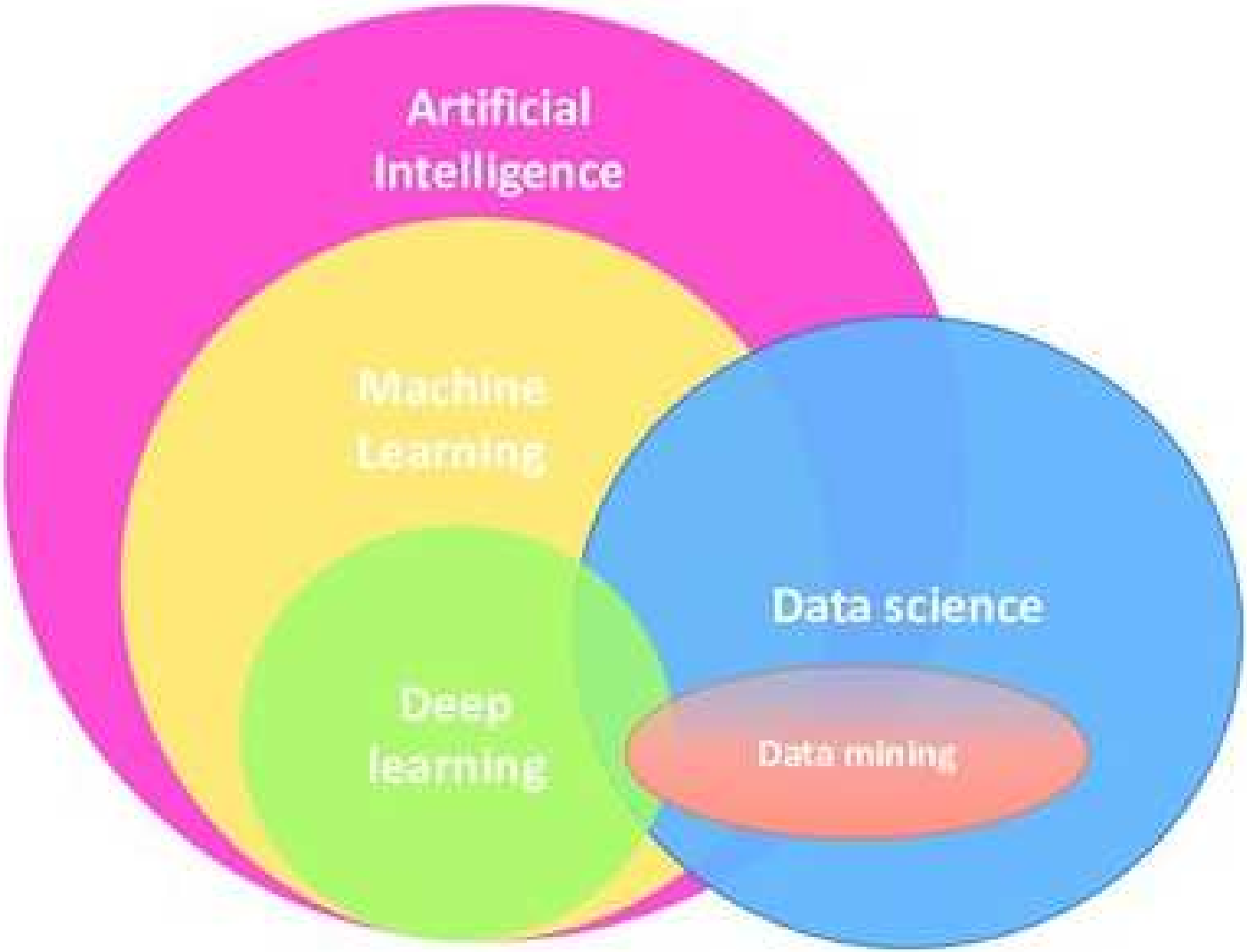
Нейросети и машинное обучение

Лектор: Постовалов Сергей Николаевич

s.postovalov@g.nsu.ru

Структура курса

- Машинное обучение
- Нейронные сети
  - Полносвязные
  - Сверточные
  - Реккурентные
- Применение нейронных сетей
  - Автоэнкодеры
  - Генеративно-состязательные сети
  - Идентификация личности
  - Машинный перевод
  - Обучение с подкреплением
  - Рекомендательные системы


Artificial Intelligence
Machine Learning
Deep learning
Data science
Data mining

Раздел 1. Машинное обучение

- Курс для магистрантов ММФ: «Математические методы анализа данных»
  - Лекции: Бериков В.Б.
  - Лабораторные работы и семинары: Постовалов С.Н., Неделько В.М.

1.1. Основные понятия

Пусть имеется N объектов, у каждого объекта есть n характеристик (feature, attribute, factor, variable) X_j.

D_j - множество значений X_j

Типы переменных:

1. Бинарный: $D_j = \{0,1\}$ или D_j={ *false*, *true*};
2. Категориальный: $D_j = \{b_1,\ldots,b_{l_j}\}$ – некоторое множество символов, $l_j \geq 2$.
3. Порядковый: $D_j = \{d_1,\ldots,d_{l_j}\}$, – множество упорядоченных значений.
4. Циклический: $D_j = \{d_1,\ldots,d_{l_j}\}$ - множество циклически повторяющихся значений;
   - угол [0,360]
5. Вещественный: $D_j = \mathbf{R}$.

- *Множество переменных* $X_1,...,X_n, \quad D_X = D_1 \times ... \times D_n$;
- Y *–целевая переменная ; область значений* Y: D_Y;

- *Вектор наблюдений* $x=x(o)=X_1(o),...,X_n(o), \; y=Y(o)$;
- *Таблица данных* $T=\{ x_j^{(i)}, y^{(i)} \}$, где

  $x_j^{(i)} = X_j(o^{(i)}), \; j=1,...,n, y^{(i)} = Y(o^{(i)}) \; i=1,...,N$;

  (N строк и n колонок):

  - Обучающая выборка – Y известно;
  - Контрольная выборка – Y неизвестно.

- Временной ряд – множество наблюдений в моменты времени $t^{(1)},...,t^{(N)}$.

Цель: предсказать Y в зависимости от X для произвольного объекта.

Feature Engineering

- Извлечение признаков из сырых данных
- Удаление неинформативных признаков
- Масштабирование (приведение значений к интервалу [0,1])
- Нормализация среднего (преобразование входных данных так, чтобы среднее было равно 0)
- Создание новых признаков

Основные типы задач

- *Y* - категориальная: классификация.
- *Y* - вещественная: регрессия.
- *Y* - вещественная, зависящая от времени: предсказание временного ряда;
- Y - категориальная, неизвестна: кластерный анализ, классификация без учителя.

- Обучение с подкреплением (reinforcement learning)

Основные шаги

1. Описание переменных, цель
2. Сбор и очистка данных;
3. Построение модели (зависимости Y от X);
4. Проверка качества модели;
5. Использование модели для предсказания.
6. Обучение с подкреплением: на шаг 2

Качество решения

$L(f(x), y)$ - *функция потерь (loss function)*

Индикаторная функция потерь

$$l_{i,j} = \mathbf{I}(i \neq j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}.$$

$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} L(f(x^{(i)}), y^{(i)})$ - эмпирический риск.

Для индикаторной функции

$$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}(f(x^{(i)}) \neq y^{(i)}) = P_{emp}$$

- доля неправильно классифицированных объектов.

Для вещественной целевой переменной $Y \in R$
и некоторой решающей функции

$$y = f(x).$$

$e^{(i)} = y^{(i)} - f(x^{(i)})$ - ошибка для i-го наблюдения.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |e^{(i)}|;$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (e^{(i)})^2;$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (e^{(i)})^2}.$$

Бинарная задача классификации

$$D_Y = \{True, False\} = \{+, -\},$$

.

Confusion matrix

True class	Predicted class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

accuracy = (TP+TN)/(TP+TN+FP+FN);

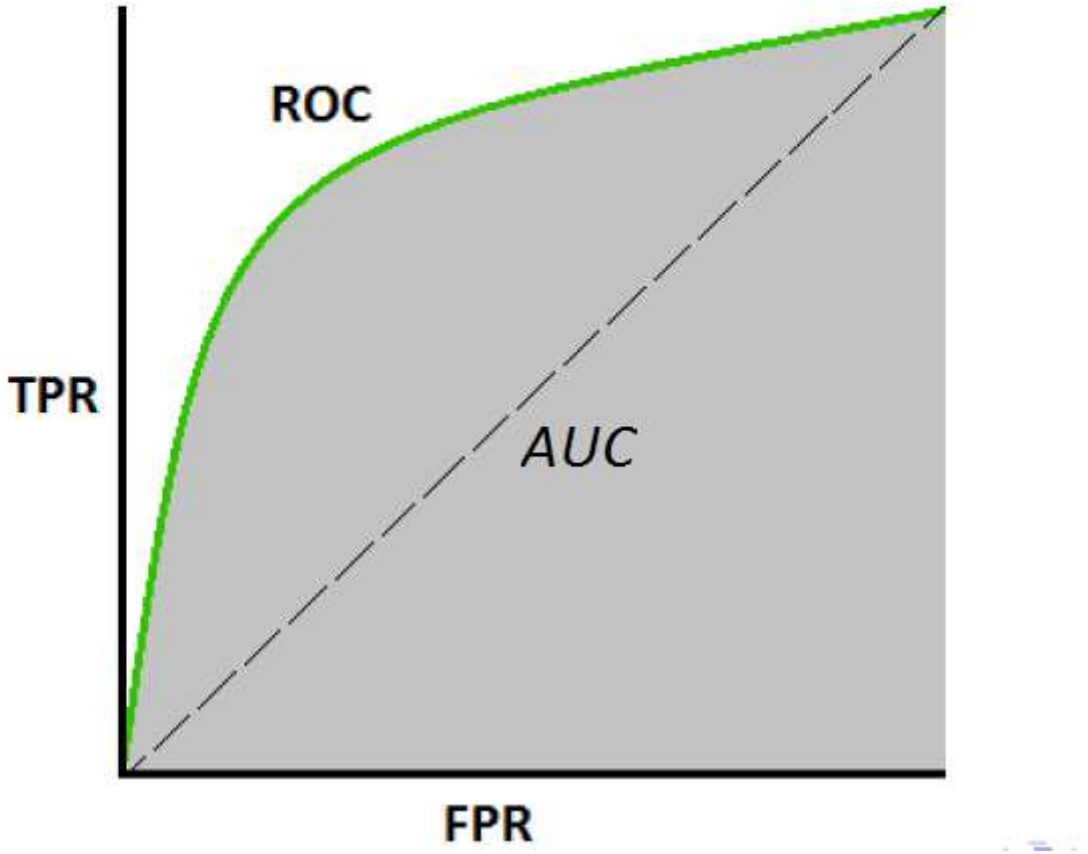
recall = TP/(TP+FN) (true positive rate, TPR, sensitivity, полнота);

precision = TP/(FP+TN);

F1= 2•(precision·recall)/(precision+recall)
гармоническое среднее от precision и recall.
FPR= FP/(TN+FP) (false positive classification rate)

ROC-curves analysis (Receiver Operating Characteristic)

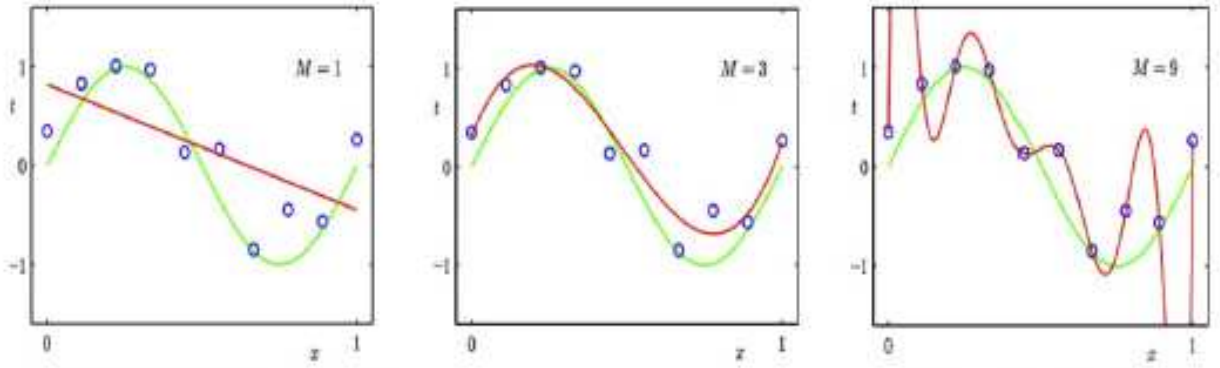
$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN}$$



Проблема недо- и пере- обучения

Under- and Over-fitting examples

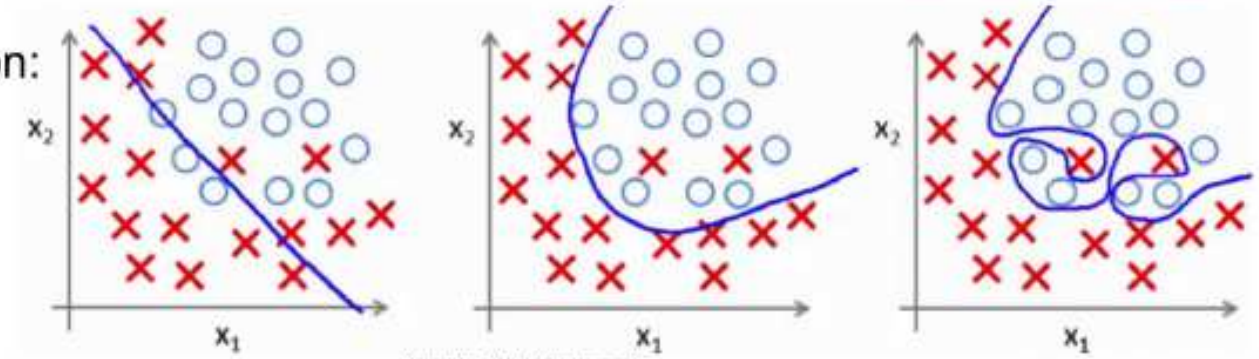
Regression:



predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:



Как обнаружить переобучение?

Holdout method.

Разбиение выборки на обучающую и контрольную.

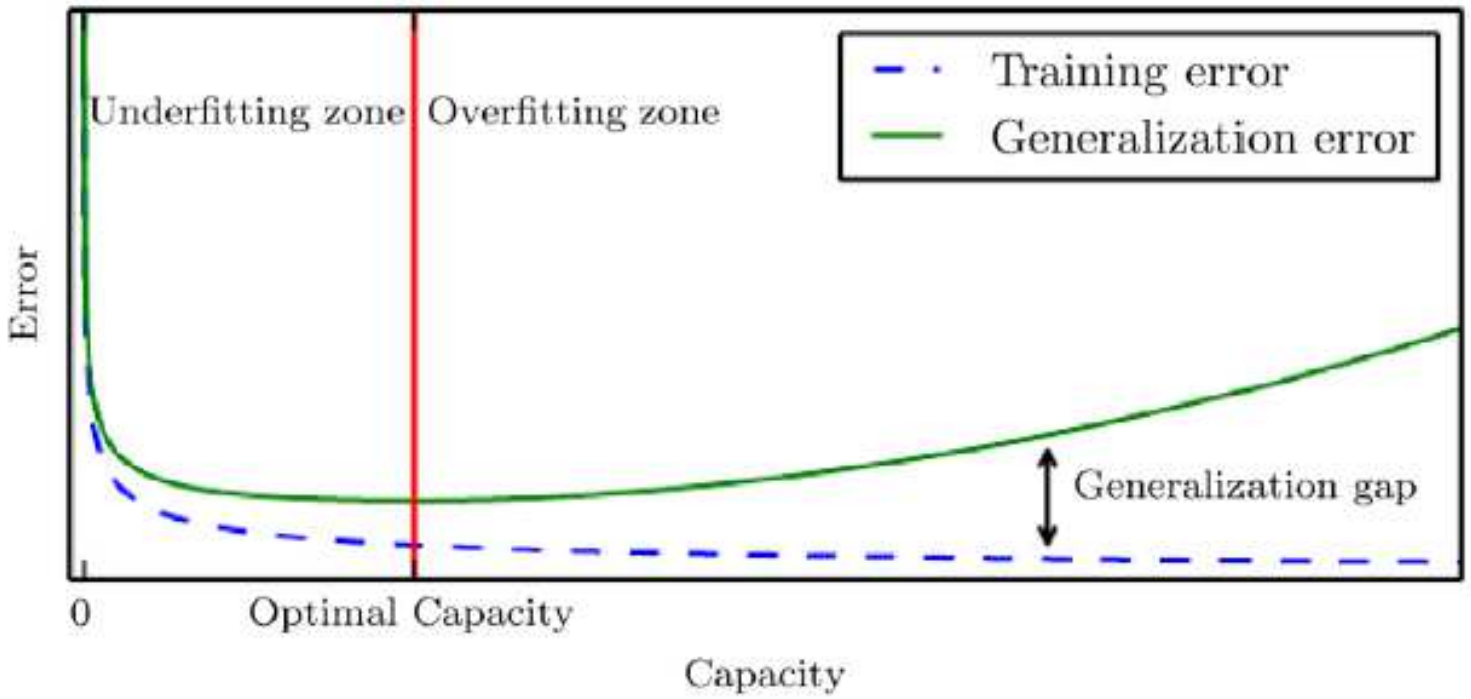
One-leave-out method.

- Один объект вынимается из выборки, а по оставшимся делается построение модели.
- Делается оценка качества предсказания по этому объекту
- Объект возвращается и вынимается другой.
- Процедура повторяется N раз.
- Оценка качества усредняется

L-fold cross-validation.

- Выборка делится на L частей
- По L-1 части делается построение модели, а по оставшейся части делается оценка качества модели.
- Процедура повторяется L раз
- Оценка качества усредняется

Как решить проблему переобучения?



Основные подходы к решению задач машинного обучения

- Вероятностный
  - параметрический
  - непараметрический
- Геометрический
  - Метод ближайшего соседа
  - Метод опорных векторов
- Логический
  - Конъюнкции предикатов
  - Дерево решений
- Кибернетический
  - Нейронные сети