1. 개요

Naïve Bayes를 이용한 Spam Filtering

사용언어: C++

IDE: Visual Studio 2013

2. 필요한 기능

- A. Bayes learning
- B. Parser
- C. Classification

3. 구현

Bayes Learning.cpp

Spambase.txt를 분석하여 6부분으로 나눈데이터에 대해 스팸 일 때, 아닐 때 단어들의 value를 구한다. 구한 parameter들을 "parameter.h"로 전달하였다.

parameter.h

Classification.cpp

Parser

입력 받은 경로의 spam과 ham 디렉토리에 있는 txt파일을 읽는다. 54개의 단어에서 해당 단어가 존재하는지 vector에 기록한다.

 P(spam)과 P(ham)은 learning에 사용한 spam의 개수와 ham의 개수로 초기화

 vector에서 단어가 존재하면 P(spam) *= P(해당단어|spam), P(ham) *= P(해당단어|ham)

 없으면 P(spam) *= 1 - P(단어|spam), P(ham) *= 1 - P(단어|ham)

 모든 단어에 대해 수행한 후, 확률이 더 큰 쪽으로 분류한다.

5-fold cross-validation

총 2788개의 ham과 1813개의 spam을 아래의 개수로 나눠 학습하였다.

case	1	2	3	4	5	6
ham	400	800	1200	1700	2200	2788
spam	300	600	900	1200	1500	1813

4. 결과 (공개된 데이터 셋)

C:\Users\HODONG\Documents\Uisual Studio 2013\Projects\SpamFilter\Kelease>SpamFi ONG\Desktop\p\resource\								
12111622 이호동								
Classifier 1:	1607	1537	463	393	0.776329	0.8035		
Classifier 2:	1586	1452	548	414	0.743205	0.793		
Classifier 3:	1600	1477	523	400	0.75365 0.8			
Classifier 4:	1591	1470	530	409	0.750118	0.7955		
Classifier 5:	1607	1459	541	393	0.748138	0.8035		
Classifier 6:	1542	1525	475	458	0.764502	0.771		

5. 개선 (mySpamfilter)

12111622 이호동	5					
Classifier 1:	1597	1596	404	403	0.798101	0.7985
Classifier 2:	1603	1568	432	397	0.787715	0.8015
Classifier 3:	1616	1604	396	384	0.803181	0.808
Classifier 4:	1608	1596	404	392	0.799205	0.804
Classifier 5:	1633	1595	405	367	0.801276	0.8165
Classifier 6:	1586	1581	419	414	0.791022	0.793

개선한 점:

어떤 단어가 spam과 ham에서 나올 확률차이가 클 때 약간의 가중치를 부여 parsing 해서 나온 단어가 parameter에서 확률이 낮을 때 P(spam)과 P(ham)의 확률을 더 많이 낮추도록 가중치를 부여

parsing 해서 나오지 않은 단어가 parameter 값이 크다고 해서 P(spam)과 P(ham)의 확률을 너무 많이 낮추지 않도록 보정