AIML428

- Admin
  - Presentation sign up
    - Edit the shared file, hard copy or email
  - Topics
    - Hot topics, new systems or applications
      - e.g. Sora, Gemini, ChatGPT,
    - New algorithms or technologies; or something you know well
      - e.g. Transformer, attention, BERT, Bi-LSTM,

Marking guide

[3] Topic:

related to course, most recent research

[3] Slides:

concise/clarity/readability

informative/effective/using images/demos

structure and organisation, colour scheme

[4] Presentation skills

clear/show deeper understanding/insights

body language, voice tone, pace

confidence/well prepared, attract attention/interesting

effective question answering

Learning materials:

- Machine learning algorithms: Wikipedia
- Teaching Videos
- Python and implementation: online tutorials
  - I installed Python using Anaconda
  - I use Jupyter Notebook to run Python
- Papers: google scholar
- Technical reports

Today

- Review on supervised machine learning
- An example: KNN in python

- Features
- Text features

Supervised Machine Learning

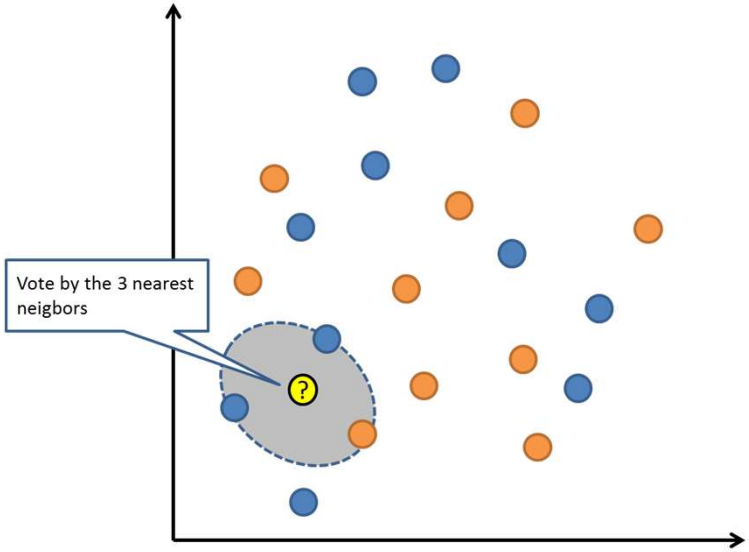
1. Build or get a representative corpus
2. Label it
3. Define features
4. Represent the instances
5. Learn and analyse
6. Go to 3 until accuracy is acceptable

Test on unseen instances

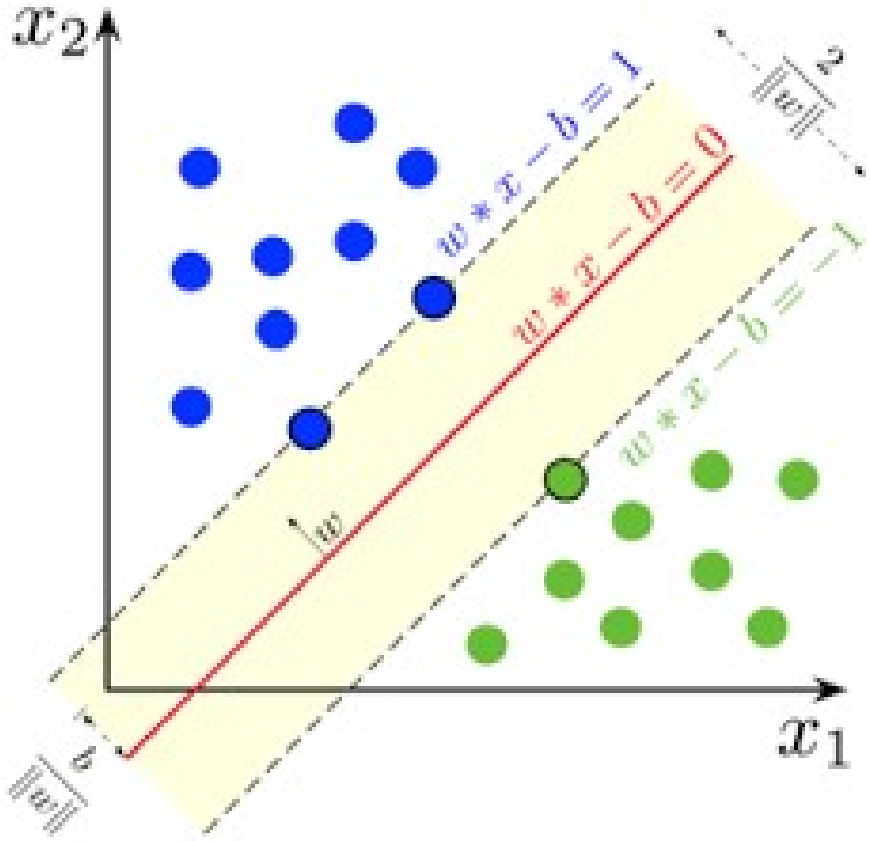
Classify the fruits, what features?

mass	width	height	color_score	label
192	8.4	7.3	0.55	0
180	8.0	6.8	0.59	0
86	6.2	4.7	0.80	2
176	7.4	7.2	0.60	0
90	7.1	5.6	0.25	1

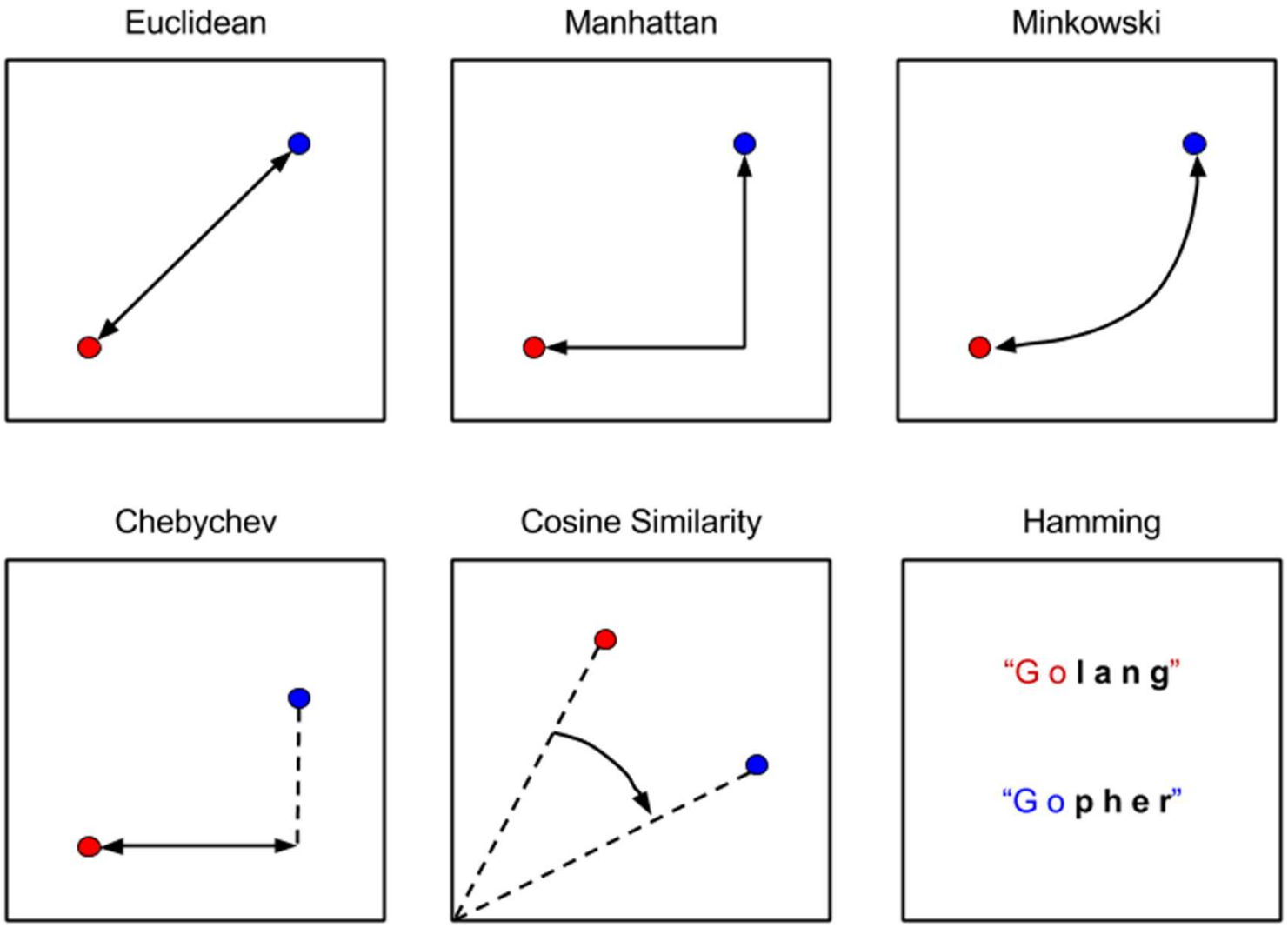
Supervised learning



Many algorithms are distance based



Distance measures (similarity measures)



Similarity Measures

Cosine similarity

- For two vectors $\vec{x}$ and $\vec{y}$, the cosine similarity between them is given by:

$$\cos(\angle(\vec{x}, \vec{y})) = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

- Here $\vec{x} \bullet \vec{y}$ is the vector product of $\vec{x}$ and $\vec{y}$, calculated by multiplying corresponding frequencies together
- The cosine measure calculates the angle between the vectors in a high-dimensional virtual space

Supervised learning algorithms

- K-nearest neighbour (KNN)
- Support Vector Machines (SVM)
- Decision tree learning, e.g. C4.5
- Naïve Bayes (NB)

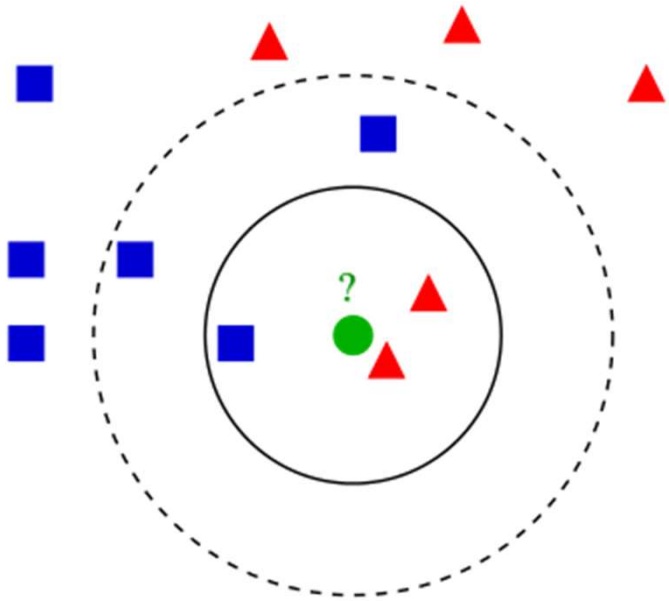
- Neural Networks, shallow, deep, variants
  - Convolutional Neural Network (CNN)
  - Long Short Term Memory (LSTM)
  - Gated Recurrent Unit (GRU)
  - Bidirectional RNN
  - Recurrent Convolutional Neural Network (RCNN)

- Genetic Algorithms (GA), Genetic Programming (GP), Particle Swarm Optimisation (PSO)

- Top 10 algorithms in data mining in 2007 paper by X Wu etc.
  - C4.5, K-means, SVM, Apriori, PageRank, EM, AdaBoost, KNN, NB, CART

KNN in python

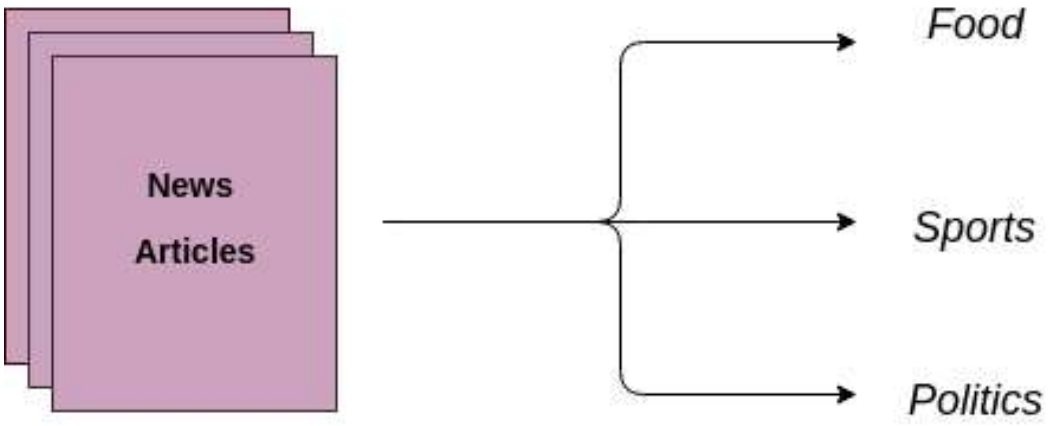
- KNN: An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours
- The data file
- Features, target
- Training, testing



- KNN model
- Fit with training data
- Predict on testing data
- Evaluation

Text classification

- The goal of text classification is to automatically classify the text documents into one or more pre-defined categories.
- Typical applications
  - Categorization of news articles into defined topics.
  - Understanding audience sentiment from social media,
  - Detection of spam and non-spam emails,
  - Auto tagging of customer queries



Classification

Supervised learning

1. Build or get a representative corpus
2. Label all instances, split into training set and testing set

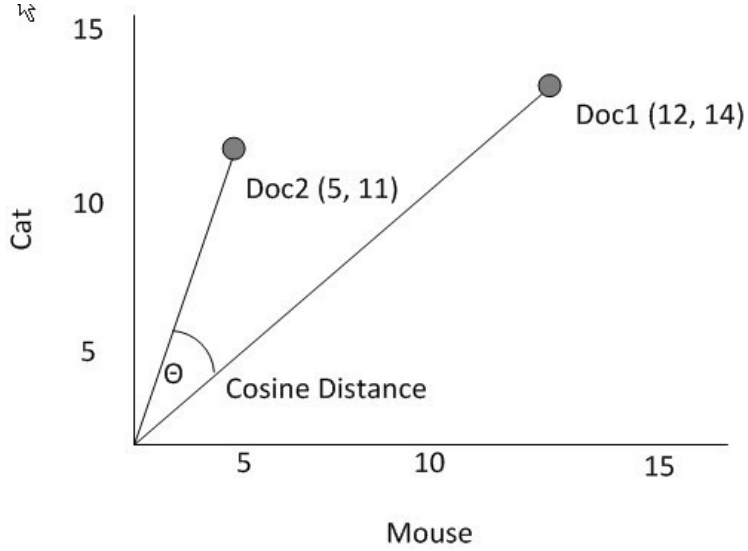
3. Define features (Text features are special)
4. Represent the instances
5. Learn and analyse
6. Go to 3 until accuracy is acceptable

Evaluate on test dataset (unseen instances)

Vector Space Model

Any text object can be represented by a vector.

Example
Doc1: 0.3, 0.1, 0.4, …
Doc2: 0.8, 0.5, 0.6, …



Vector Space Similarity: Cosine of the angle between the two vectors

$$\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$$

$$\frac{0.3 * 0.8 + 0.1*0.5 + 0.4*0.6}{\sqrt{0.3^2 + 0.1^2 + 0.4^2}\ \sqrt{0.8^2 + 0.5^2 + 0.6^2}}$$

Text data

- What are the features?

- How to change the features into numbers

Why features are important?

- Classify our group

- Features
  - Gender
  - Age
  - Nationality
  - Hair colour
  - Eye colour
  - Height
  - Weight
  - …

  - Assignment score
- Labels: