MSiA414 SEC01
Text Analytics
Lab 4 - Text Classification

Timo Wang

Northwestern University

October 14, 2020

Overview

- What is text classification?
- How do we approach classification?
- How do we approach text classification?

What is classification?

The goal of text classification is to assign pieces of text (such as reviews, emails, etc.) to one or more categories such as positive/negative sentiment, spam email or not, etc..

What is text classification?

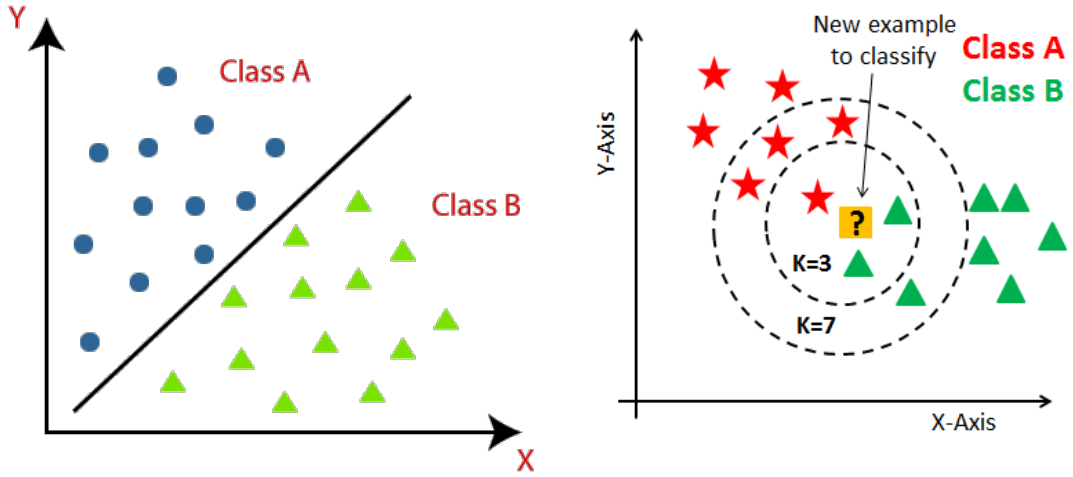

Figure: Given a set of feature vectors, e.g. sentence embeddings, word embeddings, bag-of-words, etc., separate them from each other, either linearly (left) or non-linearly (right)

How do we approach classification?

Simple solution

Use existing packages/libraries!

- **scikit-learn** is a "simple and efficient tools for predictive data analysis".
- **keras** is a high-level API over machine learning frameworks such as TensorFlow or Theano that is easier to use and has simpler syntax.
- **FastText** is a library/command line tool that learns text representations and serves as text classifiers.

How do we approach classification?

Data preparation

X: a 2-d array of size (n, d), where n is the number of training examples and d is the size of the feature.
y: a 1-d array of length n, where n is the number of training examples.

How do we approach classification?

Model fitting and prediction

Python

```
model.fit(X[:8,:], y[:8])
y_pred = model.predict(X[8:,:], y[8:])

accuracy_score(y, y_pred)
f1_score(y, y_pred)
```

How do we approach text classification?

General steps

Step 1 Study the content of your dataset and identify your task.

Step 2 Transform input text into some vectorized representation. (bag-of-word, BERT, average of word embeddings, etc.)

Step 3 Study the vectorized representations (through visualization using matplotlib)

Step 4 Choose a classifier model. (logistic regression, SVM, fasttext, etc.)

Note: if you use FastText, you can skip the first three steps.

Useful resources

1. Text classification with scikit-learn
2. Text classification with Keras
3. Text classification with fastText