Введение в машинное обучение

Machine Learning

what society thinks I do

what my friends think I do

what my parents think I do

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

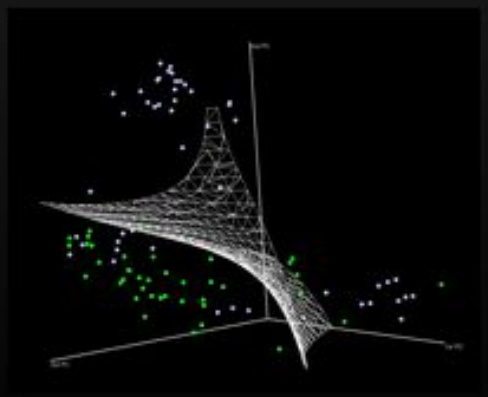
$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

what other programmers think I do



what I think I do

```
>>> from scipy import SVM
```

what I really do

Почему сейчас? Есть данные

- **Веб-данные (клики или клики в соответствии с данными)**
  - Лучше понять пользователей
  - Предсказать их поведение
- **Медицинские записи**
  - Электронные записи -> превращаем записи в знания
- **Биологические данные**
  - Генные цепочки. ML дает лучшее понимание человеческого генома.
- **Инженерная информация**
  - Данные сенсоров, логи, фотографии, камеры и т.д.

Зачем сейчас?

Сложные программы, которые мы не можем создавать вручную:

- Дроны
- Распознавание почерка
- Обработка естественного языка
- Компьютерное зрение

Зачем сейчас?

Программы, учитывающие персональные предпочтения автоматически:

–Netflix

–Amazon

–iTunes genius

–Используют пользовательскую информацию, учатся, основываясь на поведении пользователей

Что такое машинное обучение

- Arthur Samuel (1959)
  - **"Field of study that gives computers the ability to learn without being explicitly programmed«**
- Samuel написал программу, играющую в шахматы
- Программа сыграла сама с собой 10000 партий
- Сделала вывод о том, какие позиции на доске хороши/ плохи в зависимости от выигрышей / проигрышей

Что такое машинное обучение

- Tom Michel (1999)
  *"**A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."***
- Программа игры в шахматы

  E = 10000 партий

  T - игра в шахматы

  P - выигрыши/ проигрыши

Задача обучения по прецедентам

X — множество *объектов*;
Y — множество *ответов*;
$y^*: X \to Y$ — неизвестная зависимость (target function).

Дано:
$\{x_1, \ldots, x_\ell\} \subset X$ — *обучающая выборка* (training sample);
$y_i = y^*(x_i), \ \ i = 1, \ldots, \ell$ — известные ответы.

Найти:
$a: X \to Y$ — алгоритм, решающую функцию (decision function), приближающую y^* на всём множестве X.

Признаки (фичи) объектов

$f_j : X \to D_j,\ \ j = 1, \ldots, n$ — признаки объектов.

Типы признаков:

- $D_j = \{0, 1\}$ — *бинарный* признак f_j;
- $|D_j| < \infty$ — *номинальный* признак f_j;
- $|D_j| < \infty$, D_j упорядочено — *порядковый* признак f_j;
- $D_j = \mathbb{R}$ — *количественный* признак f_j.

Вектор $\big(f_1(x), \ldots, f_n(x)\big)$ — *признаковое описание* объекта x.

Матрица «*объекты–признаки*»

$$F = \big\|f_j(x_i)\big\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}$$

Основные проблемы машинного обучения

- **Недообучение:** слишком мало объектов, чтобы построить по ним зависимость – плохое решение даже при обучающей выборке

- **Переобучение:** построено отличное решение, но на примерах, которые не входят в обучающую выборку, решение не работает

- Пример алгоритма, не ошибающегося на обучающей выборке, но бесполезного на практике
  - Если пример принадлежит обучающему множеству, то вернуть известный класс
  - Если не принадлежит, то вернуть случайный класс

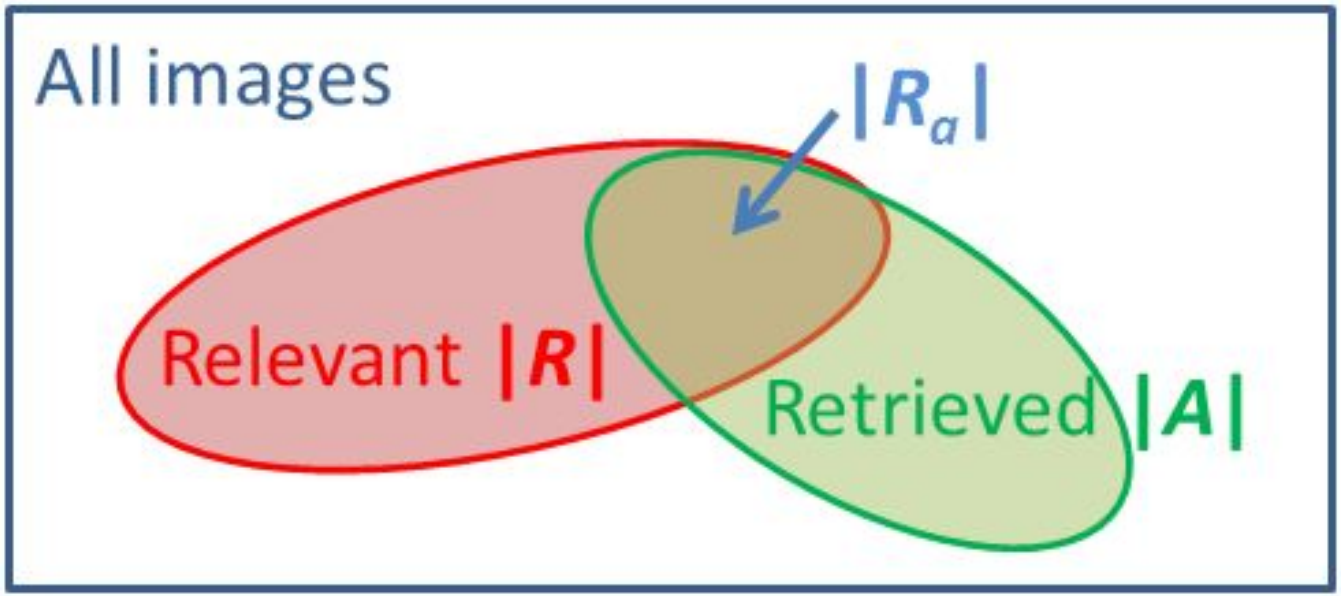
Меры качества

- Внутренняя мера качества
  - Функция, которая оптимизируется (минимизируется или максимизируется) для подбора решения=функционал качества
- Внешняя мера качества
  - Качество, которое обеспечивается в решении задачи

Полнота и точность

Полнота $|R_a|/|R|$

Точность $|R_a|/|A|$



Как проверить успешность применения

Cross-validation -- данные разбиваются на k частей.

Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования.

Процедура повторяется k раз.

Типы задач, которые решаются машинным обучением

- **Обучение с учителем (supervised learning)**
  - прецеденты -- пара «объект, ответ»
  - найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ.
- **Обучение без учителя (unsupervised learning)**
  - ответы не задаются, и требуется искать зависимости между объектами
- **Sequence labeling**
  - есть цепочка данных (например, предложение), нужно найти наилучший способ назначить тэги.

Supervised learning: типы задач

Задача классификации (classification)

–множество допустимых ответов конечно: *метки классов.* Класс — это множество всех объектов с данным значением метки.

Задача регрессии (regression)

–допустимым ответом является действительное число или числовой вектор.

Unsupervised learning: типы задач

Задача кластеризации (clustering)

–сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов.

Задача поиска ассоциативных правил (association rules learning)

–исходные данные – признаковые описания. Найти такие наборы признаков и такие значения этих признаков, которые особенно часто (неслучайно часто) встречаются в признаковых описаниях объектов.

Unsupervised learning: типы задач

Задача фильтрации выбросов

— обнаружение в обучающей выборке небольшого числа нетипичных объектов.

Задача построения доверительной области (quantile estimation)

— найти область минимального объёма с достаточно гладкой границей, содержащей заданную долю выборки.

Unsupervised learning: типы задач

Задача сокращения размерности (dimensionality reduction)

–по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки.

Задача заполнения пропущенных значений (missing values)

— замена недостающих значений в матрице “объекты–признаки” их прогнозными значениями.

Основные типы задач, востребованные в лингвистике

- Задачи классификации
- Задачи регрессии
- Задачи кластеризации

Задачи классификации при обработке текстов

- Является ли точка концом предложения
- К какой части речи относится словоформа «стали»
- Разрешение лексической неоднозначности
- Классификация текстов
  - Тематическая классификация
  - Жанровая классификация
  - Определение авторства
  - Анализ тональности
  - Классификация почтового спама
- и многие другие

Примеры задач: регрессия

- Предсказать числовую величину или построить упорядоченный список
- Примеры:
  - Формирование поисковой выдачи в информационной системе (learning-to-rank)
  - Упорядоченный список извлеченных слов (словосочетаний) по мере их терминологичности