IN323: Probability and Statistics

Fall, 2015

<Deadline for submission: 2015. 12. 3(목)>

1. Introduction

a. Objectives

- 확률 모델링에 의한 문제 해결방법 습득.
- Naïve Bayes 분류기에 대한 이해.

b. Overview

Bayes 분류기를 이용하여 주어진 이메일 데이터에 대해 스팸 필터링.

c. Attachment

- .\resource\Spambase.txt
  : 단어 출현에 따른 스팸여부 판단 데이터베이스(4601개의 메일)
  54개의 단어, 숫자 및 특수문자의 출현여부, 1개의 스팸 여부에 대한 Labeling으로 이루어짐. 즉, 한 메일에 대한 분석 데이터는 55개의 0 또는 1 값을 가짐.
- .\resource\Spambase.xlsx: Spambase.txt파일의 엑셀문서(데이터베이스 분석용)
- .\resource\ham*.txt: 스팸이 아닌 메일의 테스트 셋(2000개의 메일)
- .\resource\spam*.txt: 스팸 메일의 테스트 셋(2000개의 메일)

2. Assignment

주어진 스팸 이메일 데이터베이스를 분석하여 Bayes 분류기를 설계 및 구현하고, 테스트 메일의 스팸여부를 판별.

1) Spam filtering 수행과정

a. Bayes learning

데이터베이스로 주어진 파일(Spambase.txt)을 이용하여 분류기에서 사용될 매개변수를 산출.

- Input: Spamdata.txt
- Output: Naïve Bayes Parameter (Likelihood, Prior, …….)

b. Parser

테스트 셋으로 주어진 메일을 분석하여 단어 출현여부에 대한 벡터를 구함.

- Input: Test set E-mail text file
- Output: Vector

c. Classification

이전 Parser 단계에서 출력된 벡터와 Bayes 단계의 매개변수들을 이용하여 스팸여부를 판별.

- Input: Vector from the previous step(Parser).
- Output: Classify – Spam/Non-spam

d. 5-fold cross-validation 수행

각 클래스에 해당하는 학습 데이터를 5부분으로 나눠 총 6번의 분류 수행(5번은 일부 학습 데이터로 학습된 분류기, 6번째는 모든 학습 데이터로 학습된 분류기로 분류된 결과).

e. mySpamfilter 제작

Bayes 모델을 확장하여 나만의 mySpamfilter 분류기 제작

2) 제출방법

a. 실행파일 만들기: spamfilter.exe 혹은 spamfilter.bat

b. 실행파일 사용법: spamfilter [테스트데이터 디렉토리]

예) spamfilter D:\TestData\

※ 테스트데이터 디렉토리 하위에는 .\spam과 .\ham 두 개의 디렉토리만 존재.

c. 화면 출력

학번

Classifier 1: True positive/negative, False positive/negative, Precision, Recall

Classifier 2: True positive/negative, False positive/negative, Precision, Recall

...

Classifier 6: True positive/negative, False positive/negative, Precision, Recall

(콘솔 출력)

d. 제출사항: 실행파일, 소스코드, 보고서(자유양식)을 zip파일로 압축하여 업로드.

3) 평가방법 : 공개된 데이터 셋 테스트 (30%) 비공개 데이터 셋 테스트 (70%)

※ 참고자료

· True positive/negative, False positive/negative

A true positive test result is one that detects the condition when the condition is present.
A true negative test result is one that does not detect the condition when the condition is absent.
A false positive test result is one that detects the condition when the condition is absent.
A false negative test result is one that does not detect the condition when the condition is present.

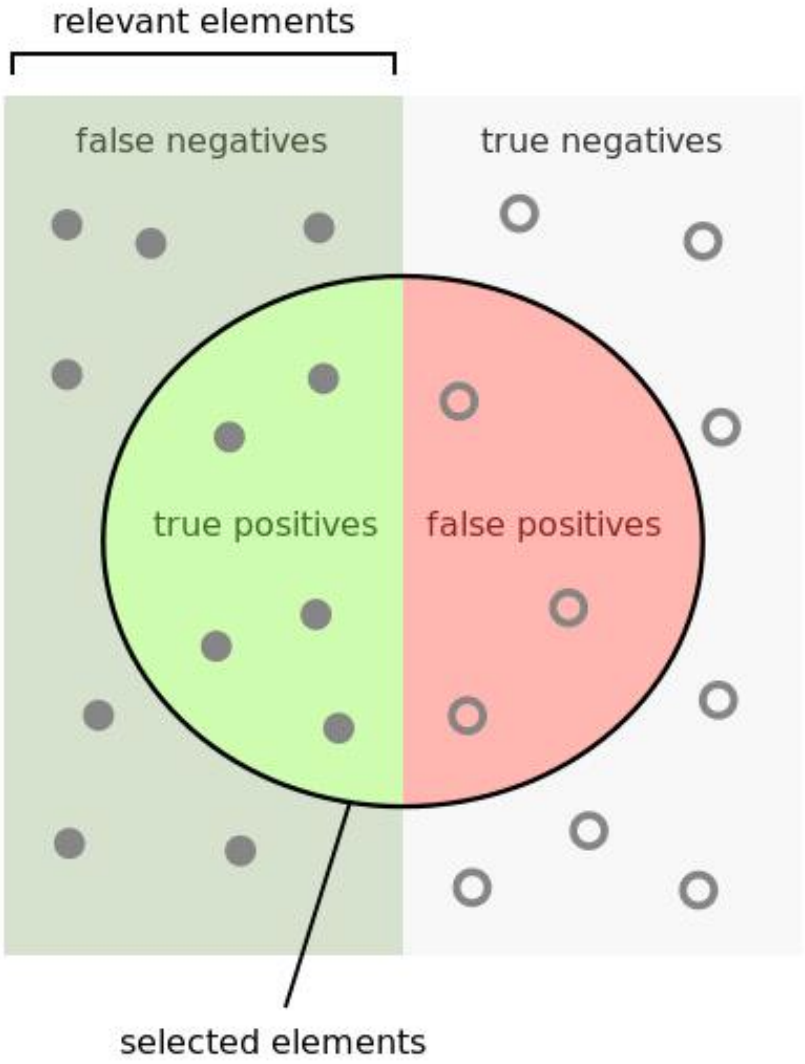
		Condition (Gold standard)	
		True	False
Test outcome	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{True positive rate} = \frac{TP}{TP+FN} \quad \text{true negative rate} = \frac{TN}{TN+FP}$$

$$\text{(total) accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$$

$$\text{positive predictive value} = \frac{TP}{TP+FP} \quad \text{negative predictive value} = \frac{TN}{TN+FN}$$

· Precision/Recall



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$