Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e

Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Human expertise is too expensive
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What We Talk About When We Talk About“Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought “Blink” also bought “Outliers” (www.amazon.com)

- Build a model that is *a good and useful approximation* to the data.

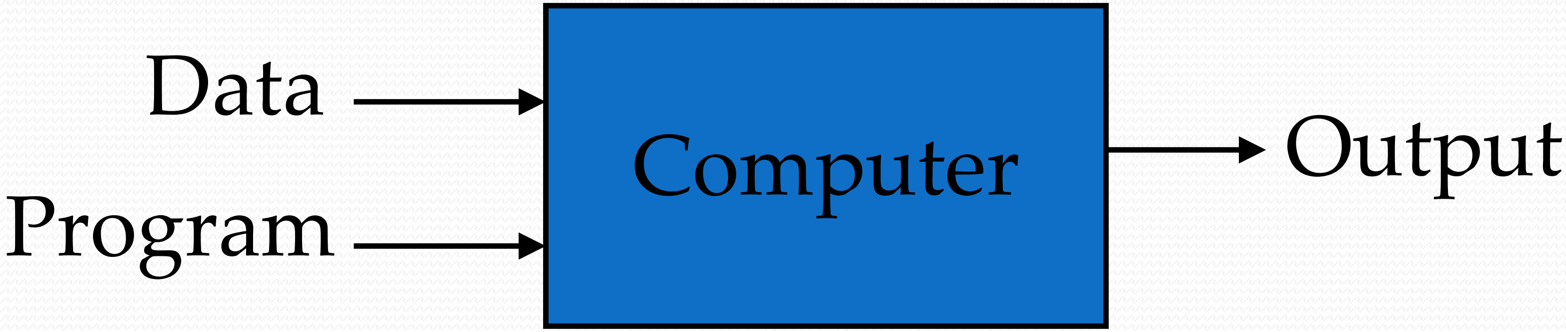
Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Web mining: Search engines, recommendations, similarity detection
- NLP: Automatic translations
- ...

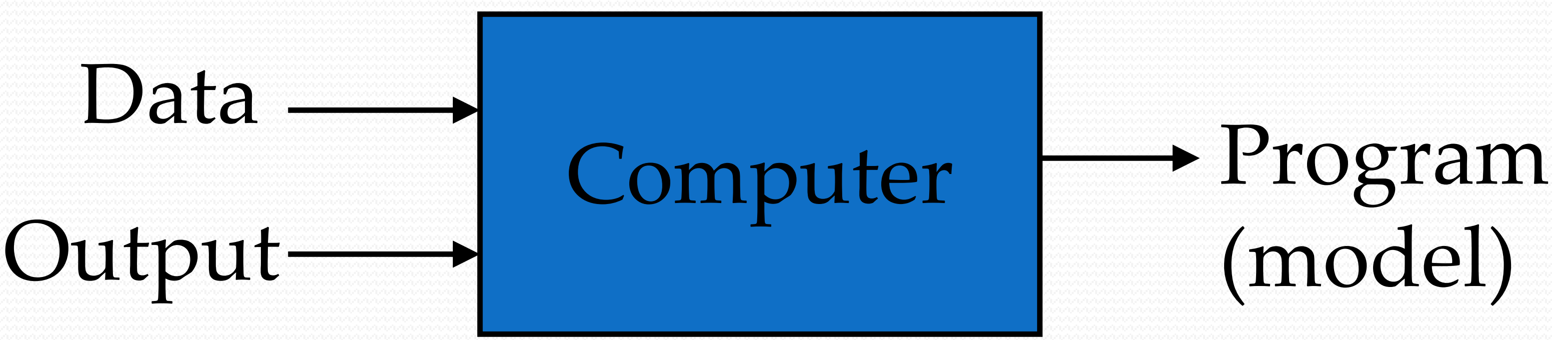
What is Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

Traditional Programming



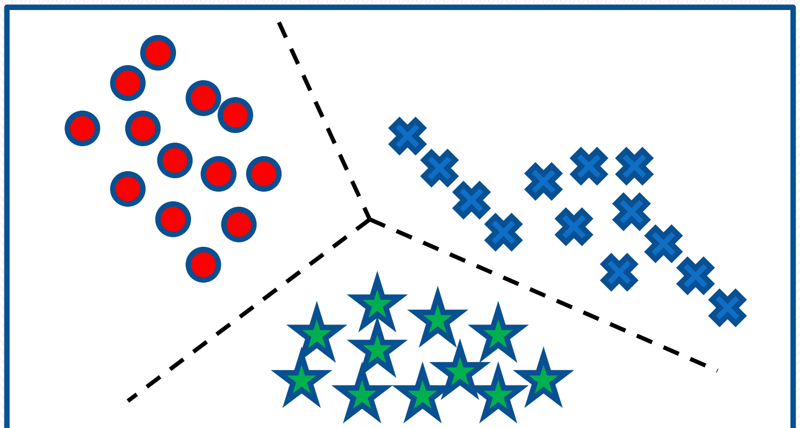
(Supervised) machine Learning



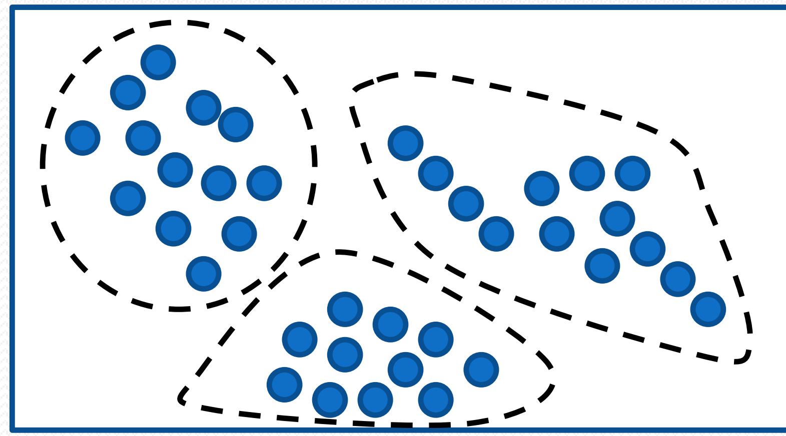
Applications

- Association
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

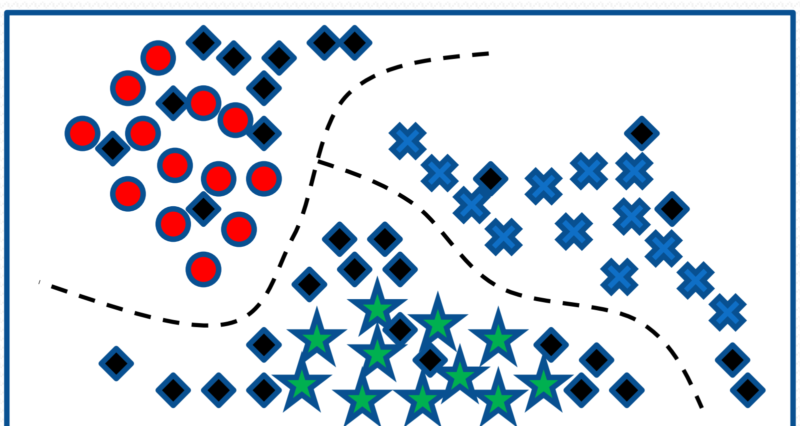
Algorithms



Supervised learning



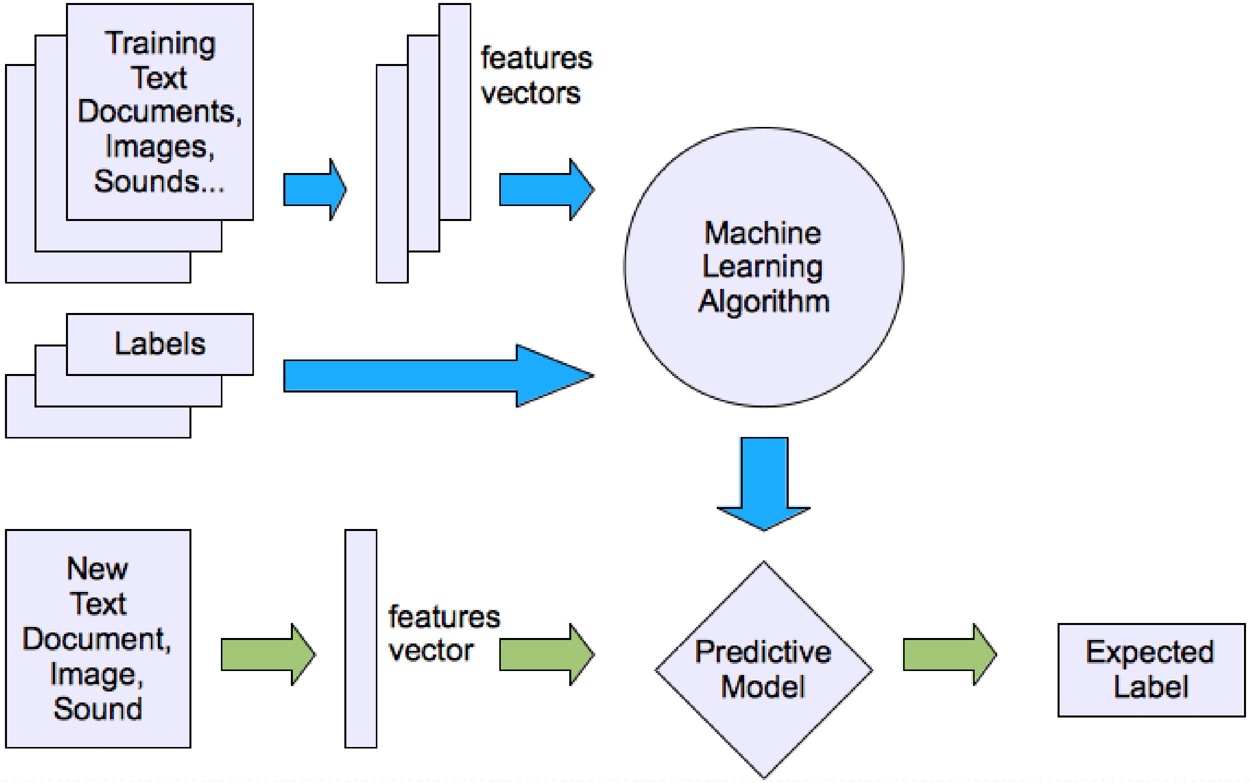
Unsupervised learning



Semi-supervised learning

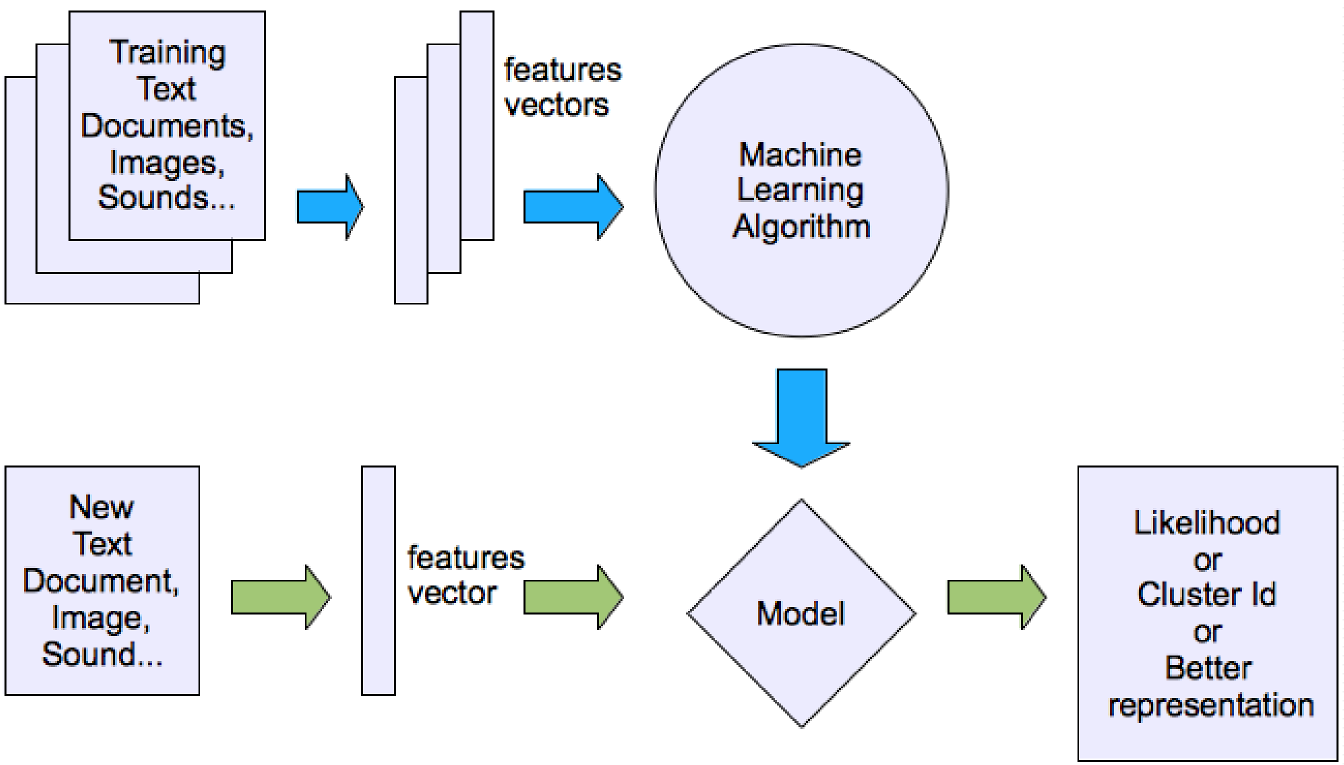
Machine learning structure

- Supervised learning



Machine learning structure

- Unsupervised learning



Learning Associations

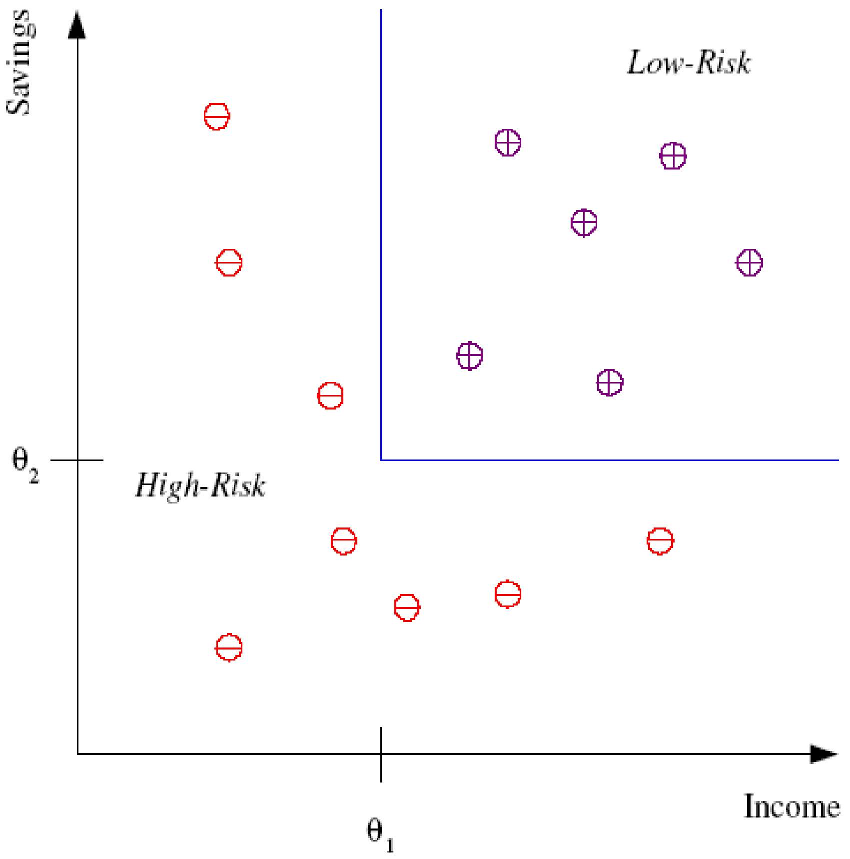
- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: $P(\text{chips} \mid \text{beer}) = 0.7$

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

Regression

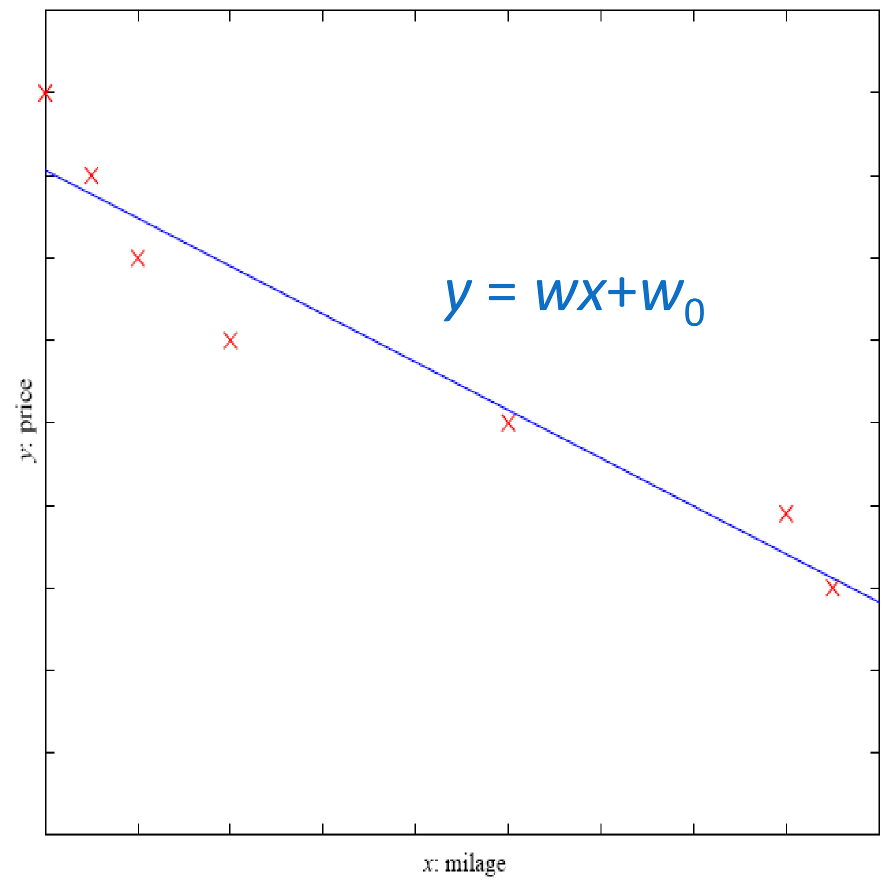
- Example: Price of a used car
- x : car attributes

  y : price

  $$y = g\,(x \mid \theta)$$

  g ( ) model,

  θ parameters



Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning "what normally happens"
- No output
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs

Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Delve: http://www.cs.utoronto.ca/~delve/
- Kaggle challenges: https://www.kaggle.com/

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on Artificial Neural Networks (ICANN)
- International Conference on AI & Statistics (AISTATS)
- International Conference on Pattern Recognition (ICPR)
- ACM Recommender Systems (RecSys)
- ...

Supervised Learning

- Input:
  - Explaining features (data)
  - Target variable (label)
- Output:
  - Function f(data -> label)

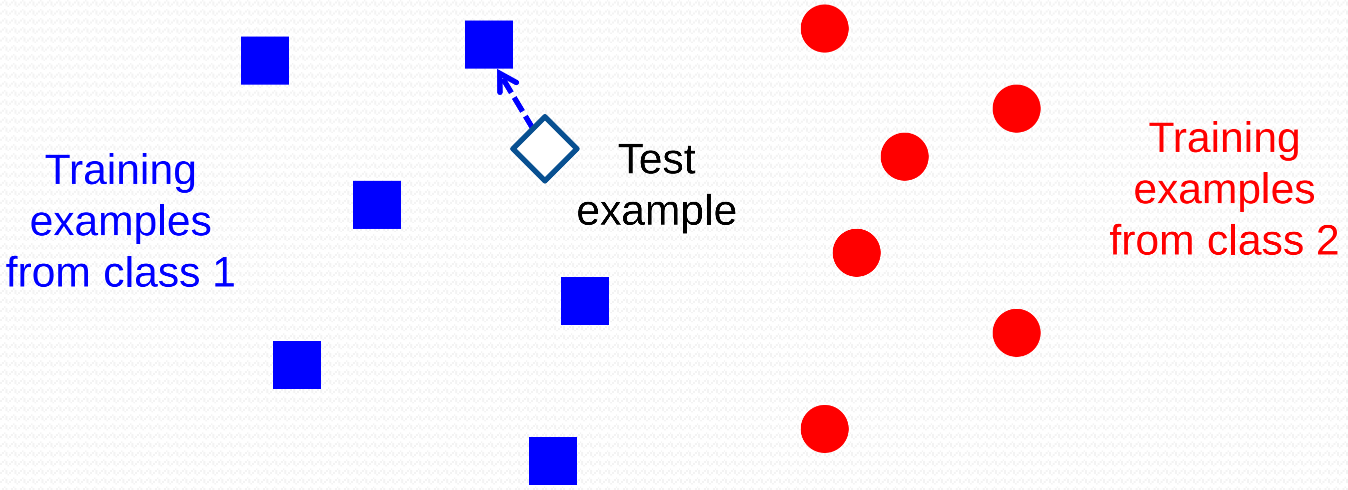
Supervised Learning - Examples

- Regression: income prediction
  - Data: Age, gender, location, education, profession,…
  - Target: income
- Classification: spam detection
  - Data: Length, from, to, metadata, text representation
  - Target: spam or not
- Multiclass classification: bank transactions labeling
  - Data: Payment ammount, target account details, text repr.
  - Target: Expense category

Supervised Learning

- Two basic approaches
  - Linear Regression
  - Nearest Neighbor Method

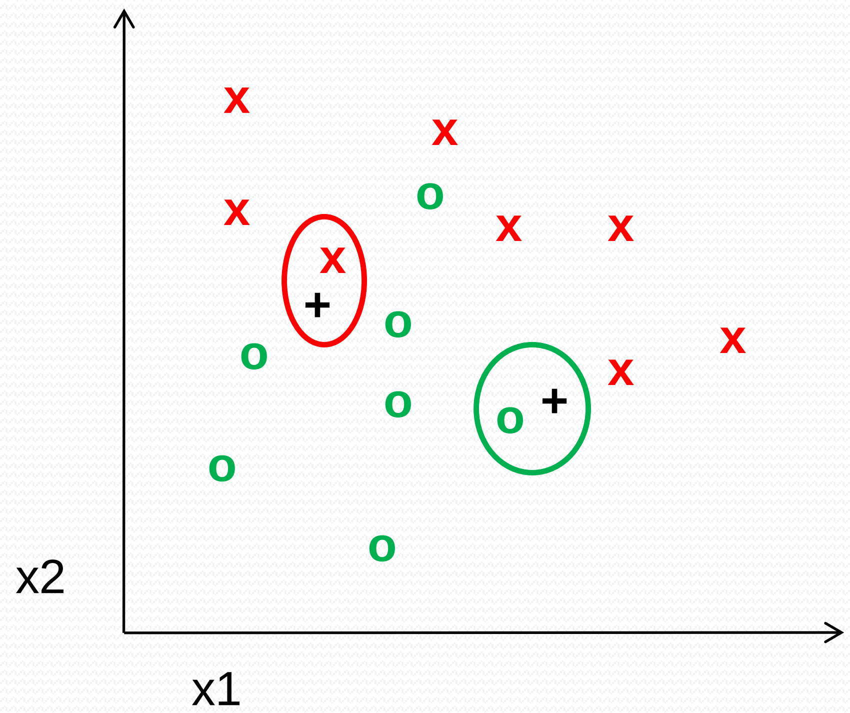
Classifiers: Nearest neighbor



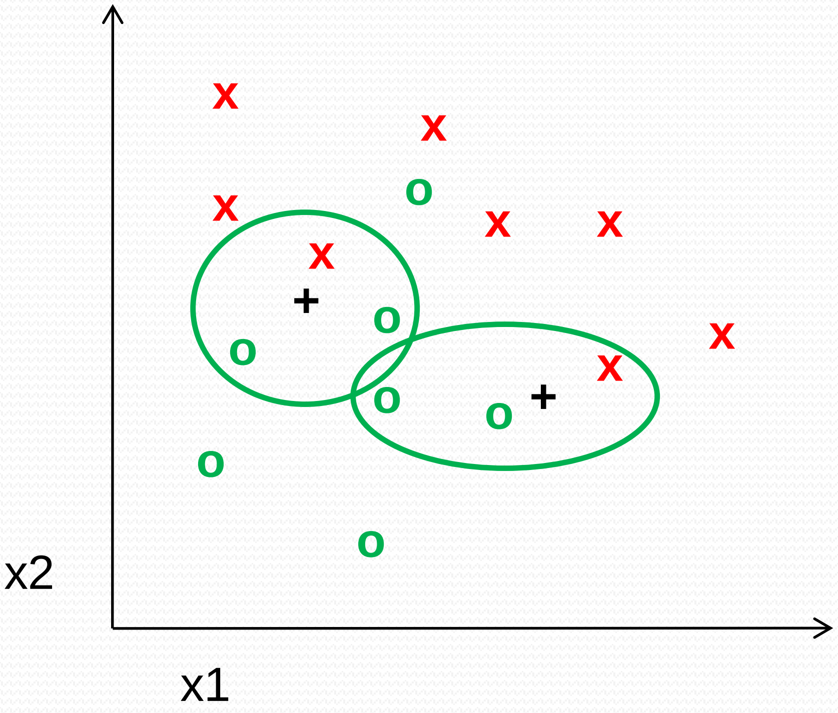
f(**x**) = label of the training example nearest to **x**

- All we need is a distance function for our inputs
- No training required!

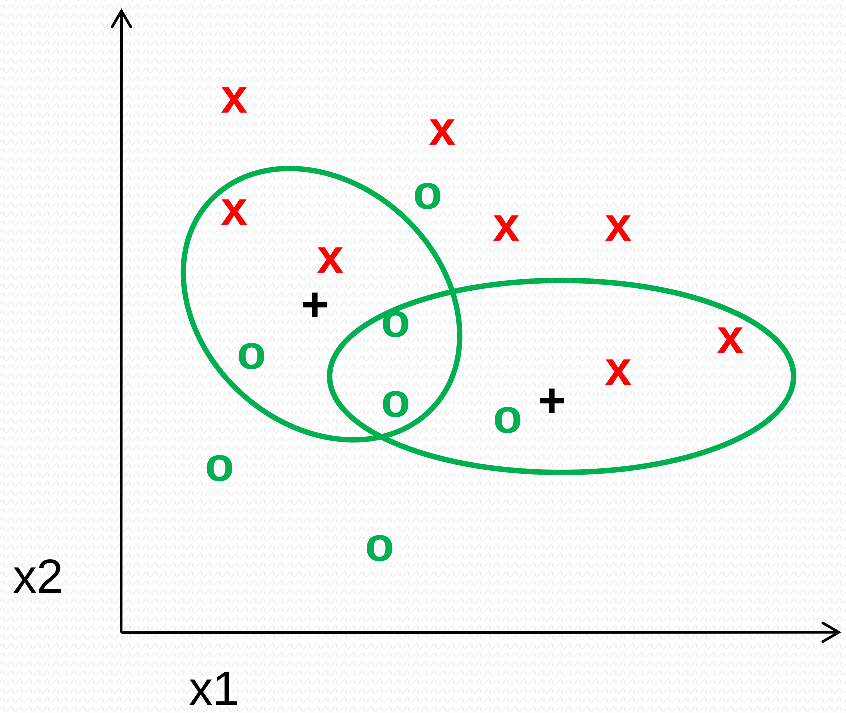
1-nearest neighbor



3-nearest neighbor



5-nearest neighbor



Using K-NN

- Simple, a good one to try first
- With infinite examples, 1-NN provably has error that is at most twice Bayes optimal error
- You need to define a distance metric among objects!

K-NN Algorithm

- „Training“
  - Just store data and labels

- Classify new example O
  - Obtain *K closest* neighbors to the O *(may be time consuming)*
  - Use majority vote of neighbors to define label of O

Linear Regression – Least Squares

- Target attribute is a linear combination of source data
  - Up to a certain noise with normal distribution

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

- How to estimate β?
  - Minimize the residual square error

$$S(b) = \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}} b)^2 = (y - Xb)^{\mathrm{T}} (y - Xb),$$

  - Derive by b, set = 0

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} S(b) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \,.$$

Úkol č.3

- Predikce žánrů knih (multi-label klasifikace)
  - Vyjděte z předchozího úkolu s knihami, načtěte prvních 1000 řádků datasetu. Pro knihy na řádcích 0-100 chcete predikovat jejich žánry na základě znalostí z ostatních knih.
  - Použijte *k*-NN algoritmus a vhodně definovanou podobnost na knihách. Vyzkoušejte vícero variant *k*, případně různé podobnostní metriky.
  - Pro každou testovanou knihu vyhodnoťte výslednou predikci jako Jaccardovu podobnost skutečných a predikovaných žánrů ($A \cap B$ / $A \cup B$). Zkuste do konce hodiny dosáhnout co nejlepší průměrné Jaccard sim.

- Otázky na zamyšlení:
  - Jaký počet žánrů predikovat?
  - Jaký je efekt *k v k*-NN?
  - Jaké jsou nevýhody k-NN?

More on KNN

- A Complete Guide to K-Nearest-Neighbors with Applications in Python and R: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/ (nice tutorial with further references)

- Hubness aware k-NN: https://www.sciencedirect.com/science/article/pii/S0950705115002282 (deal with „bad hubs“ in k-NN)

- Optimality of k-NN (if sufficient data is available): http://cseweb.ucsd.edu/~elkan/151/nearestn.pdf

- Curse of dimensionality (why we do not have sufficient data): https://en.wikipedia.org/wiki/Curse_of_dimensionality

More on KNN

- A Complete Guide to K-Nearest-Neighbors with Applications in Python and R: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/ (nice tutorial with further references)

- Hubness aware k-NN: https://www.sciencedirect.com/science/article/pii/S0950705115002282 (deal with „bad hubs“ in k-NN)

- Optimality of k-NN (if sufficient data is available): http://cseweb.ucsd.edu/~elkan/151/nearestn.pdf

- Curse of dimensionality (why we do not have sufficient data): https://en.wikipedia.org/wiki/Curse_of_dimensionality, https://math.stackexchange.com/questions/346775/confusion-related-to-curse-of-dimensionality-in-k-nearest-neighbor