Assignment 1, Cross-Validation, Text Vectorization

COMP 4211 - Tutorial 03

Chun-Kit Yeung

Hong Kong University of Science and Technology

2018-03-02

Assignment 1

Assignment 1

- There are two problem in the assignment, one written problem (Q1) and one programming problem (Q2).
- You should hand in the **hardcopy** of both Q1 and Q2 at the beginning of the tutorial on Mar 16 (Fri), 2018.
- Also, you have to submit your code (in .ipynb format) to **CASS system**.[1]
- For details, you can refer to the .pdf file of assignment 1.

[1] CASS user guide:
`http://cssystem.cse.ust.hk/UGuides/cass/student.html`

Dataset for Q2

20 newsgroups dataset:

- It is a collection newsgroup documents.
- It has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.
- It comprises around 18000 newsgroups posts on 20 topics, and you will get a better understanding in the following.

Q2 Task

Specifically, only two groups of news are required in assignment 1

1. `comp.sys.mac.hardware`
2. `sci.space`

Your task is to classify a given document to either `comp.sys.mac.hardware` or `sci.space` using naïve Bayes classifier with Laplace smoothing.

Eventually, evaluate your model with 5-fold cross-validation.

Cross-Validation

Machine Learning Workflow

In the last tutorial, we mentioned the general machine workflow is as follow:

1. Collecting data
2. Preparing data
3. Choosing a model
4. Training a model
5. Evaluating a model $\leftarrow$ Using train-test-split

Machine Learning Workflow

In this tutorial, we going to evaluate our model by cross-validation:

1. Collecting data
2. Preparing data
3. Choosing a model
4. Training a model
5. **Evaluating a model** ← Using cross-validation

How cross-validation works

- Cross-validation randomly divides the training data into a number of partitions, also called **folds**.
- For example, in ten-fold cross-validation, ten models are actually created in cross-validation. Each model trained using 9/10 of the data, and tested on the remaining 1/10.

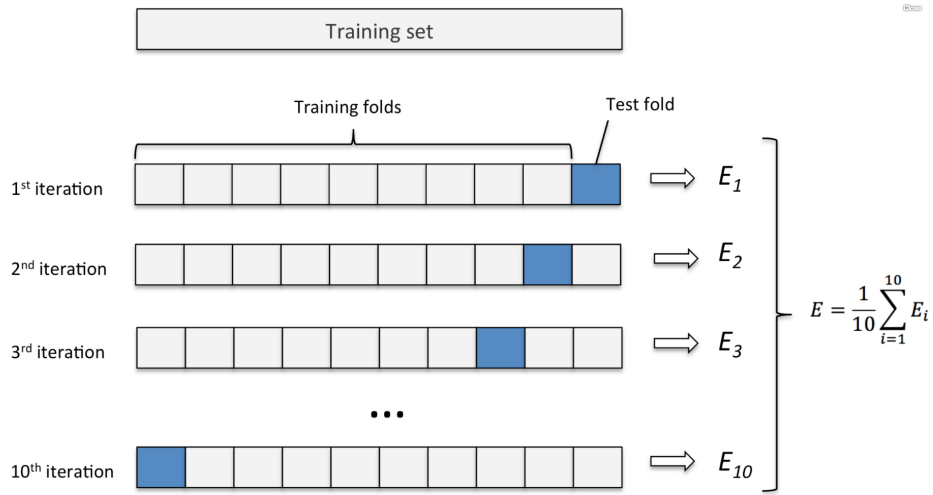

Image extracted from Quora.

Pros and Cons of cross-validation

Pros:

- Use more test data - cross-validation uses the entire training dataset for both training and evaluation, instead of some portion.
- Evaluates the dataset as well as the model - cross-validation gives you some idea of how representative the dataset is and how sensitive the model might be to variations in the data.

Cons:

- Take longer time than the usual train-test split introduced in last tutorial $\rightarrow$ computationally intensive.

Text Vectorization

Bag-of-Words representation

- Term as the basis for vector space
  - Doc1: Text mining is to identify useful text information.
  - Doc2: Useful text information is mined from text.
  - Doc3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

Bag-of-Word with Word Counts

Bag-of-Words with Counts

- Term as the basis for vector space
  - Doc1: Text mining is to identify useful text information.
  - Doc2: Useful text information is mined from text.
  - Doc3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	2	1	1	1	0	1	1	1	0	0	0
Doc2	2	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

Stop Words Removal

Stop words usually refers to the most common words in a language, such as *a, an, the, is, are, from, to, ...,* and so on. It contains not much information about a text and therefore we usually remove them in text mining.

Question: What is the resulted vector representation in the above example after removing the stop words? (Given {a, an, the, is, are, from, to} is the stop words.)

Let's code.

To better understand today tutorial, the following .ipynb is covered:

- T03a_countervectorizer . . . using_sentiment_analysis.ipynb
- T03b_intro_to_assignment1_dataset.ipynb