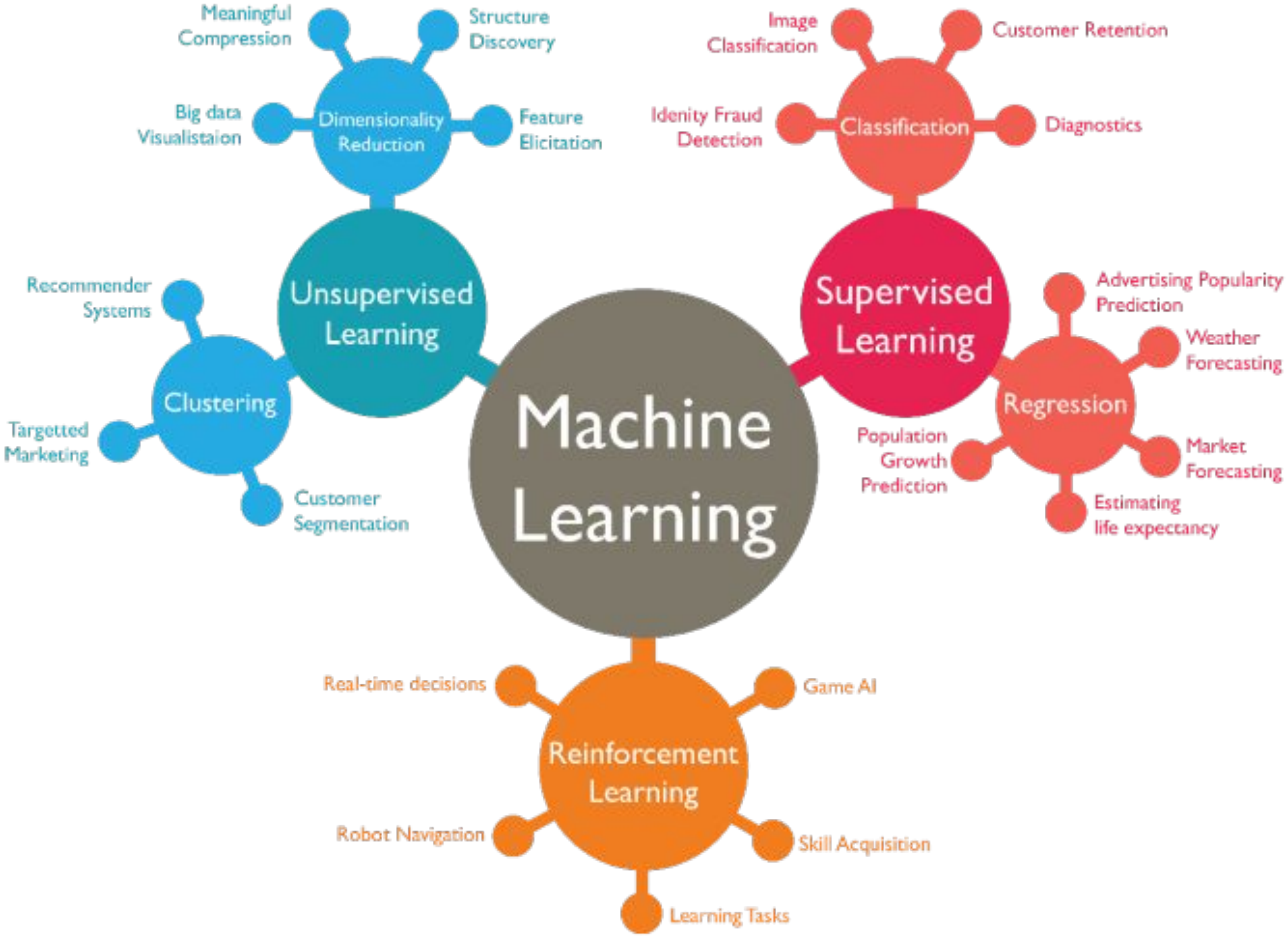
Машинное обучение и майнинг данных
Лекция 2: Линейная и Логистическая регрессия

Полина Полунина

Структура презентации

- Задачи машинного обучения: обзор
- Простая парная регрессия
- Множественная регрессия
- Логистическая регрессия
- Градиентный спуск



Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Обучение с учителем и без учителя

Обучение с учителем <=> размеченные данные

Цель: построить модель, которая обучается предсказывать целевую переменную

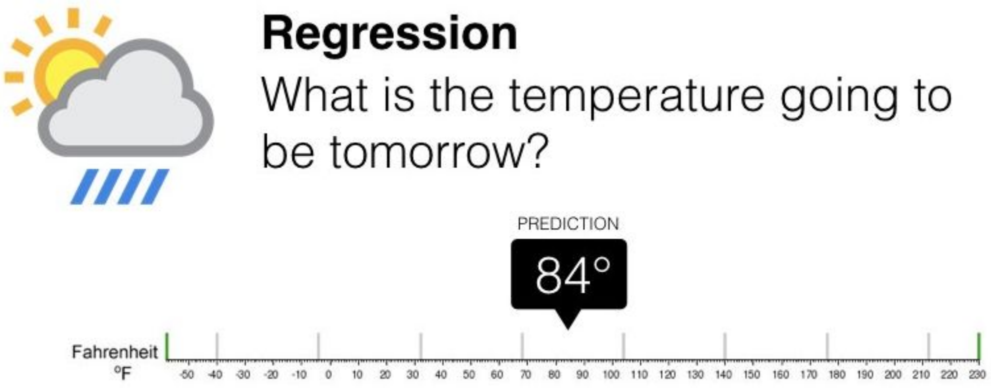
Соответствующие задачи: Регрессия, Классификация

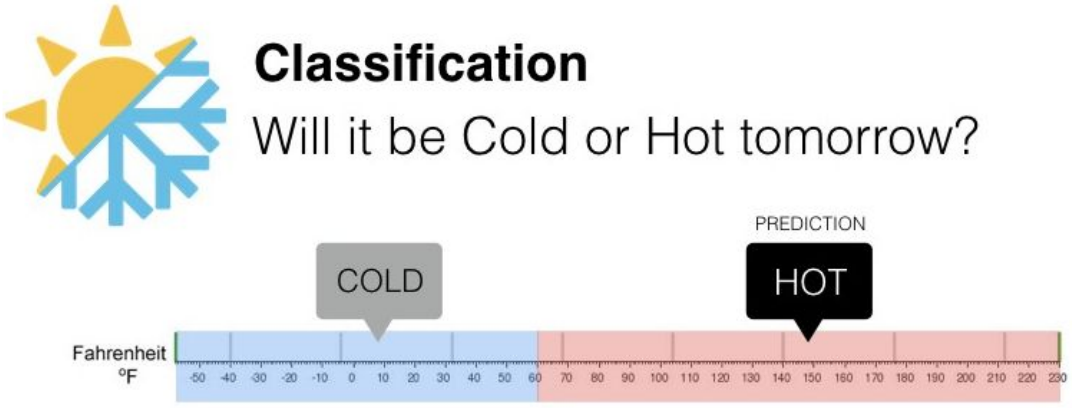
Обучение без учителя <=> данные без разметки

Цель: найти скрытые закономерности в данных

Соответствующие задачи: Кластеризация, Понижение размерности

Регрессия vs Классификация





Регрессия ⇔ непрерывная целевая переменная

Цель: предсказать значения

Example: цены на квартиры, цены на акцию, выходная мощность ветряных турбин, температура воздуха, etc.

Классификация ⇔ дискретная целевая переменная

Цель: предсказать класс

Example: кошка/собака, настроения рынка, типы сейсмических событий, спам, музыкальные жанры, etc.

Модель парной регрессии

- Задача: построить модель, описывающую зависимость между ценой дома и его общей жилой площадью

X – total living area

Y – historical price values

- Модель:

$\mathrm{Y} = a + bX + \varepsilon$

$\mathrm{Y} = \hat{a} + \hat{b}X + r,$

где ε – настоящие ошибки (errors); r – остатки (residuals), т.е. оценка случайных ошибок

- Как найти коэффициенты a и b?

Минимизация ошибки: $\sum_{i=1}^{l}(y_i - a - bx_i)^2 \rightarrow min$ - a.k.a. MSE, Mean Squared Error

Решение в явном виде: $\hat{b} = \frac{cov(X,Y)}{var(X,Y)}, \hat{a} = \bar{Y} - \hat{b}\bar{X}$

Простая линейная регрессия

Линейная регрессия: Предпосылки модели

Теорема Гаусса-Маркова:

$\hat{a}, \hat{b}$ наилучшие линейные несмещенные оценки настоящих значений коэффициентов a and b в модели линейной регрессии, если:

- $E\varepsilon_i = 0, E\varepsilon_i^2 = Var(\varepsilon_i) = \sigma^2$ - гомоскедастичность ошибок
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \; for \; i \neq j$ – некоррелированность ошибок
- Вектор признаков $(x_1, x_2, \dots, x_l)^T$ не коллинеарен единичному вектору $(1,1,\dots,1)^T$

Множественная регрессия

- Задача: предсказание цены дома по каким-либо признакам (total living area, year built, roof style, garage area, etc.)

X – набор признаков

Y – исторические значения цены на дома

- Модель:

$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_l x_l + \varepsilon = Xb + \varepsilon$, где X – матрица

- Как найти вектор b?

Минимизация суммы квадрата ошибок:

$$\sum_{i=1}^{l} r_i^2 = \sum_{i=1}^{l} (y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij})^2 = (y - Xb)'(y - Xb) \rightarrow min$$

Решение в явном виде:

$$\hat{b} = (X'X)^{-1}X'y$$

Классификация: Логистическая регрессия

- Задача: научиться классифицировать виды ирисов по каким-либо признакам

X – длина и ширина лепестков (petal), длина и ширина чашелистиков

Y – три вида ирисов: setosa, versicolor, virginica

Iris Versicolor **Iris Setosa** **Iris Virginica**

Логистическая регрессия: Maximum Likelihood Estimation

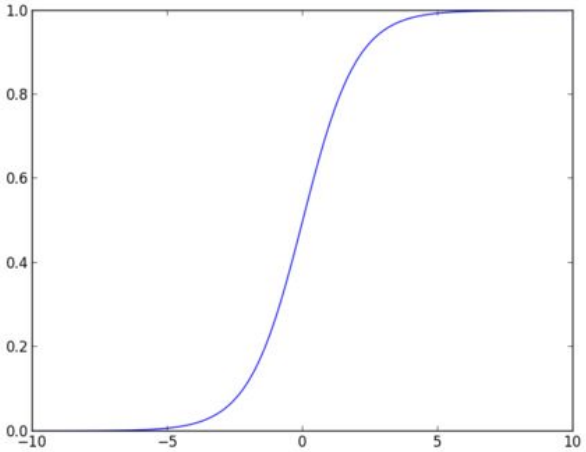
- Модель для двух классов:

$$P(y_i = 1|X) = \sigma(Xb) = \frac{1}{1 + e^{-Xb}}$$

Функция плотности вероятности:

$$P(Y = y|X) = \sigma(Xb)^y[1 - \sigma(Xb)]^{1-y}$$

$$f(x) = \frac{1}{1 + e^{-(x)}}$$



- Правдоподобие (Likelihood):

$$L(b) = \prod_{i=1}^{l} P\left(Y = y^{(i)} | X = x^{(i)}\right) = \prod_{i=1}^{l} \sigma\left(x^{(i)}b\right)^{y^{(i)}} \left[1 - \sigma\left(x^{(i)}b\right)\right]^{1-y^{(i)}}$$

- Log-likelihood:

$$\ln\left(L(b)\right) = \sum_{i=1}^{l} y^{(i)} ln\sigma\left(x^{(i)}b\right) + \left(1 - y^{(i)}\right) ln\left[1 - \sigma\left(x^{(i)}b\right)\right] \rightarrow max$$

Логистическая регрессия: Градиентный спуск

- Формула для обновления параметров на каждой итерации: $b_j^{new} = b_j^{old} + \eta \frac{\partial LL(b^{old})}{\partial b_j^{old}}$

```
Initialize: θ_j = 0 for all 0 ≤ j ≤ m

Repeat many times:
    gradient[j] = 0 for all 0 ≤ j ≤ m

    For each training example (x, y):
        For each parameter j:
            gradient[j]  += x_j (y − 1 / (1 + e^(−θ^T x)))

    θ_j += η * gradient[j] for all 0 ≤ j ≤ m
```

$$\text{gradient[j]} \mathrel{+}= x_j\left(y - \frac{1}{1+e^{-\theta^{\mathrm{T}}\mathbf{x}}}\right)$$

- Ссылка на источник

Градиентный спуск