Name:

Giovanni Zavalza

Netid:

gzava3

CS 441 - HW3: PDFs and Outliers

Complete the sections below. You do not need to fill out the checklist.

Total Points Available	**[ ] / 150**
1. Spam Detection	
a. Classifier Training, Val Accuracy	[ ] / 15
b. Data exploration	[ ] / 20
c. Precision-recall trade-off	[ ] / 15
2. Robust Estimation	
a. Assume no noise	[ ] / 10
b. Robust estimation with percentiles	[ ] / 15
c. Robust estimation with EM	[ ] / 25
3. Stretch Goals	
a. Improvements to Spam Detection	[ ] / 20
b. Impact of school on salary	[ ] / 15
c. Impact of experience on salary	[ ] / 15

1. Spam Detection

a. Classifier Training, Val Accuracy

Validation accuracy (xx.x%)

98.13%

Should be higher than 95%

b. Data Exploration

10 spammiest words

['18' 'cs' '16' '500' 'tone' 'www' '150p' 'uk' 'prize' 'claim']

‘prize’ is one of the words

10 hammiest words

```
['gt' 'lt' 'he' 'but' 'lor' 'da' 'she' 'later' 'ì_' 'wat']
```

Spammiest ham (highest spam score, true label ham)

> Waqt se pehle or naseeb se zyada kisi ko kuch nahi milta,Zindgi wo nahi he jo hum sochte hai Zindgi wo hai jo ham jeetey hai..........

Hammiest spam (lowest spam score, true label spam)

> LIFE has never been this much fun and great until you came in. You made it truly special for me. I won't forget you! enjoy @ one gbp/sms

Spammiest spam (highest spam score, true label spam)

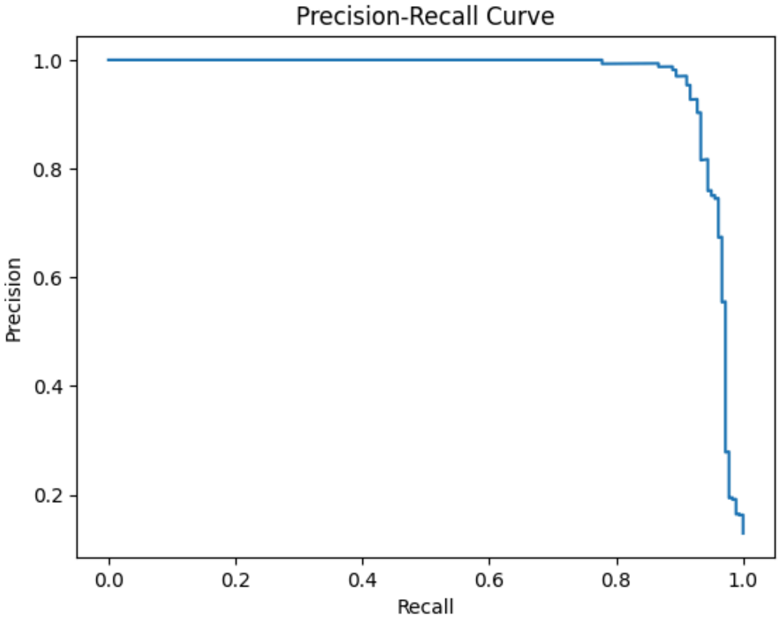
> FREE for 1st week! No1 Nokia tone 4 ur mob every week just txt NOKIA to 8007 Get txting and tell ur mates www.getzed.co.uk POBox 36504 W45WQ norm150p/tone 16+

Hammiest ham (lowest spam score, true label ham)

> Sad story of a Man - Last week was my b'day. My Wife did'nt wish me. My Parents forgot n so did my Kids . I went to work. Even my Colleagues did not wish. As I entered my cabin my PA said, " Happy B'day Boss !!". I felt special. She askd me 4 lunch. After lunch she invited me to her apartment. We went there. She said," do u mind if I go into the bedroom for a minute ? " "OK", I sed in a sexy mood. She came out 5 minuts latr wid a cake...n My Wife, My Parents, My Kidz, My Friends n My Colleagues. All screaming.. SURPRISE !! and I was waiting on the sofa.. ... . ' NAKED...!

c. Precision-recall trade-off

Precision-Recall Curve on val set



Selected threshold (x.xx) using val PR curve (P>0.99)

6.10

Test Accuracy (xx.x%) with threshold

97.6%

Test Precision (x.xx) with threshold

1.00

Test Recall (x.xx) with threshold

0.81

2. Robust Estimation

Round to nearest whole number.

	a. No noise	b. Percentiles	c. EM
Min	64694	75,494	64,694
Mean	123,750	113,879	111,984

Std	61,954	15,876	17,966
Max	611,494	159,901	169,008

For 2b, answer could also be 159,160 for a different way of getting percentiles.

First five indices of invalid data (based on EM solution, you add last 3)

18	28	49	127	128

3. Stretch Goals

a. Spam detection improvements

What did you try to improve the spam detection?

What worked best to improve the spam detection?

What was your test precision and recall at the threshold selected based on the validation set?

Precision

Recall

b. Impact of school on salary

Report mean salary overall and for each school

	Average Salary
Overall	
School 0 (UIUC)	
School 1 (MIT)	
School 2 (Cornell)	

Describe your approach to estimate this.

c. Impact of years of experience on salary

How much are salaries expected to increase with one year of experience?

Describe your approach to estimate this.

Acknowledgments / Attribution

List any outside sources for code or ideas or “None”.