Workshop 3

COMP90051 Natural Language Processing

Semester 1, 2020

- Online lectures and tutorials
- Recording
- Questions

Materials

- Download files
  - Workshop-03.pdf
  - 03-classification.ipynb
  - 04-ngram.ipynb
- From Canvas - Modules - Workshops - Worksheets/ Notebooks

Learning Outcomes

- Text classification
  - Definition, Applications, Challenge …
  - Algorithms
- N-gram language model
  - Different N
  - Smooth vs Non-Smooth
  - Back-off and Interpolation

Text Classification

Text classification

1. What is **text classification**? Give some examples.

Text classification

1. What is **text classification**? Give some examples.

- Text classification is the task of classifying text documents into different labels.
  - Input
  - Output

Text classification

1. What is **text classification**? Give some examples.

- Text classification is the task of classifying text documents into different labels.
  - Input
    - a document *d*
    - a fixed set of labels C
  - Output
    - A predicted class c ∈ C

Sentiment Analysis

- Document d: I like this movie.
- Labels: Positive, Negative

Text classification

1. What is **text classification**? Give some examples.

- Examples

Text classification

1. What is **text classification**? Give some examples.

- Examples
  - Topic classification
  - Sentiment analysis
  - Authorship attribution
  - Native-language identification
  - Automatic fact-checking

Text classification

(a) Why is text classification generally a difficult problem? What are some hurdles that need to be overcome?

Text classification

(a) Why is text classification generally a difficult problem? What are some hurdles that need to be overcome?

- Problem
  - Document representation
    - BOW

Bag-Of-Words

- Document A: I like natural language processing
- Document B: I am playing a game
- Document C: The aims for this subject is to develop an understanding of natural language processing

	I	Like	Natural	Language	Processing	Am	Playing	A	Game	The	aims	For	…	Of
Doc A	1	1	1	1	1	0	0	0	0	0	0	0	…	0
Doc B	1	0	0	0	0	1	1	1	1	0	0	0	…	0
Doc C	1	0	1	1	1	0	0	0	0	1	1	1	…	1

- What is the length of vectors?

Text classification

(a) Why is text classification generally a difficult problem? What are some hurdles that need to be overcome?

- Document representation
  - BOW
- Feature selection
- Sparse data problem

Classifier

- (b) Consider some (supervised) text classification problem, and discuss whether the following (supervised) machine learning models would be suitable:

Classifier

- (b) Consider some (supervised) text classification problem, and discuss whether the following (supervised) machine learning models would be suitable:

  i. k-Nearest Neighbour using Euclidean distance

  ii. k-Nearest Neighbour using Cosine similarity

  iii. Decision Trees using Information Gain

  iv. Naive Bayes

  v. Logistic Regression

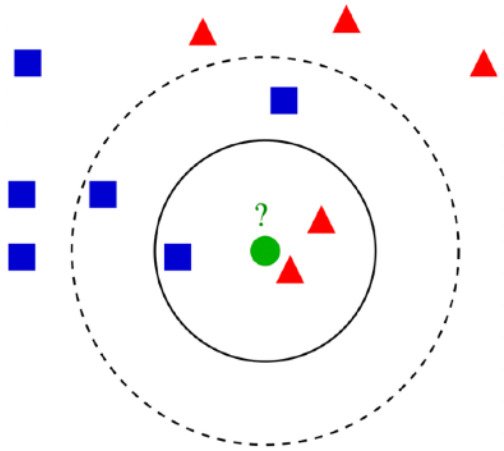
  vi. Support Vector Machines

Classifier

- It depends on
  - Number of Features
  - Number of classes
  - Number of instances
  - Underlying assumption
  - Complexity
  - Speed
  - …

KNN

- Classify based on majority class of *k*-nearest training examples in feature space



- High-dimensionality problems

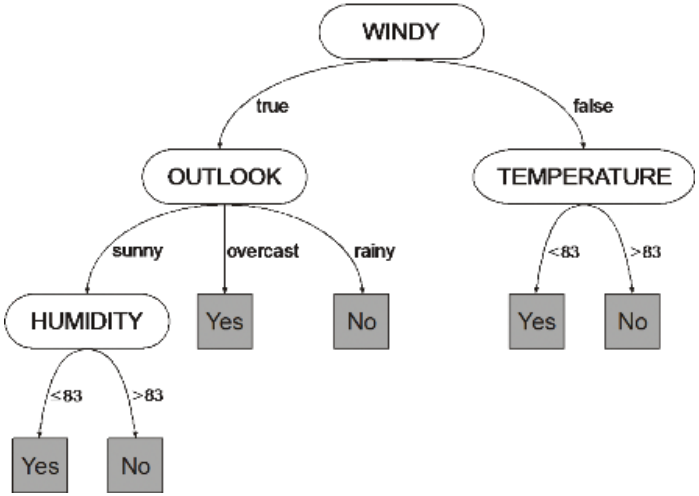
Euclidean distance vs Cosine similarity

- Doc A: 11111000000000000000
- Doc B: 10000111100000000000
- Doc C: 10111000011111111111

- Euclidean distance:
  - d(q,p) $= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$
- Cosine similarity
  - c(a,b) $= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$

Decision Tree

- Construct a tree where nodes correspond to tests on individual features



- Feature selection
- Information Gain
  - It tends to prefer rare features

Naive Baye

- Finds the class with the highest likelihood under Bayes law

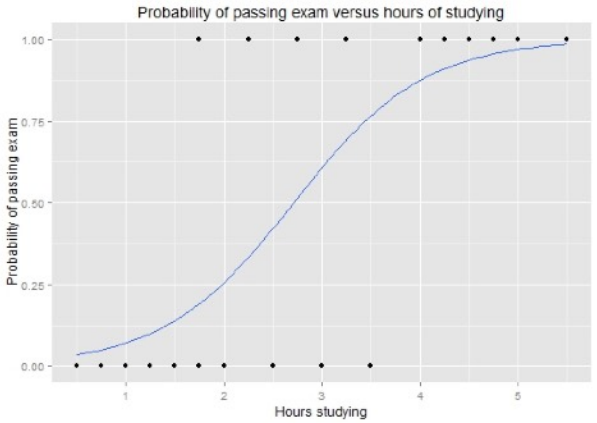
- Assumption of NB

Naive Baye

- Finds the class with the highest likelihood under Bayes law
- Assumption of NB
  - The conditional independence of features and classes
- Sensitive to a large feature set

Logistic Regression

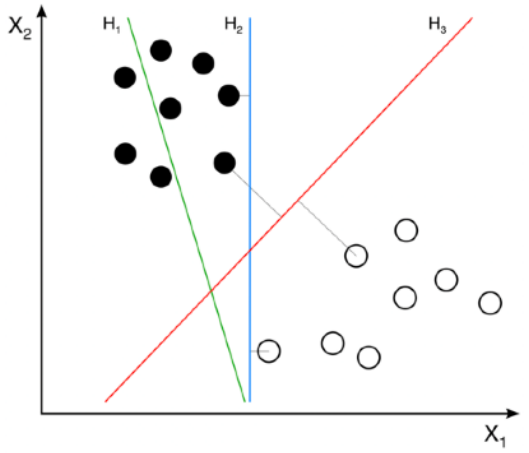
- A linear classifier



- Relaxes the conditional independence
- Handle large numbers of mostly useless features

Support Vector Machines

- Finds hyperplane which separates the training data with maximum margin



- Multiple classes

Language model

Language model

- What is language model
  - Models that assign probabilities to sequences of words
- 2. For the following "corpus" of two documents:

  1. how much wood would a wood chuck chuck if a wood chuck would chuck wood
  2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

  - (a) Which of the following sentences: a wood could chuck; wood would a chuck; is more probable, accoding to:

    i. An unsmoothed uni-gram language model?
    ii. A uni-gram language model, with Laplacian ("add-one") smoothing?
    iii. An unsmoothed bi-gram language model?
    iv. A bi-gram language model, with Laplacian smoothing?
    v. An unsmoothed tri-gram language model?
    vi. A tri-gram language model, with Laplacian smoothing?