UPPSALA
UNIVERSITET

Information Retrieval (5LN712)

Vector space classification

2020-05-27

Ali Basirat

Department of Linguistics and Philology

UPPSALA UNIVERSITET

Today

- Document as vector
- How to classify document vectors
- How many document classes
- What classification methods should be used

UPPSALA UNIVERSITET

Document representation

- Each document is considered as a vector in a continuous vector space
- Usually a weighting mechanism is used
- We can classify the vector documents based on the contiguity hypothesis
- Documents in the same class form a contiguous region and regions of different classes do not overlap

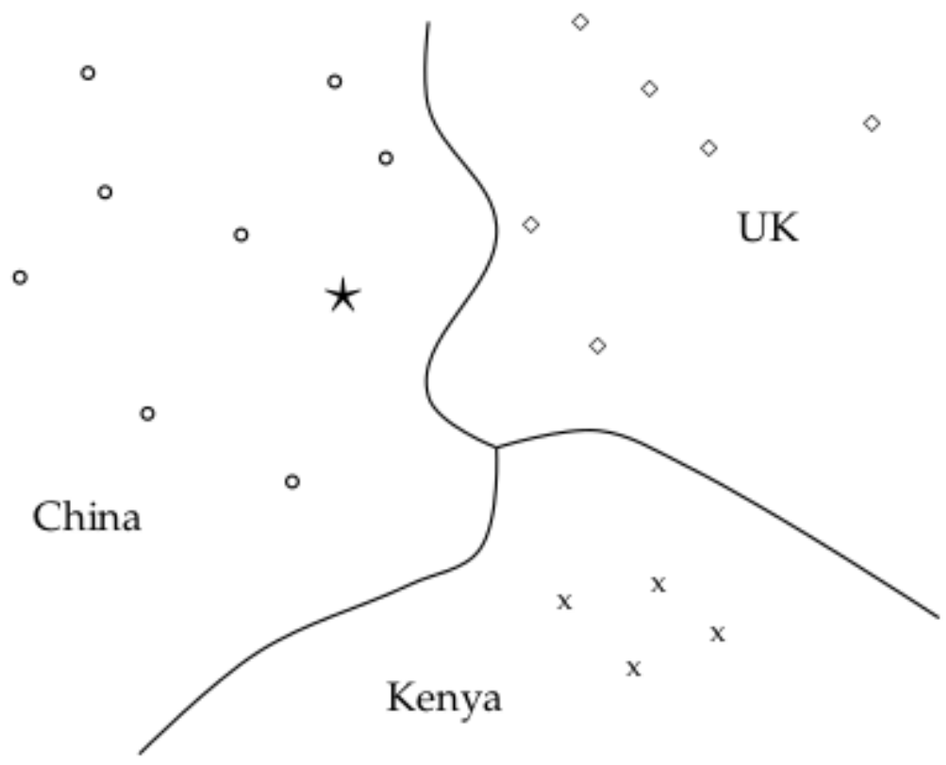
UPPSALA UNIVERSITET

Document representation

- The document classes can be distinguished by word patterns
- Some dimensions have higher weights for some documents
- The documents that tend to have high weights on similar dimensions (terms) belong to the same document class

Document representation

- Documents in class Chinese tend to have values on dimensions like Chinese, Beijing, and Mao
- Documents in class UK tend to have values on dimensions like London, British, and Queen



UPPSALA UNIVERSITET

Document representation

- The distance between vectors can be measured in different ways
- In practice the vectors are normalized so:
  - They lie on a sphere
  - The cosine distance and the Euclidian distance are related to each other

UPPSALA UNIVERSITET

Vector classification

- Decision boundary
- We should find a decision boundary that maximizes the classification accuracy
- Rocchio algorithm is based on centroids
- KNN algorithm is based on nearest neighbors

UPPSALA UNIVERSITET

Rocchio classification

- The centroid of a class is the vector average or the center of the mass of its members

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

- The boundary between two classes is a line (hyperplane) with equal distance between the centroid of the classes
- Documents are assigned to classes based on their position on the decision boundary
- Or their distance to the centroids of classes

UPPSALA UNIVERSITET

K nearest neighbor

- Each document is assigned to the majority class of its k closest neighbors
- We can measure the probability of memberships
- The parameter k is chosen based on experience or prior knowledge
- We can weights the memberships with cosines

$$score(c,d) = \sum_{d' \in S_k} I_c(d') \cos(\vec{v}(d'), \vec{v}(d))$$

$$I_c(d') = 1 \Leftrightarrow d' \in c$$

UPPSALA UNIVERSITET

Linear vs nonlinear classification

- Document space: each document is a vector
- Linear classification: to find a hyperplane that classifies the documents
- The decision hyperplane:

$$\sum_{i=1}^{n} w_i x_i = b$$

$$\vec{w}^T \vec{x} = b$$

- $\vec{x} = [x_1, \ldots, x_n]^T$ is a document vector
- $\vec{w} = [w_1, \ldots, w_n]^T$ is a feature weight vector
- Decision making

$$\sum_{i=1}^{n} w_i x_i - b > 0$$

UPPSALA UNIVERSITET

Linear vs nonlinear classification

- The Rocchio algorithm is a linear classifier
- The decision boundary is a hyperplane with equal distance to the mean of the classes

$$|\vec{\mu}(c_1) - \vec{x}| = |\vec{\mu}(c_2) - \vec{x}|$$

$$\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$$

$$b = \frac{1}{2}(|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2)$$

$$\vec{w}^T \vec{x} = \mathrm{b}$$

UPPSALA UNIVERSITET

Linear vs nonlinear classification

	Doc ID	Words in documents	class
Training set	1	Chinese Beijing Chinese	1
	2	Chinese Chinese Shanghai	1
	3	Chinese Macao	1
	4	Tokyo Japan Chinese	0
Test set	5	Chinese Chinese Chinese Tokyo Japan	?

UPPSALA UNIVERSITET

Linear vs nonlinear classification

- The Naïve Bayes classifier is linear
- Classify documents with respect to $\hat{P}(c|d)$

$$\hat{P}(c|d) \propto \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

- If we aim at classifying between a class and its complement

$$\log \frac{\hat{P}(c|d)}{\hat{P}(\bar{c}|d)} = \log \frac{\hat{P}(c)}{\widehat{P(\bar{c})}} + \sum_{1 \leq k \leq n_d} \log \frac{\hat{P}(t_k|c)}{\hat{P}(t_k|\bar{c})}$$

- The class c is selected if the log of odds is greater than 0

UPPSALA UNIVERSITET

Linear vs nonlinear classification

- The linear weights of the multinomial NB classifier is:

$$w_i = \log[\frac{\hat{p}(t_i|c)}{\hat{p}(t_i|\bar{c})}]$$

$$x_i = \sum_{t \in d} t == t_i$$

$$b = -\log[\frac{\hat{P}(c)}{\hat{P}(\bar{c})}]$$

UPPSALA UNIVERSITET

Linear vs nonlinear classification

	Doc ID	Words in documents	class
Training set	1	Chinese Beijing Chinese	1
	2	Chinese Chinese Shanghai	1
	3	Chinese Macao	1
	4	Tokyo Japan Chinese	0
Test set	5	Chinese Chinese Chinese Tokyo Japan	?

UPPSALA UNIVERSITET

Linear vs nonlinear classification

- The nonlinear boundary classes are complex
- kNN is a non-linear classifier
- Kernel methods, Gaussian processes, and neural networks
- If linear classifiers fail, then try non-linear ones

UPPSALA UNIVERSITET

Classification with more than two classes

- Any-of classification: a document can belong to several or no classes

1. Build a binary classifier for each class
2. At test time, apply each classifier separetally

- One-of classification: a document can belong to only one class (e.g., kNN)
- Not commonly used for documents classification
- Used for example to determine the language of a document

UPPSALA UNIVERSITET

The bias-variance tradeoff

- Should we always use non-linear classifiers?
- Mean squared error: the expected error made by a classifier γ

$$MSE(\gamma) = E_d[\gamma(d) - P(c|d)]^2$$

- An optimal classifier minimizes MSE
- A good learning method should minimize the learning error over labelled training sets

$$LE(\Gamma) = E_D[MSE(\Gamma(D))]$$

UPPSALA UNIVERSITET

The bias-variance tradeoff

- The learning-error is the sum of two terms: bias plus variance

$$E_d[[E_D\Gamma_D(d) - P(c|d)]^2 + E_D[\Gamma_D(d) - E_D\Gamma_D(d)]^2]$$

- Bias is the expected error over training sets
- Variance is the expected variation of the learned classifiers over training sets – how much the classifier changes if we change the training data

UPPSALA UNIVERSITET

The bias-variance tradeoff

- The flexible methods increase variance and decrease bias
- A high value of bias means the classification method tends to make a lot of mistakes (underfitting)
- A high value of variance means the classification method is sensitive to the changes in training sets (overfitting)
- Linear methods have high bias for non-linear problems, but low variance
- Non-linear methods have high variance, but low bias

UPPSALA UNIVERSITET

The bias-variance tradeoff

- Linear methods work better for text classification
- High-dimensional data are more likely to be linearly separable
- Noises are handled better by linear methods

UPPSALA UNIVERSITET

Summary

- Documents can be represented as vectors
- Documents classification meets vector classifications
- Rocchio classification
- kNN classification
- Linear vs nonlinear classification
- Classification with more than two classes
- The bias-variance tradeoff