Основные понятия и задачи машинного обучения

Виктор Владимирович Китов

v.v.kitov@yandex.ru
victorkitov.github.io

Содержание

Актуальность

- "Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, искусственном интеллекте и машинном обучении." Клаус Мартин Шваб, президент Всемирного экономического форума (2016).

Идея машинного обучения

- Машинное обучение - область знания, позволяющая компьютерам обучаться принимать сложные решения без явной спецификации алгоритма принятия решения.
  - алгоритм находится в большом параметризованном семействе моделей.

- Компьютер учится на заданном опыте решать некоторый класс задач, относительно некоторого показателя качества, если показатель качества растет на классе задач после получения опыта.

Примеры

- Фильтрация спама
  - если отправитель осуществляет массовую рассылку, на которую пользователи не отвечают, а тело письма содержит ключевые слова "уникальное предложение" и "приобретите сейчас" -> спам
  - если пользователь уже отвечал на письмо -> не спам
  - если в подписи письма отправитель и получатель из одной организации -> не спам
  - ...
- Разметка частей речи.

  - если слово заканчивается на "еть"-> глагол (смотреть, хотеть)
  - если предыдущее слово "к" -> существительное (к солнцу, к делу)
  - ...

Ручной подход и машинное обучение

Ручной подход:

- сложно найти специалистов
- дорого
- медленно
- неточно (только простые правила)

Эти проблемы решает использование машинного обучения.

- но нужны ресурсы для формализованного описания опыта.

Где машинное обучение дает преимущество

- **сложно сформулировать явную зависимость**
  - слишком много наблюдений
    - логи сайта
  - слишком много признаков
    - категоризация текстов
  - сложные взаимосвязи признаков
    - классификация изображений
- **нужна быстрая скорость адаптации** к изменяющимся условиям
  - предсказание цен на акции
- **необходимо построить много моделей**
  - свою под каждого пользователя голосового помощника

Формальные определения

- Есть класс объектов Z
- Каждый объект описывается вектором известных характеристик (признаков) $\mathbf{x} \in \mathcal{X}$ и предсказываемых характеристик (откликов) $y \in \mathcal{Y}$.

$$z = (\mathbf{x}, y) \in Z$$

- Задача: найти отображение f, которое бы точно описывало взаимосвязь $\mathcal{X} \to \mathcal{Y}$.
  - используя конечный набор известных пар $(\mathbf{x}, y)$ обучающей выборки.
  - для применения к любым новым $\mathbf{x}$ в тестовой выборке.
- Тестовая выборка может быть известна или неизвестна заранее.

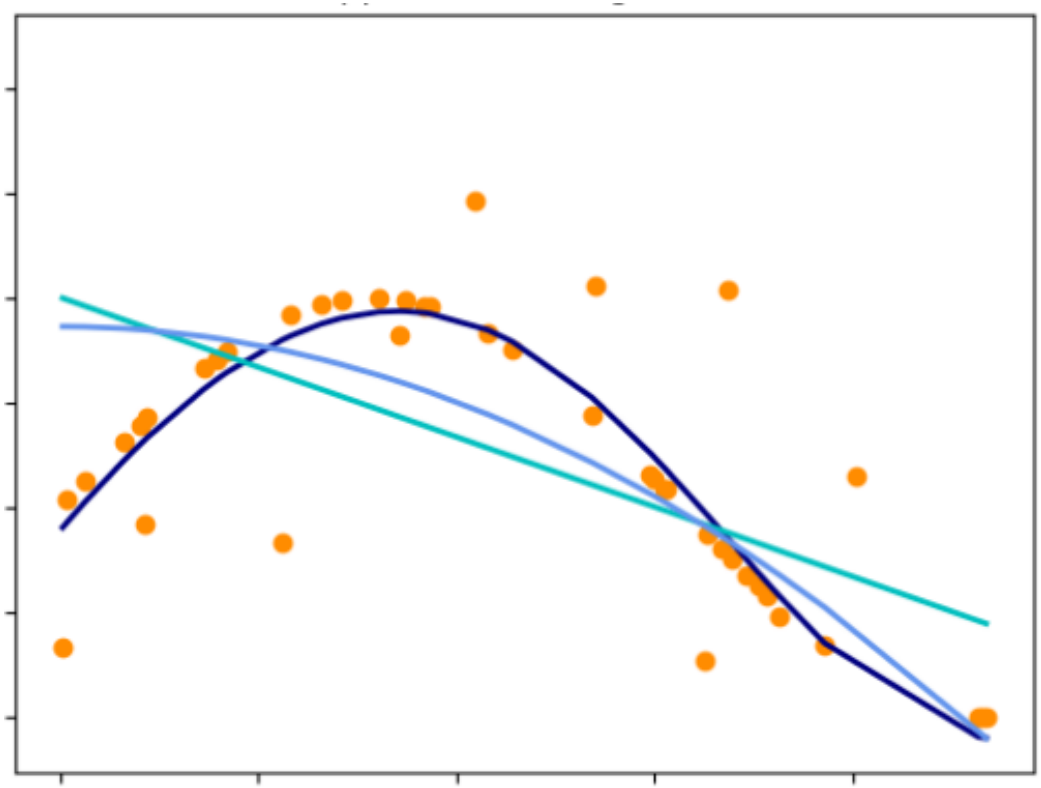
Основные типы признаков

- Входное описание объекта $\mathbf{x} \in \mathcal{X}$ состоит из индивидуальных признаков $x^i \in \mathcal{X}_i$.
- Типы каждого признака (например, в задаче кредитного скоринга):
  - $\mathcal{X}_i = \mathbb{R}$ - вещественный (количественный) признак
    - например, возраст, зарплата.
  - $\mathcal{X}_i = \{0, 1\}$ - бинарный признак
    - пример: есть ли у должника просрочки по платежам?
  - $|\mathcal{X}_i| < \infty$ - дискретный категориальный признак
    - пример: профессия.
  - $|\mathcal{X}_i| < \infty$ и $\mathcal{X}_i$ упорядоченный дискретный (порядковый) признак
    - пример: уровень образования.

Возможные постановки задачи

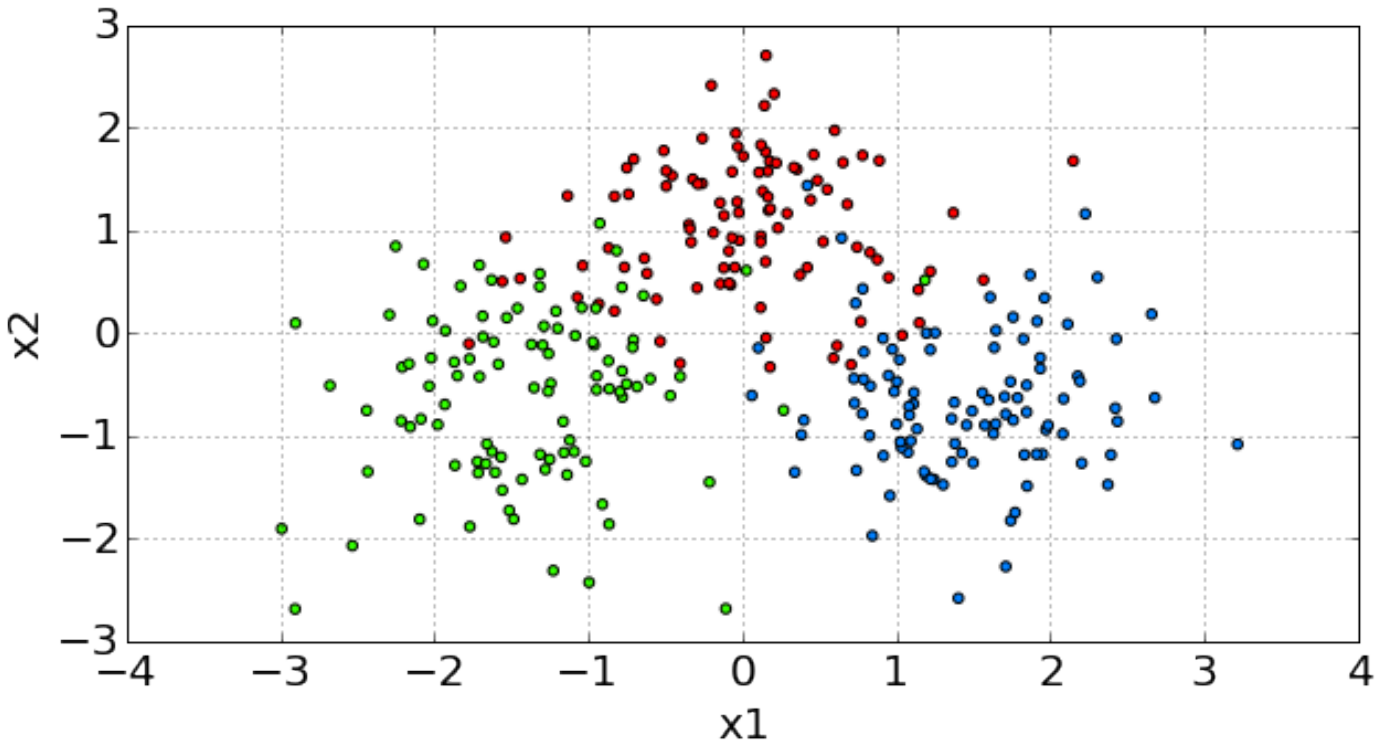
- **Обучение с учителем** (supervised learning):
  - обучающая выборка: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_N, y_N)$
  - **трансдуктивное обучение**: входы x тестовой выборки известны заранее.
- **Обучение без учителя** (unsupervised learning):
  - обучающая выборка: $\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N$
- **Частичное обучение** (semi-supervised learning):
  - обучающая выборка: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_N, y_N)$, $\mathbf{x}_{N+1}$, $\mathbf{x}_{N+2}, ...\mathbf{x}_{N+M}$
- **Обучение с подкреплением** (reinforcement learning):
  - обучающая выборка $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_N, y_N)$ формируется динамически
  - зависит от предыдущих действий агента
  - применяется, например, в робототехнике.

Обучение с учителем - регрессия



Предсказать $y \in \mathbb{R}$ для любого x.

Обучение с учителем - классификация



Предсказать дискретный y, обозначенный цветом, в каждой точке.

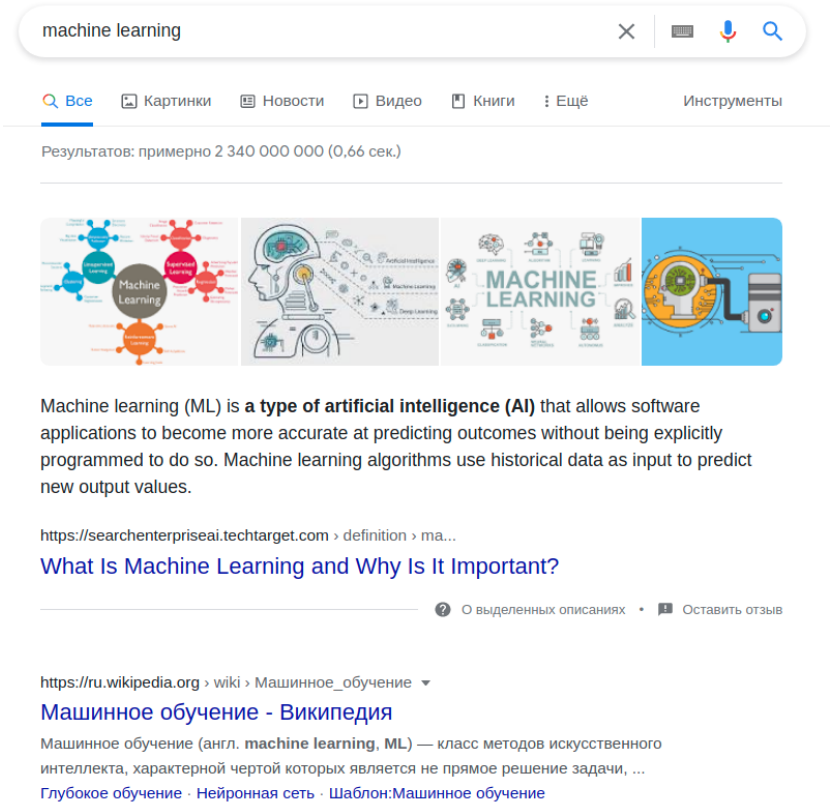
Пример: медицинские приложения

- **объекты**: пациенты
- **признаки**:
  - вещественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза лекарства.
  - бинарные: пол, наличие головной боли, слабости, тошноты
  - категориальные: перенесенные болезни
  - порядковые: тяжесть состояния
- **возможные отклики**:
  - классификация: определить тип болезни, способ лечения.
  - регрессия: длительность лечения и выздоровления.

Пример: прогнозирование поведения клиентов

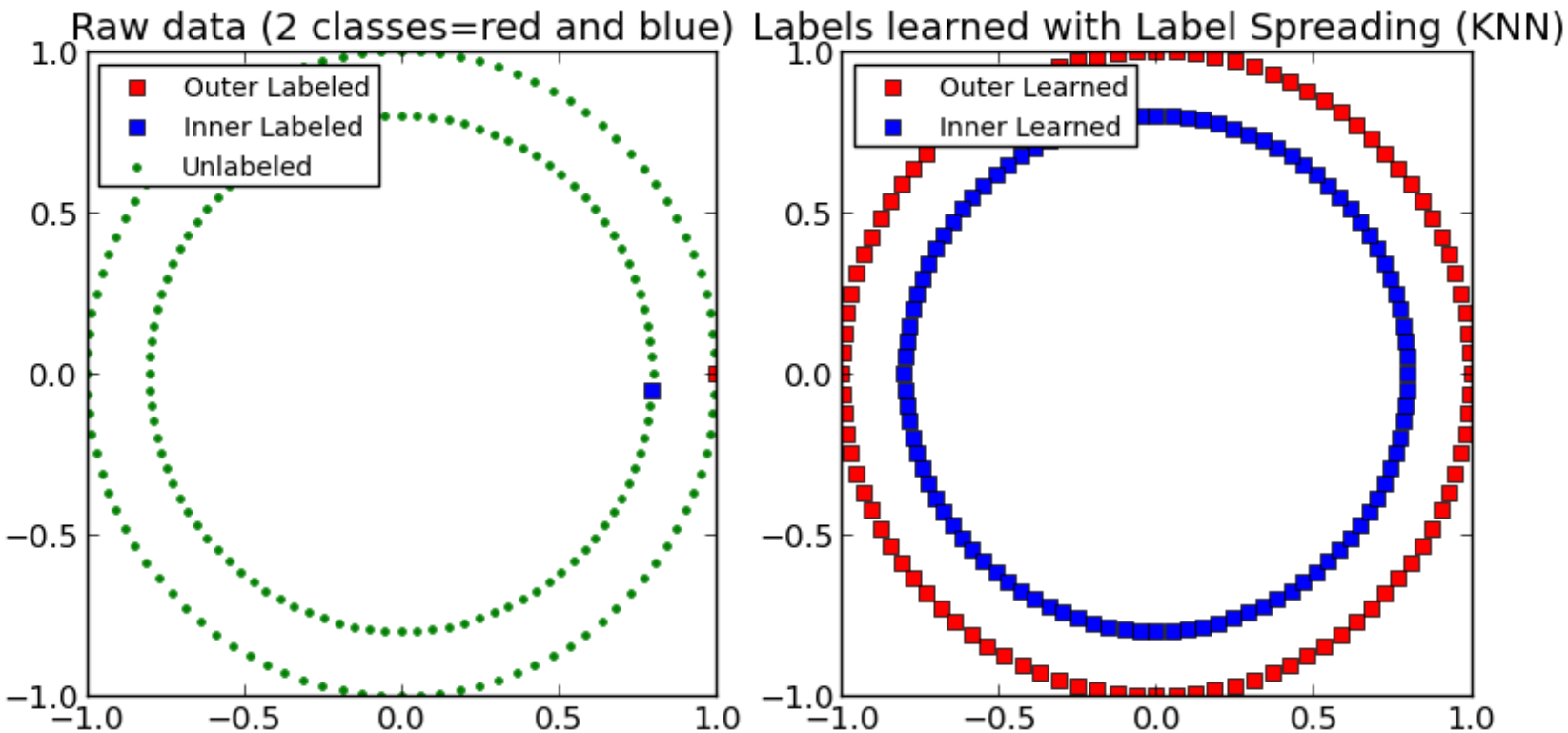
- **объекты**: клиент в текущий момент времени.
- **признаки**:
  - вещественные: возраст, историческая частота пользования услугами и траты на услуги.
  - бинарные: пол, были ли задолженности по платежам.
  - категориальные: какими услугами пользуется.
  - порядковые: оценка компании по мнению клиента.
  - возможные отклики:
- **классификация**: уйдет ли клиент к конкурентам? подключит ли услугу?
- **регрессия**: сколько раз воспользуется услугой? сколько денег внесет на счёт?

Обучение с учителем - ранжирование



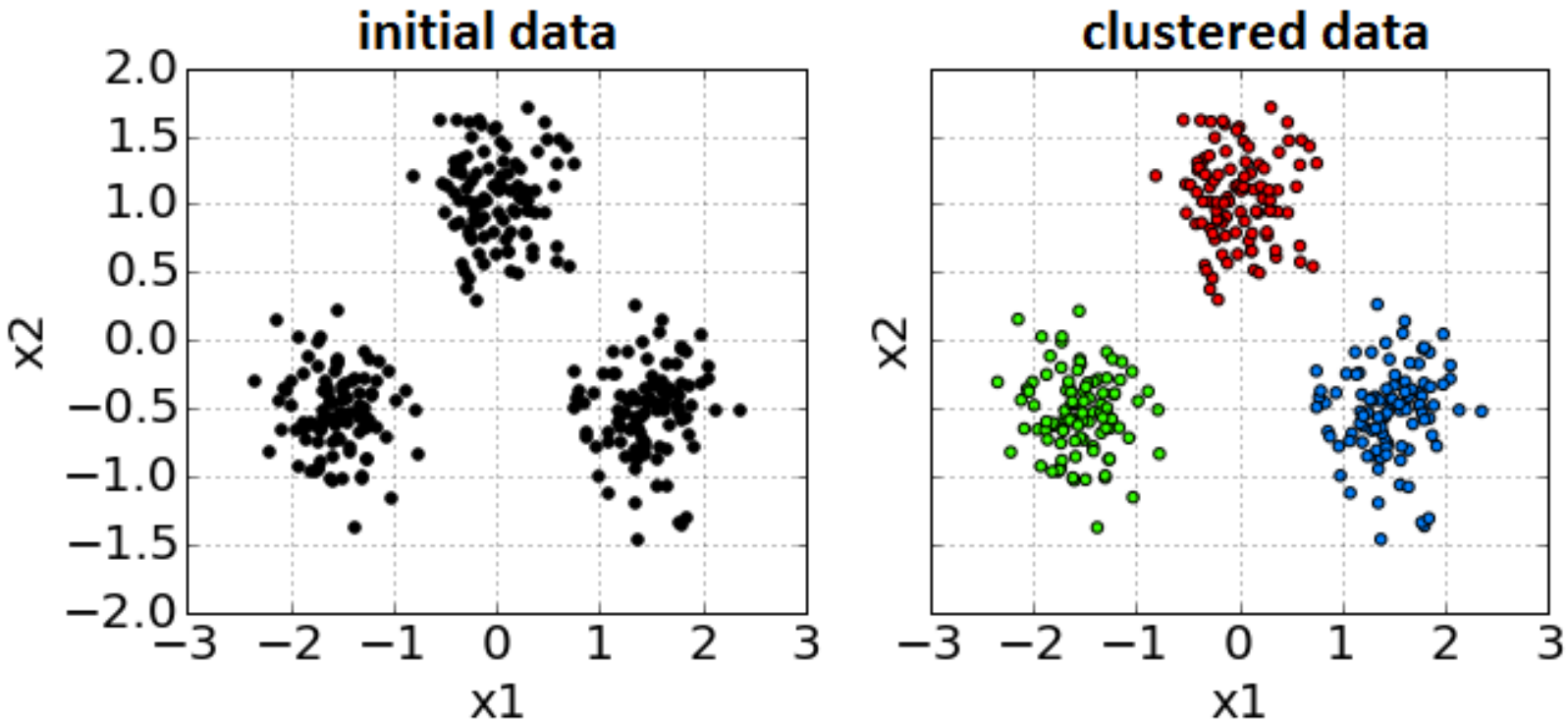
Как можно было бы решить задачу ранжирования через регрессию или классификацию?

Частичное обучение - классификация



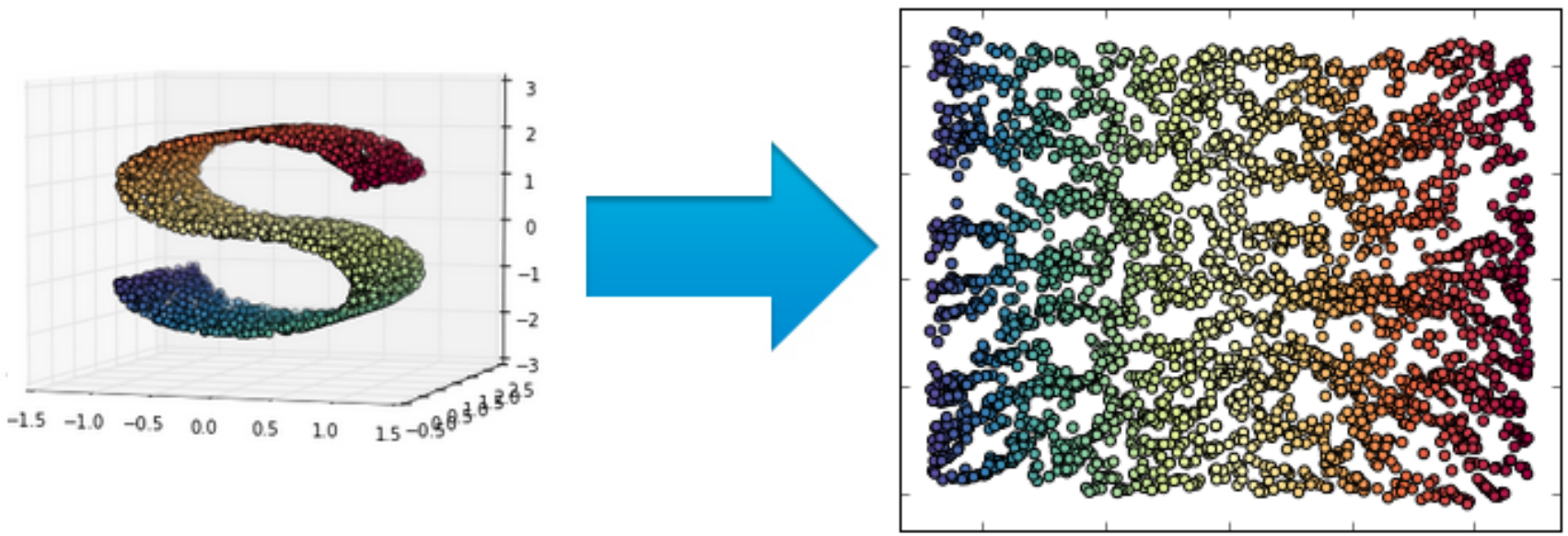
Предположение-близкие точки принадлежат одному классу.

Обучение без учителя - кластеризация



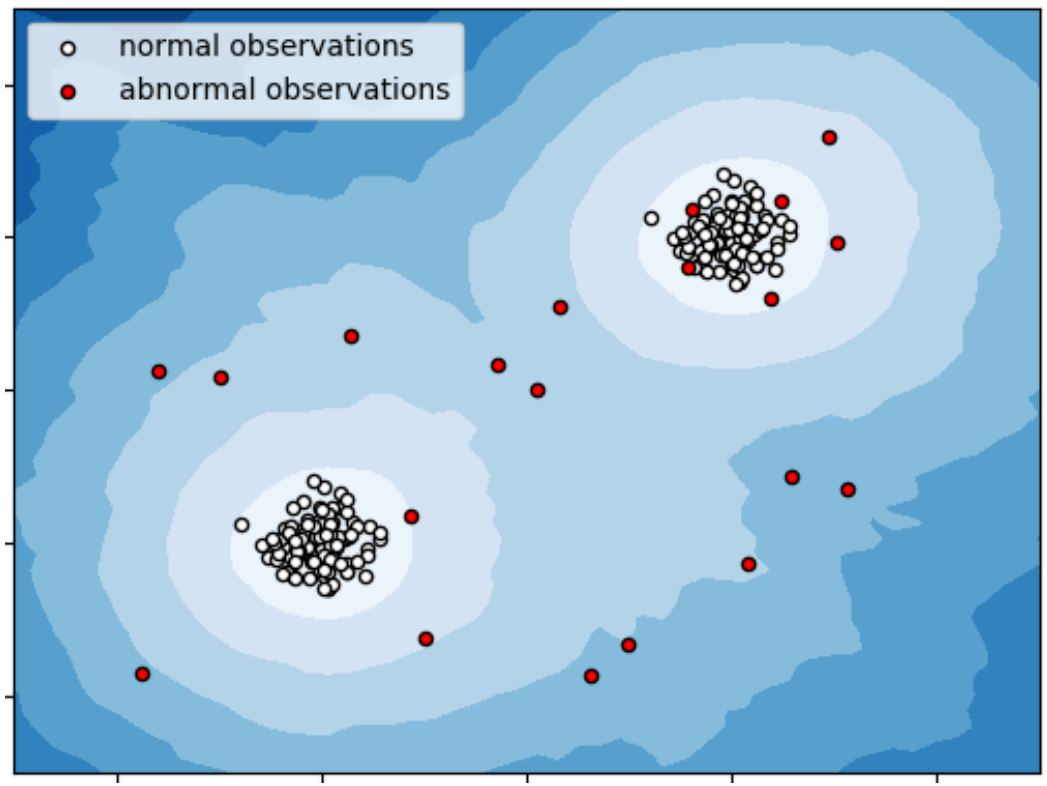
Разбивка объектов на похожие группы.

Обучение без учителя - снижение размерности



Переход из 3D в 2D с минимальным искажением геометрии.

Обучение без учителя - обнаружение аномалий



Выделение нетипичных объектов.

Поиск ассоциативных правил

rhs	lhs	support	confidence	lift
bottled beer	liquor, red/blush wine	0.0016268429	0.9411765	19.53
	liquor, red/blush wine, soda	0.0006100661	1.0000000	20.75
	liquor, soda	0.0010167768	0.7692308	15.96
bottled water	bottled beer, misc. beverages	0.0005083884	0.5000000	8.29
citrus fruit	meat, turkey	0.0004067107	0.5714286	6.97
coffee	condensed milk, sugar	0.0004067107	0.8000000	25.71
frankfurter	liver loaf, sausage	0.0005083884	0.5000000	8.48
liquor	bottled beer, red/blush wine	0.0016268429	0.6666667	81.96
	bottled beer, red/blush wine, s...	0.0006100661	1.0000000	122.94
	red/blush wine, soda	0.0006100661	0.5454545	67.06

Анализ потребительских корзин (market basket analysis)

По наборам множеств $\{a, b, c\}, \{a, d, e\}, \{a, b\}, \{a, b, g, h\}$ генерировать правила: $a \to b,\ \ b \to a, \ldots$

Обучение с учителем - задача

- Требуется найти отображение $f(x) : X \to Y$.
- Варианты использования:
  - предсказание y по x.
  - анализ зависимости $X \to Y$ на качественном уровне
  - обнаружение нетипичных объектов (где модель ошибается)
- Вопросы в настраивании модели:
  - какую целевую переменную y нужно прогнозировать?
  - какие использовать признаки x?
  - в каком классе искать отображение f?
  - в каком смысле отображение f должно приближать зависимость $X \to Y$?
  - как алгоритмически подбирать параметры f?

Основные типы откликов

- $\mathcal{Y} = \mathbb{R}$ - регрессия
  - например, цена акции
- $\mathcal{Y} = \mathbb{R}^M$ - векторная регрессия
  - например, динамика цен на квартиры
- $\mathcal{Y} = \{\omega_1, \omega_2, ...\omega_C\}$ - классификация.
  - C=2: бинарная классификация.
    - например, спам/не спам
  - C>2: многоклассовая классификация
    - например, идентификация пользователя
- $\mathcal{Y}$ - множественная классификация из $\{\omega_1, \omega_2, ...\omega_C\}$.[1]
  - например, категоризация новостей

[1] Можно ли ее решить используя обычную классификацию?

Содержание

Пример линейного класса функций.

- **Регрессия**: $\widehat{y} = g(x), \quad g(x)$ параметризовано θ.

[2] Однозначно ли определены дискриминантные ф-ции?

Пример линейного класса функций.

- **Регрессия**: $\widehat{y} = g(x), \quad g(x)$ параметризовано θ.
- **Многоклассовый классификатор** $(y \in \{1, 2, ...C\})$[2]:

$$\widehat{y}(x) = \arg\max_c g_c(x), \quad g(x) \text{ параметризовано } \theta.$$

$$\{x : g_i(x) = g_j(x)\}, \quad \text{граница между классами } i, j.$$

$$M(x, y) = g_y(x) - \max_{c \neq y} g_c(x), \quad \text{отступ (качество классификации)}$$

[2]Однозначно ли определены дискриминантные ф-ции?

Пример линейного класса функций.

- **Регрессия**: $\widehat{y} = g(x), \quad g(x)$ параметризовано θ.
- **Многоклассовый классификатор** $(y \in \{1, 2, ...C\})^2$:

$$\widehat{y}(x) = \arg\max_c g_c(x), \quad g(x) \text{ параметризовано } \theta.$$

$$\{x : g_i(x) = g_j(x)\}, \quad \text{граница между классами } i, j.$$

$$M(x, y) = g_y(x) - \max_{c \neq y} g_c(x), \quad \text{отступ (качество классификации)}$$

- **Бинарный классификатор** $(y \in \{+1, -1\})$:

$$\widehat{y}(x) = \arg\max_{c \in \{+1,-1\}} g_c(x) = \text{sign}\left(g_{+1}(x) - g_{-1}(x)\right) = \text{sign}\left(g(x)\right)$$

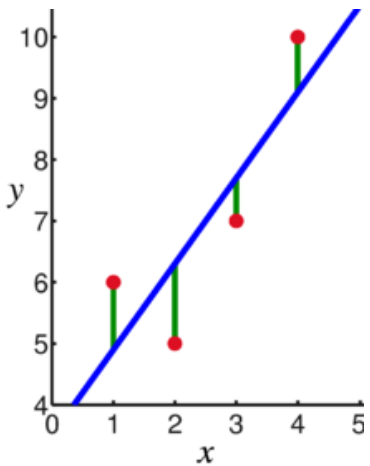
$$M(x, y) = g_y(x) - g_{-y}(x) = y\left(g_{+1}(x) - g_{-1}(x)\right) = yg(x)$$

[2] Однозначно ли определены дискриминантные ф-ции?

Примеры

линейная регрессия $y \in \mathbb{R}$:

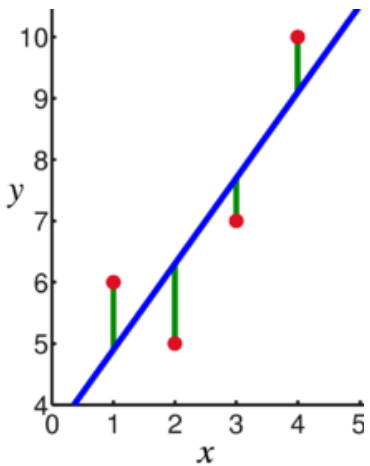
$$f(x|\theta) = \theta_0 + \theta_1 x$$



Примеры

линейная регрессия $y \in \mathbb{R}$:

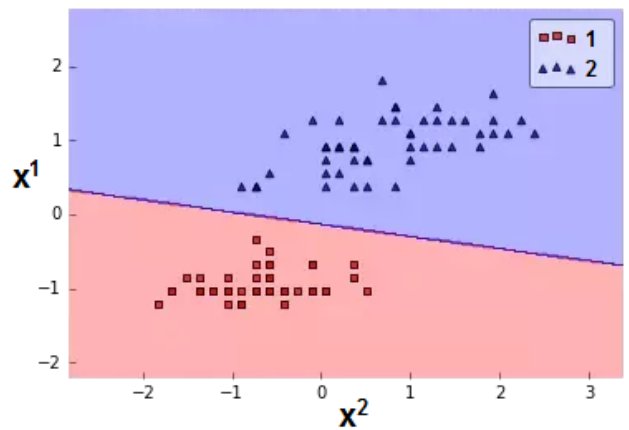
$$f(x|\theta) = \theta_0 + \theta_1 x$$



линейная классификация $y \in \{1, 2\}$:

$$g_c(\mathbf{x}|\theta) = \theta_c^0 + \theta_c^1 x^1 + \theta_c^2 x^2,\ c = 1, 2.$$

$$f(\mathbf{x}|\theta) = \arg\max_c g_c(x|\theta)$$



Функция качества / потерь

- Точность предсказаний может оцениваться:
  - **критерием качества** (score function, выше->лучше)
  - **функцией потерь** (loss function, ниже->лучше)
- $loss = F(score)$, $score = F^{-1}(loss)$ для некоторой убывающей $F(\cdot)$.
  - например, ф-ция потерь = - ф-ция качества.
- loss=loss(ошибка), score=score(ошибка), где ошибка:
  - регрессия: $(\widehat{y} - y)$
  - классификация: $-M(x, y)$.

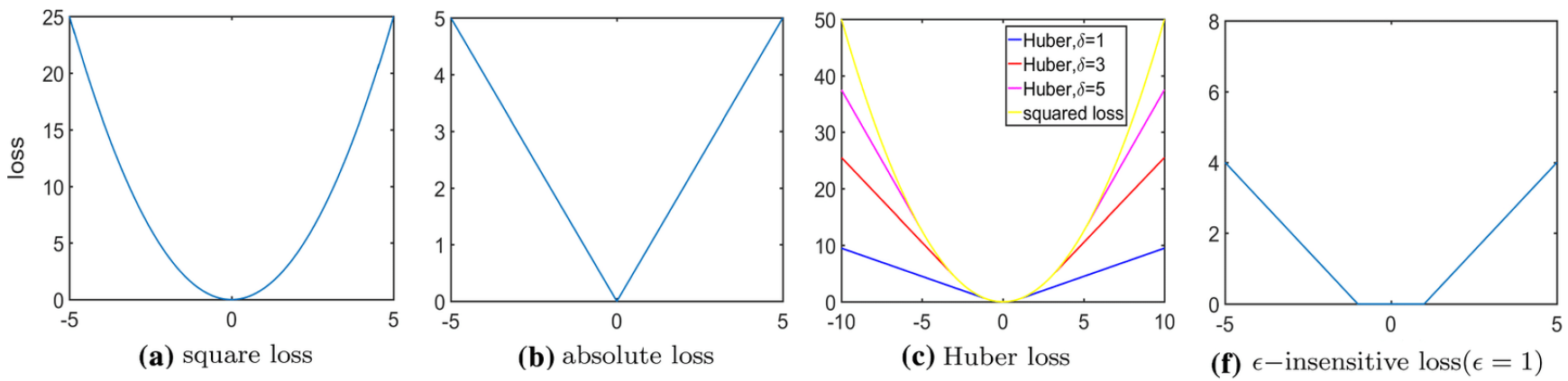
Функции потерь регрессии $F(\widehat{y}-y)$[3]

MAE: $|\widehat{y}-y|$

MSE: $(\widehat{y}-y)^2$

Huber $\begin{cases} \frac{1}{2}(\widehat{y}-y)^2, & |\widehat{y}-y| \leq \delta \\ \delta\left(|\widehat{y}-y|-\frac{1}{2}\delta\right) & |\widehat{y}-y| > \delta \end{cases}$

ε-insensitive $\max\{|\widehat{y}-y|-\varepsilon, 0\}$



(a) square loss **(b)** absolute loss **(c)** Huber loss **(f)** $\epsilon-$insensitive loss($\epsilon = 1$)

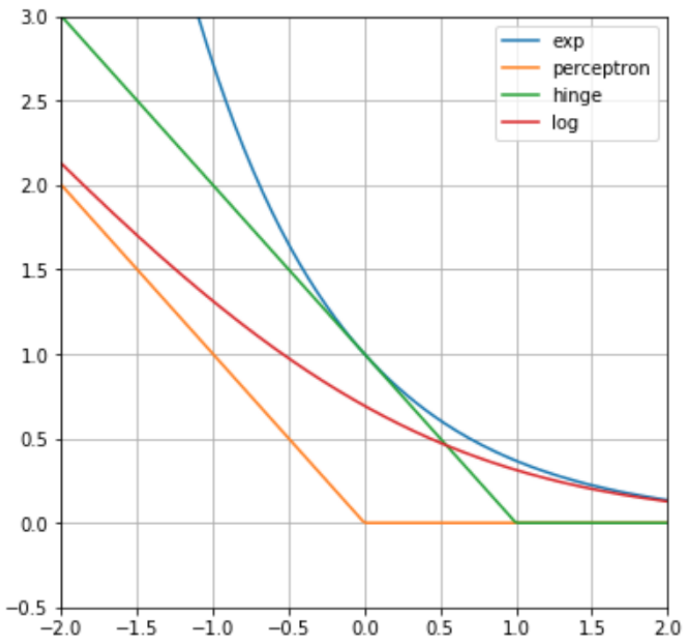
[3]Выбор ф-ции потерь на практике должен исходить от бизнес-задачи.

Функции потерь классификации $F(M)$

- $Loss = F(M)$ для некоторой убывающей $F(\cdot)$ [выше отступ->лучше].

$$\mathcal{L}_{exp}(M) = e^{-M} \quad \mathcal{L}_{perceptron}(M) = [-M]_+$$

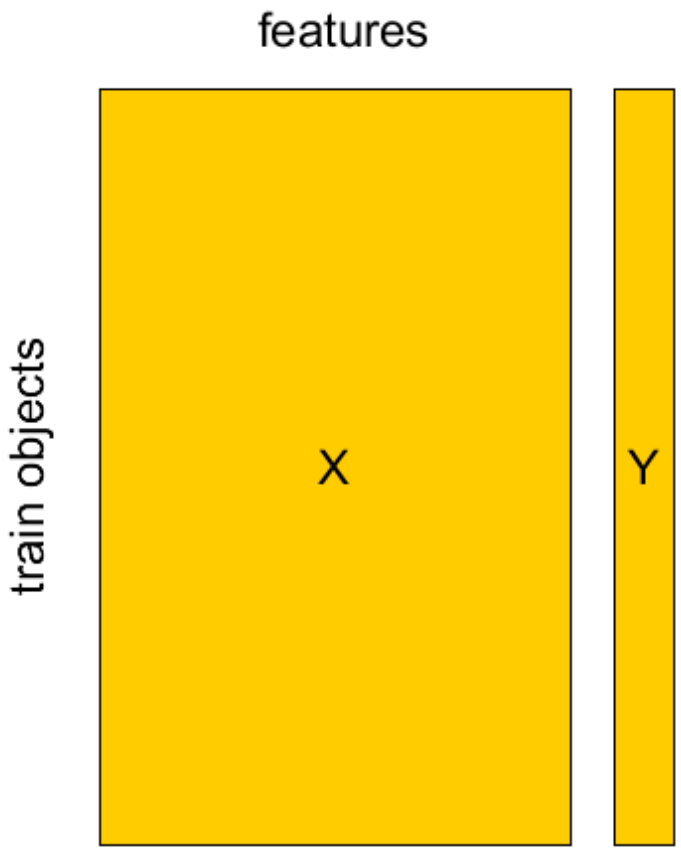
$$\mathcal{L}_{hinge}(M) = [1 - M]_+ \quad \mathcal{L}_{log}(M) = \ln\left(1 + e^{-M}\right)$$



Содержание

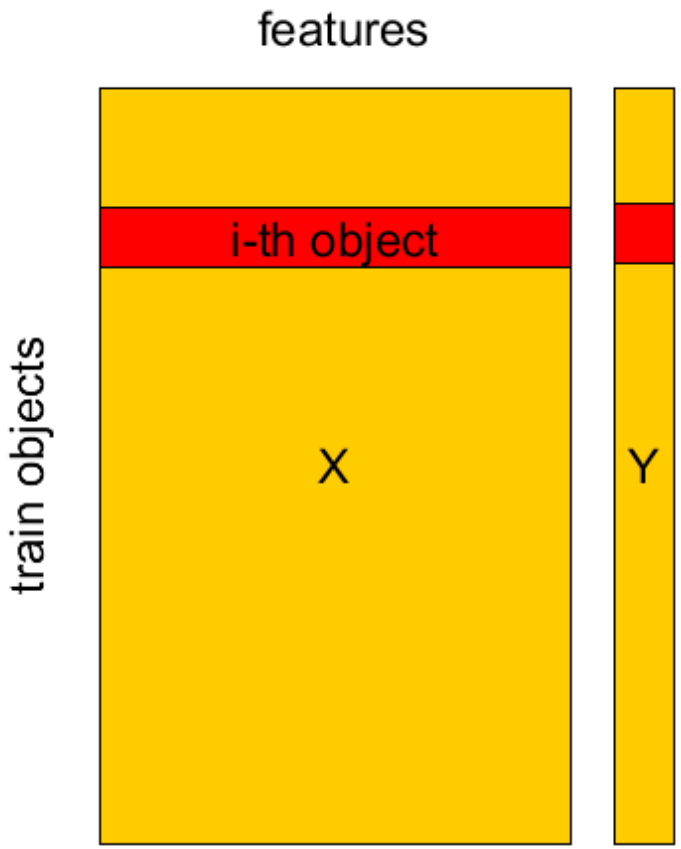
Обучающая выборка

Обучающая выборка (training set): $(\mathbf{x}_1, y_1), ...(\mathbf{x}_M, y_M)$,
Матрица признаков (design matrix) $X = [\mathbf{x}_1, ...\mathbf{x}_M]^T$, отклики (targets) $Y = [y_1, ...y_M]^T$.



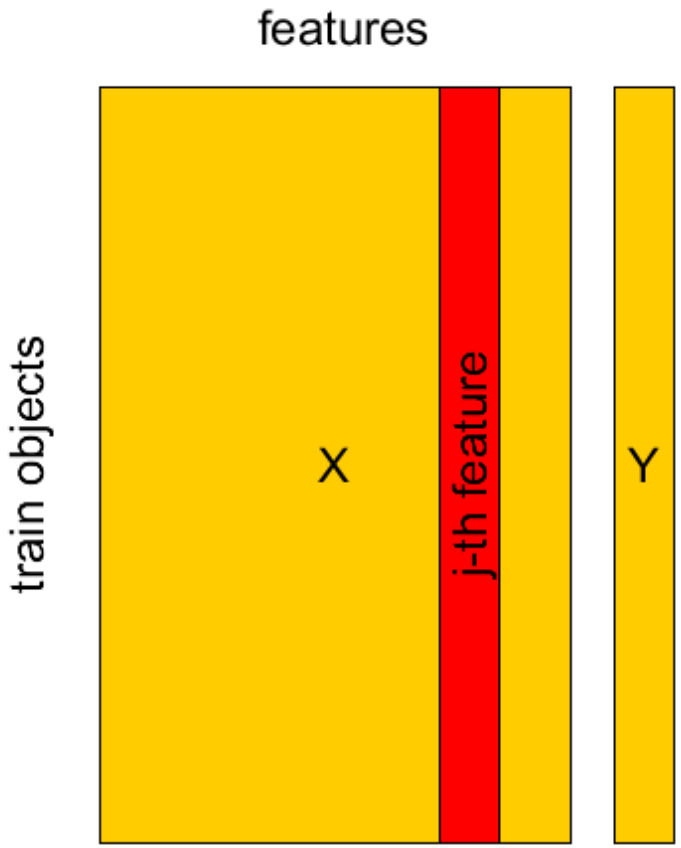
Объекты

Объект соответствует строке в матрице признаков:

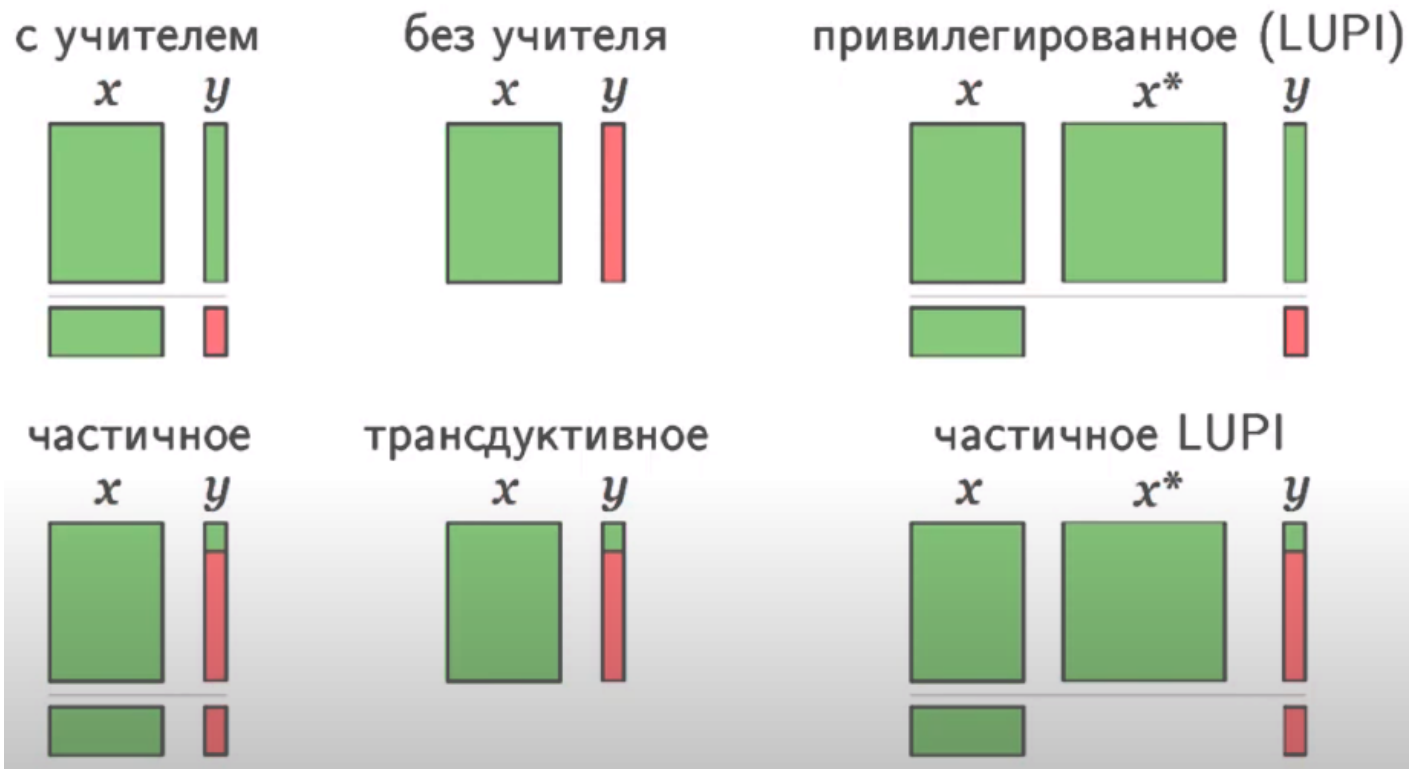


Признаки

Признак соответствует столбцу в матрице признаков:



Виды обучения

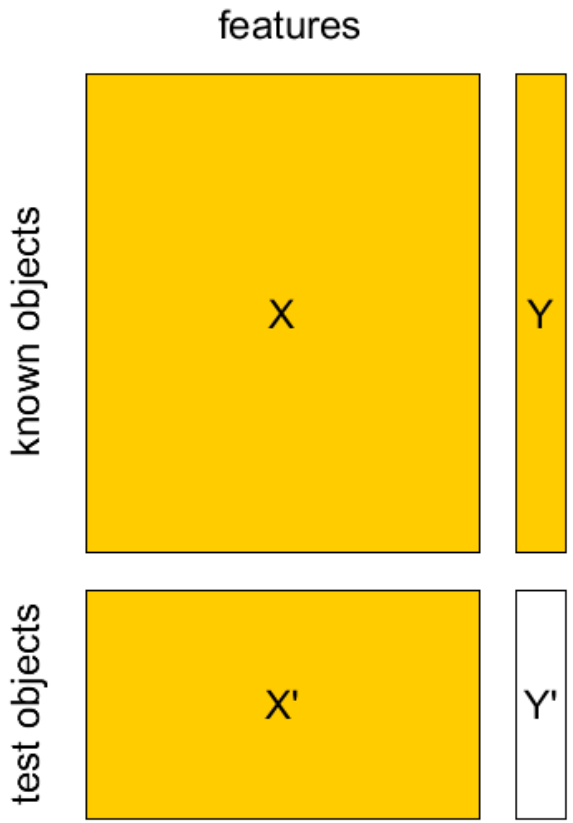


LUPI - Learning Using Priveledged Information[4].

[4]Vapnik V., Vashist A. A new learning paradigm: Learning Using Priveledged Information // Neural Networks. 2009.

Обучающая и тестовая выборка

- Обучающая выборка X, Y: $(\mathbf{x}_1, y_1), ...(\mathbf{x}_M, y_M)$
- Тестовая выборка X', Y': $(\mathbf{x}'_1, y'_1), ...(\mathbf{x}'_K, y'_K)$



Критерий оптимизации параметров модели

- Необходимо минимизировать теоретический риск:

$$\int\int \mathcal{L}(f_\theta(\mathbf{x}), y)p(\mathbf{x}, y)d\mathbf{x}dy \to \min_\theta$$

[5] Предполагаем что объекты независимы и одинаково распределены.

Критерий оптимизации параметров модели

- Необходимо минимизировать теоретический риск:

$$\int \int \mathcal{L}(f_\theta(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy \to \min_\theta$$

- Но мы можем минимизировать только эмпирический риск[5]:

$$L(\theta|X, Y) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(f_\theta(\mathbf{x}_n), y_n)$$

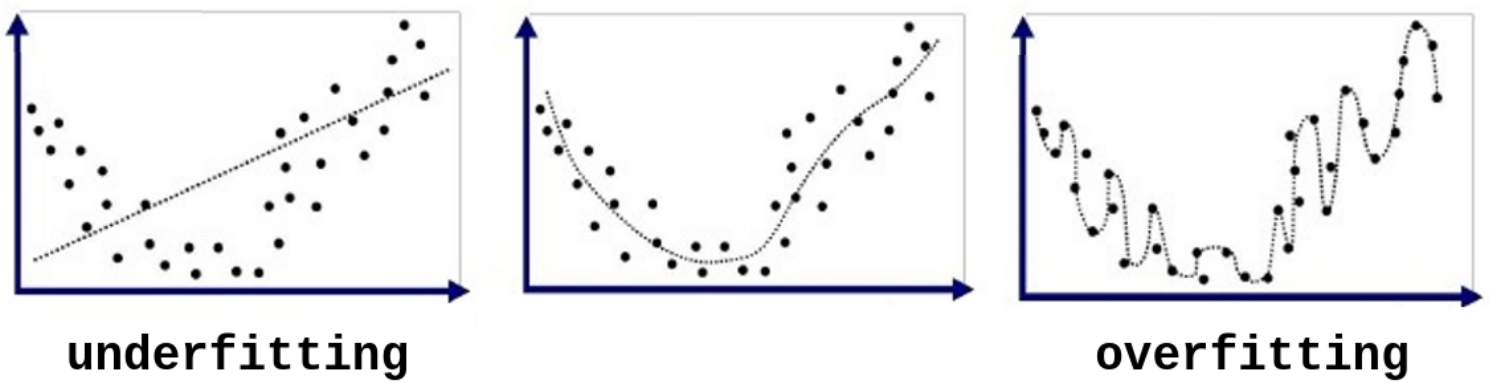
- Параметры находим из условия:

$$\widehat{\theta} = \arg\min_\theta L(\theta|X, Y)$$

[5]Предполагаем что объекты независимы и одинаково распределены.

Проблемы недообучения и переобучения

- Недообучение: модель слишком простая для реальных данных.
  - не улавливает тонких закономерностей
- Переобучение: модель слишком сложная для реальных данных.
  - настраивается на шум в измерениях



Эмпирический риск на тестовой выборке

- Как связаны $L(\widehat{\theta}|X,Y)$ и $L(\widehat{\theta}|X',Y')$?

Эмпирический риск на тестовой выборке

- Как связаны $L(\widehat{\theta}|X,Y)$ и $L(\widehat{\theta}|X',Y')$?
- В типичной ситуации

$$L(\widehat{\theta}|X,Y) < L(\widehat{\theta}|X',Y')$$

- Эффект растет с ростом переобучения.
- Как получить реалистичную оценку $L(\widehat{\theta}|X',Y')$?

Эмпирический риск на тестовой выборке

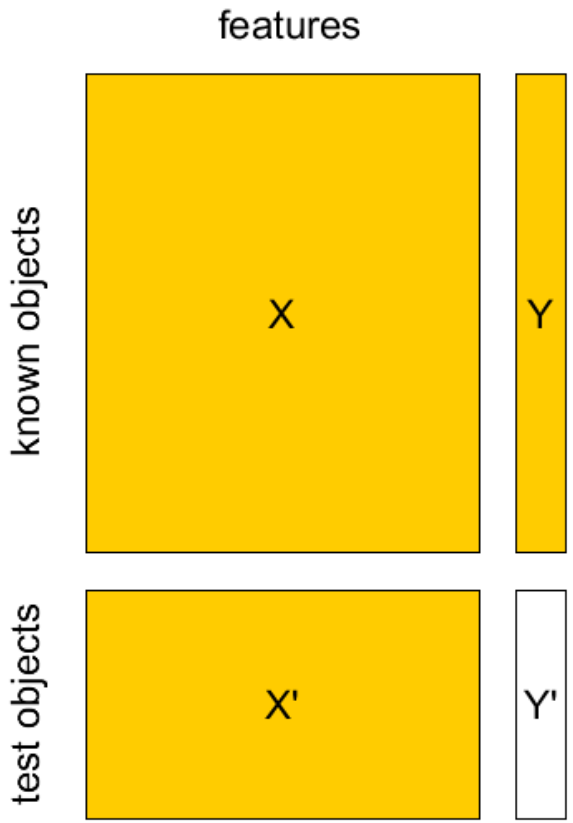
- Как связаны $L(\widehat{\theta}|X, Y)$ и $L(\widehat{\theta}|X', Y')$?
- В типичной ситуации

$$L(\widehat{\theta}|X, Y) < L(\widehat{\theta}|X', Y')$$

- Эффект растет с ростом переобучения.
- Как получить реалистичную оценку $L(\widehat{\theta}|X', Y')$?
  - на отдельной *валидационной выборке* (hold-out)
  - кросс-проверка, кросс-валидация (cross-validation)
  - скользящий контроль (leave-one-out)

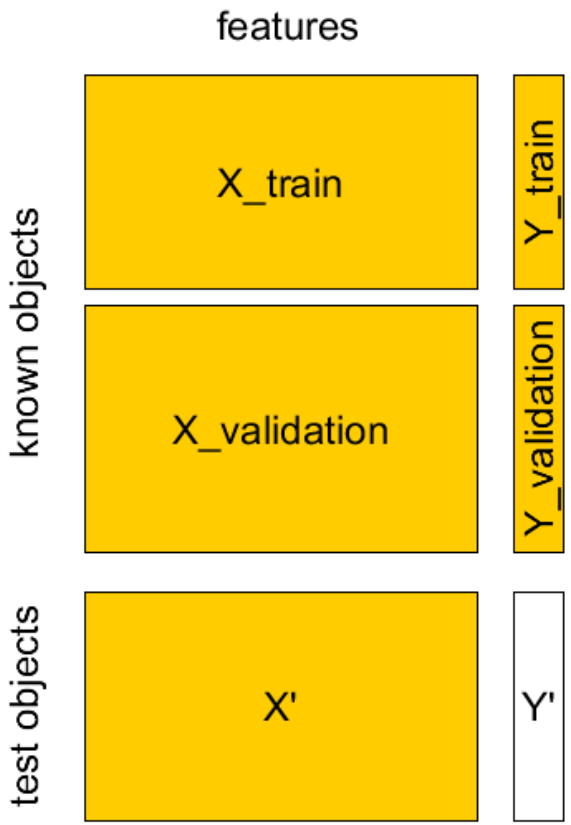
Отдельная валидационная выборка

- Обучающая выборка X, Y: $(\mathbf{x}_1, y_1), ...(\mathbf{x}_M, y_M)$
- Тестовая выборка X', Y': $(\mathbf{x}'_1, y'_1), ...(\mathbf{x}'_K, y'_K)$



Отдельная валидационная выборка

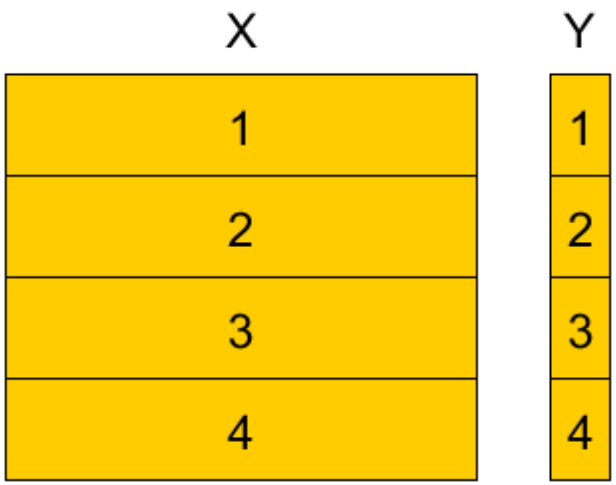
Разделим обучающую выборку на ту, где будем обучать модель и оценивать случайно (возможно, со стратификацией):



3 Оценка параметров модели

- Отдельная валидационная выборка
- Кросс-валидация

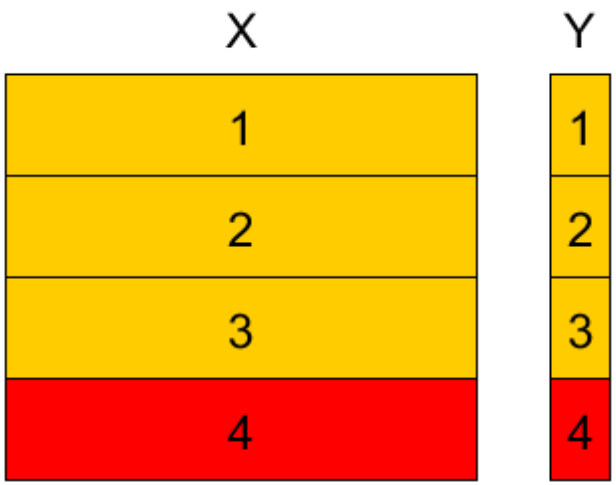
Пример: 4x блоковая кросс-валидация



Разделим обучающую выборку на K частей (блоков) ($K = 4$).

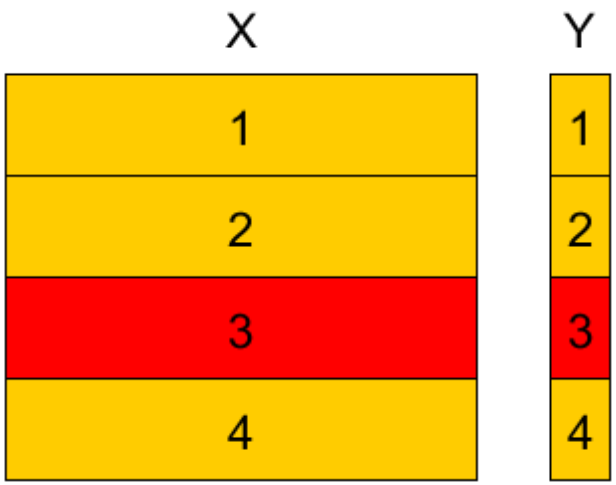
- перед разбиением важно перемешать объекты
- используется предположение независимости объектов

Пример: 4x блоковая кросс-валидация



Блоки 1,2,3 для обучения, а блок 4 - для прогнозов.

Пример: 4x блоковая кросс-валидация

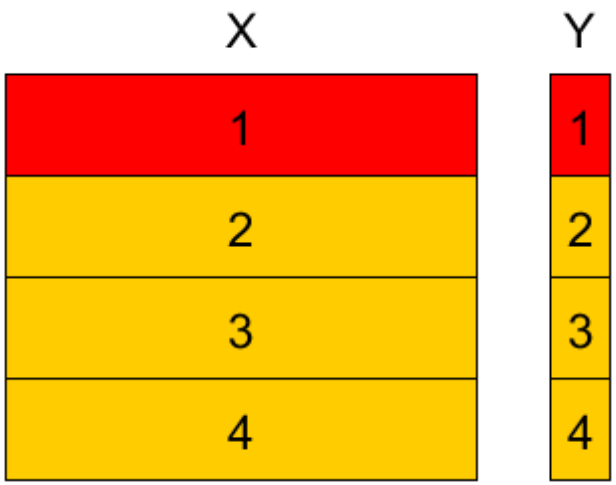


Блоки 1,2,4 для обучения, а блок 2 - для прогнозов.

Пример: 4x блоковая кросс-валидация

Блоки 1,3,4 для обучения, а блок 2 - для прогнозов.

Пример: 4x блоковая кросс-валидация



Блоки 2,3,4 для обучения, а блок 1 - для прогнозов.

Ресурсы с данными

- Соревнования по машинному обучению kaggle.com
  - полезный форум с обсуждением идей
  - много практических обучающих материалов
- Репозиторий UCI http://arhive.is.ui.edu/ml
  - данные по более чем 450 задачам.
- Есть много специализированных обучающих выборок

Этапы решения задачи[6]

- Этапы решения задачи машинного обучения:
  - Понять бизнес-проблему
  - Формализация задачи
  - Сбор данных
  - Предобработка данных
  - **Генерация признаков**
  - **Подбор модели**
  - **Оценка качества модели**
  - Внедрение модели
  - Поддержка модели

[6] Жирным выделены этапы на kaggle.com

Теория и практика

- Неясные критерии качества модели
  - заказчик не определился с целями и бизнес-процессом
- Противоречивые критерии
  - например, доходность-риск в биржевой торговле
- Грязные данные
  - ошибки измерений, сбора и обработки
- Неполные данные
  - важные признаки не собираются
- Неструктурированные данные
  - например, отчеты испытаний в свободной текстовой форме
- Данные устаревают
  - важна регулярная адаптация модели

Обозначения в курсе

- **Объекты и целевые переменные:**
  - x - вектор признаков (вход)
  - y - предсказываемая величина, отклик (выход)
  - x_i - i-й объект выборки X, y_i - i-й отклик Y.
  - x^k - k-й признак объекта x
  - x_i^k - k-й признак i-го объекта выборки x_i
- **Обучающая выборка:**
  - X - матрица объектов (объекты x признаки), $X \in \mathbb{R}^{NxD}$
  - $Y \in \mathbb{R}^N$ - вектор откликов для каждого объекта

Обозначения в курсе

- **Количественные характеристики:**
  - D - размерность признакового пространства: $x \in \mathbb{R}^D$
  - N - число (#) объектов обучающей выборки
  - C - число классов в классификации.
- Возможные классы: $\{1, 2, ...C\}$ либо $\{\omega_1, \omega_2, ...\omega_C\}$
- **Оптимизация:**
  - $\mathcal{L}(\widehat{y}, y)$ - функция потерь для одного объекта
    - y - истинный отклик, $\widehat{y}$ - прогноз.
  - $L(\theta) = \sum_{n=1}^{N} \mathcal{L}(f_\theta(x_n), y_n)$ - функция потерь на всей выборке.

Обозначения

- **Специальные функции:**
  - #[объектов] = число объектов, #[признаков] = число признаков
  - $[x]_+ = \max\{x, 0\}$ - положительная срезка
  - $\mathbb{I}[\text{условие}] = \begin{cases} 1, & \text{если условие выполнено} \\ 0, & \text{если условие не выполнено} \end{cases}$
  - $\text{sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$
- **Прочие обозначения:**
  - $\widehat{z}$ - оценка z, основанная на обучающей выборке: например, $\widehat{\theta}$ - оценка θ, $\widehat{y}$ - оценка y, и т.д.
  - $A \succcurlyeq 0$ неотрицательно определенная матрица A.
- Все вектора являются векторами столбцами, например $x \in \mathbb{R}^D$ имеет размеры Dx1.

Заключение

- Алгоритмы машинного обучения по входным признакам x прогнозируют выход y.
- Зависимость восстанавливается функцией $\widehat{y} = f_{\widehat{\theta}}(x)$ из класса $\{f_\theta(x), \theta \in \Theta\}$.
- Параметры модели контролируют сложность (гибкость) моделей.
  - бывают слишком простые и слишком сложные модели для данных
- $\widehat{\theta}$ выбирается, чтобы минимизировать эмпирический риск $\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(f_\theta(x_n), y_n)$.
- Нельзя оценивать качество модели на тех же данных, на которых она обучалась.
- Для реалистичной оценки качества нужно использовать отдельную валидационную выборку или кросс-валидацию.