MSiA490 SEC20/28
Text Analytics
Lab 5 - Text Classification

Timo Wang

Northwestern University

October 15, 2020

Overview

What is classification?

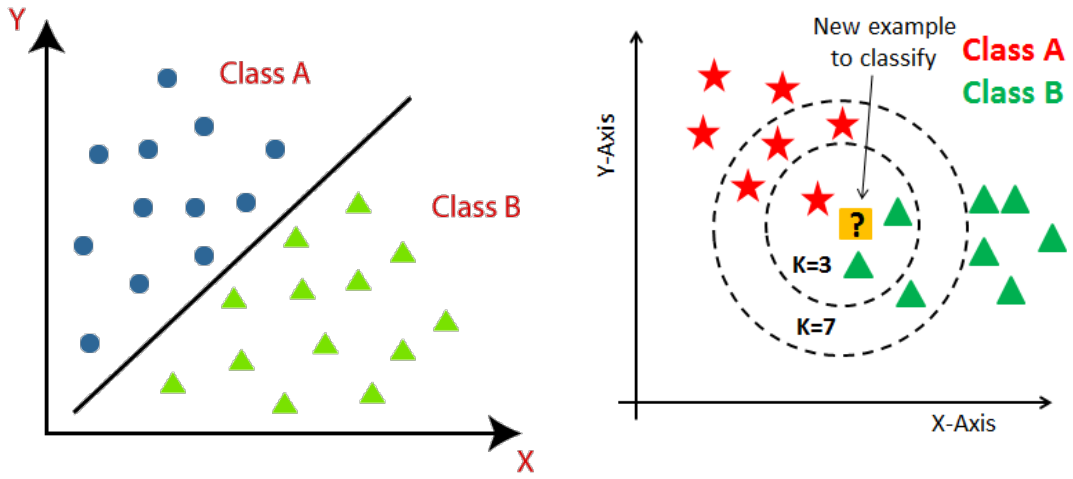

Figure: Given a set of feature vectors, e.g. sentence embeddings, word embeddings, bag-of-words, etc., separate them from each other, either linearly (left) or non-linearly (right)

How do we approach classification?

Simple solution

Use existing packages/libraries!

- **scikit-learn** provides both SVM models as well as logistic regression models.
- **FastText** is an alternative tool that also works as command line programs.

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 pipeline

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.