тестирование модели

Единственный способ узнать, насколько хорошо модель будет обобщаться на новые случаи, это фактически опробовать ее на новых примерах.

- Лучший вариант состоит в том, чтобы разделить ваши данные на два набора: обучающий набор данных и тестовый набор данных
- Тренируем модель с помощью обучающего набора и тестируем ее с помощью тестового набора
- Частота ошибок на новых примерах называется ошибкой обобщения (или ошибкой вне выборки), и, оценивая свою модель на тестовом наборе, вы получаете оценку этой ошибки.

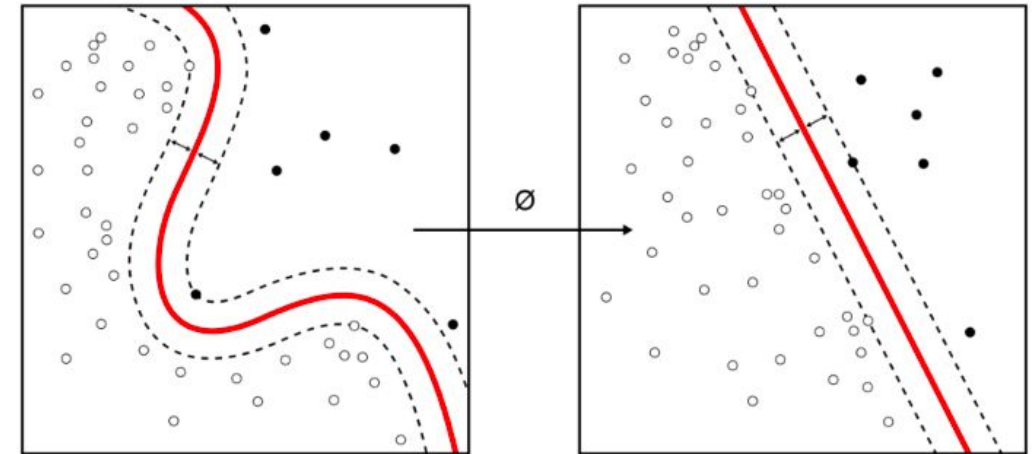
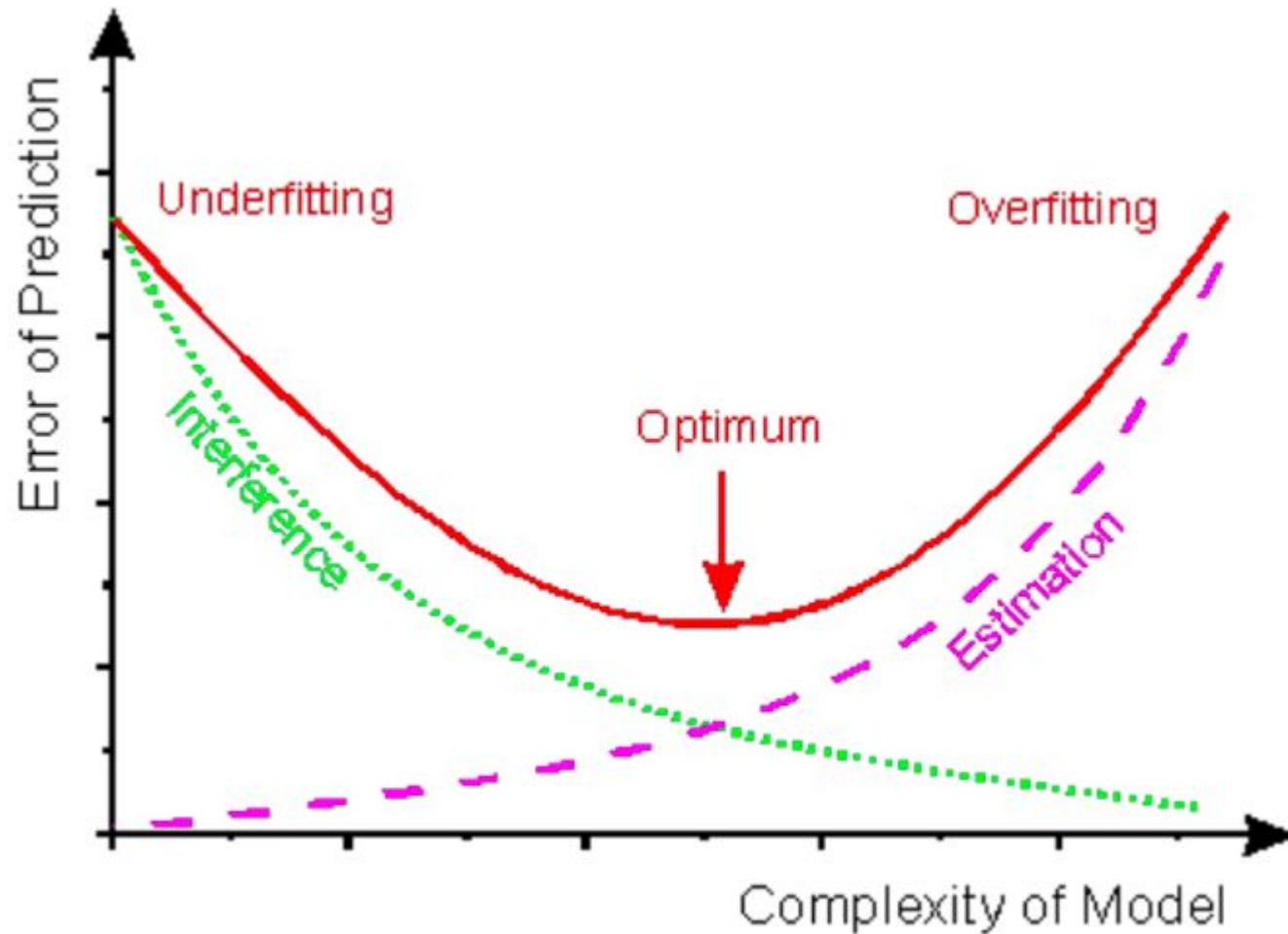


# С Какими трудностями сталкиваемся в процессе Машинного Обучения

<b>низкое качество данных</b>	<b>некорректное использование методов</b>	<b>субъективные метрики качества</b>
<b>нерелевантные признаки</b>	<b>нерепрезентативные данные</b>	<b>недообучение</b>
<b>переобучение</b>	<b>масштабирование метода с ростом числа пользователей/ запросов</b>	<b>недостаток данных</b>



# Идеальное обучение?



# Классификация

- Задача классификации
  - Имеется множество объектов (ситуаций), разделённых некоторым образом на классы.
  - Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется выборкой.
  - Классовая принадлежность остальных объектов неизвестна.
- Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества
- Типы классов
  - Двухклассовая классификация
  - Многоклассовая классификация
  - Непересекающиеся классы vs пересекающиеся классы
  - Нечёткие классы

# Метрики качества

- **Метрики**

- **точность Accuracy**
- **матрица расхождений Confusion matrix**
- **точность и полнота Precision and Recall**
- **F1-мера**
- **ROC-кривая ROC Curve**
- **ROC AUC (area under ROC curve)**

- **Почему Accuracy не лучшая метрика для классификации?**

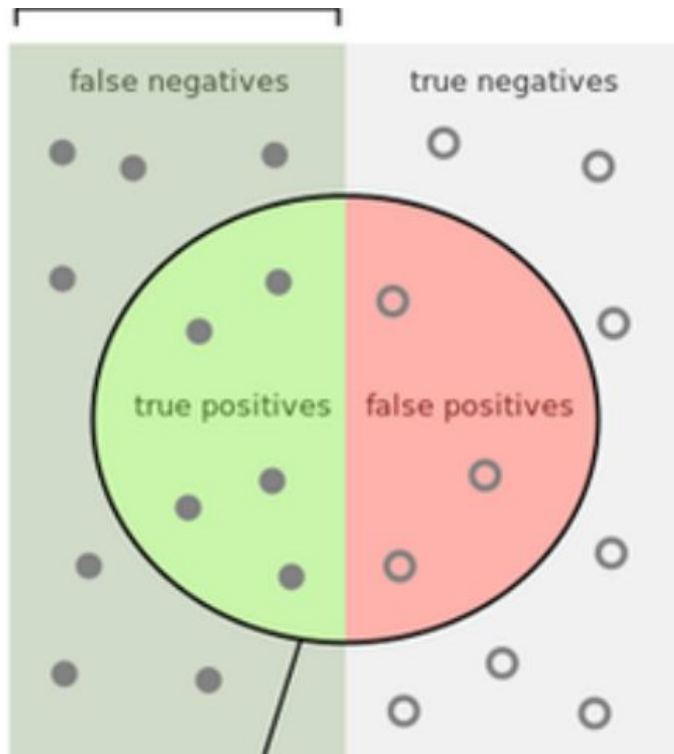
- **при несбалансированных классах 95 / 5**
- **Accuracy для “простого” классификатора будет 95%**

# Confusion матрица

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

- TN - true negatives
- TP - true positives
- FN - false negatives
- FP - false positives


# Recall и Precision



Сколько выбранных  
объектов корректны

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$


Как много  
корректных объектов  
выбрано?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$


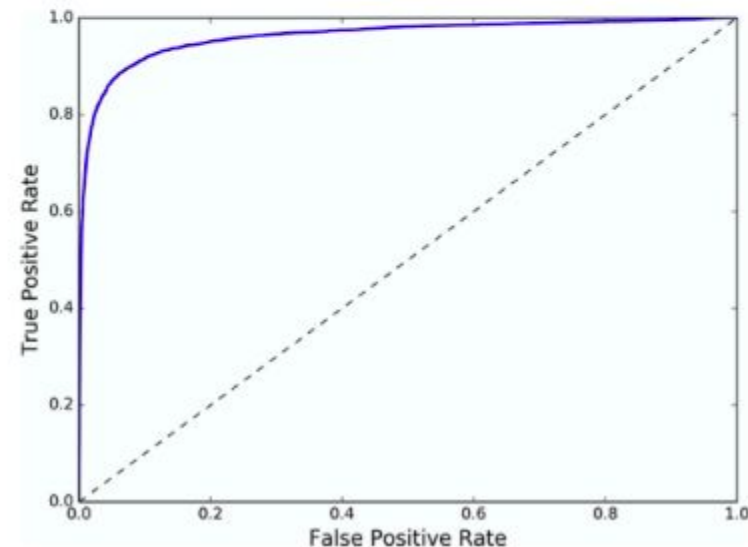
# F1 и ROC кривая

- F1-мера: гармоническое среднее точности и полноты (чувствительности)

- $$F_1 = \frac{2PR}{P + R} = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

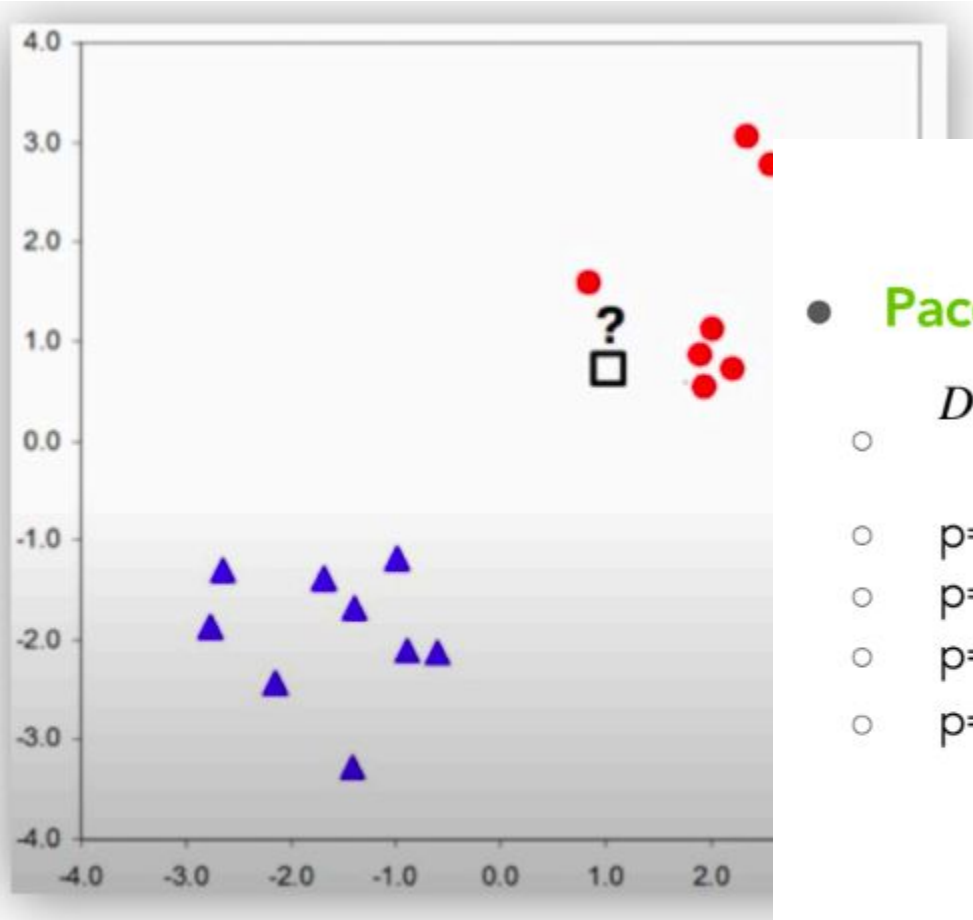
- ROC-кривая

- (ROC = receiver operating characteristic, рабочая характеристика приёмника)
- ROC AUC - площадь под ROC-кривой





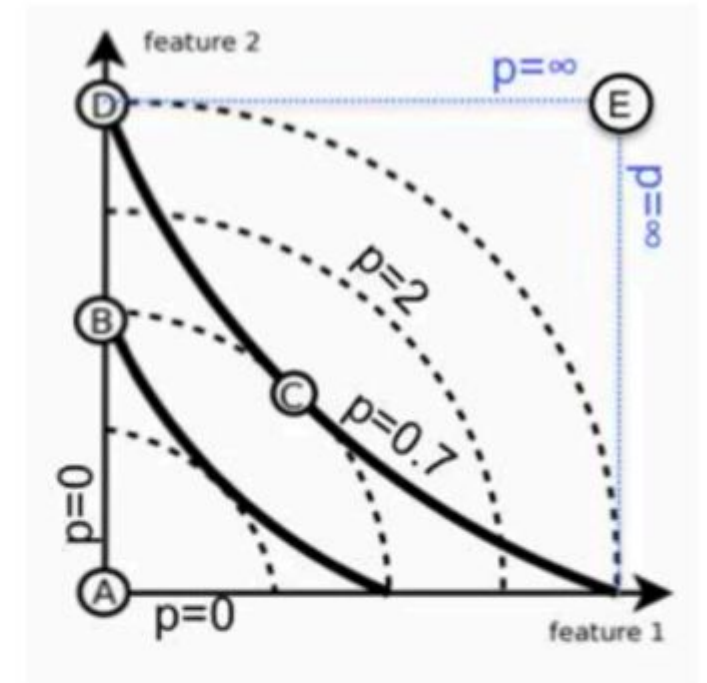
# KNN



## Расстояние Минковского

- $$D(\mathbf{x}, \mathbf{x}') = \sqrt[p]{\sum_i |x_i - x'_i|^p}$$

- $p=2$
- $p=1$
- $p=0$
- $p=\infty$



# Расстояние

## ● Расстояние Минковского

$$D(\mathbf{x}, \mathbf{x}') = \sqrt[p]{\sum_i |x_i - x'_i|^p}$$

- $p=2$
- $p=1$
- $p=0$
- $p=\infty$

