Spam Classification

Cynthia Wang

Introduction

- Report from Cloudmark: the number of unique SMS spam campaigns grew by 300% from 2011 to 2012.
- Types of spam messages
  - 92% fall into scam/fraud category
  - 63% receive a gift card spam
  - 12% iPad/iPhone test and keep spam
  - 5% need cash now spam
- Consequences
  - Personal information collected by "surveys"
  - Malicious contents downloaded to the phone
  - Pay for "products" or "services"

Dataset

- SMS Spam Collection v. 1
- 5574 text messages from UCI Machine Learning Repository
- # Spam: 747
- # Ham: 4827

ham   What you doing?how are you?
ham   Ok lar... Joking wif u oni...
ham   dun say so early hor... U c already then say...
ham   MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham   Siva is in hostel aha:-.
ham   Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam   FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam   Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B
spam   URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

.txt file

Pre-processing

- Tokenization
- Stop words removal
- Lemmatization
  - Spacy (determines POS tag by default)

Index	body	label
0	go jurong point crazy available bugis n grea…	ham
1	ok lar joke wif u oni	ham
2	free entry 2 wkly comp win fa cup final tkts…	spam
3	u dun say early hor u c already say	ham
4	nah i think go usf live around though	ham
5	freemsg hey darling 3 week word back i like …	spam
6	even brother like speak -PRON- treat like ai…	ham
7	as per request melle melle oru minnaminungin…	ham
8	winner as value network customer select rece…	spam
9	have mobile 11 month u r entitle update late…	spam
10	i gon na home soon want talk stuff anymore t…	ham
11	six chance win cash from 100 20 000 pound tx…	spam
12	urgent -PRON- 1 week free membership 100 000…	spam
13	i search right word thank breather i promise…	ham

Pre-processing

- Train/Test Split (80% training, 20% testing)
  - X_train, X_test, y_train, y_test = train_test_split(message_df, y, test_size=0.2, random_state=42)
- TF-IDF
- Principal Component Analysis
  - Target variance: 95%
  - 2196 features
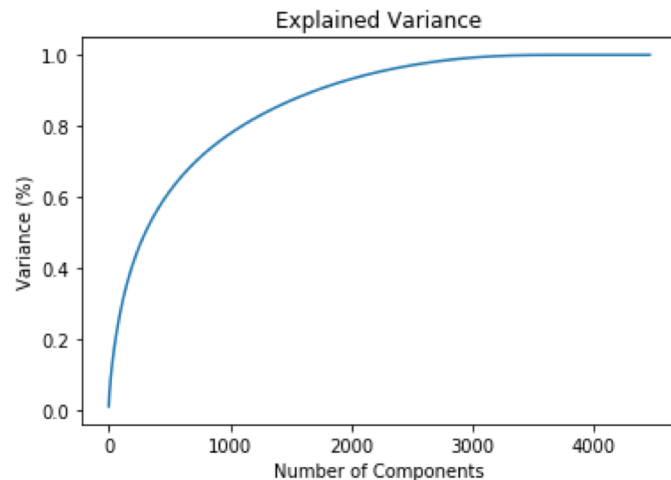
Training

- Models
  - Random Forest
  - Logistic Regression
  - Multinomial Naive Bayes
  - Support Vector Machine
- Hyperparameter tuning: Grid search
  - 5-fold cross validation
  - Except Naive Bayes

Random Forest

{'max_depth': 50, 'n_estimators': 16, 'random_state': 1234}

Accuracy: 93.5%

Precision: 100%

Recall: 54.7%

F-score: 70.7%

Validation Score: 93.0%

Index	pred:ham	pred:spam
true:ham	954	0
true:spam	73	88

Logistic Regression

{'C': 1000.0, 'penalty': 'l2'}

Accuracy: 98.5%

Precision: 98.0%

Recall: 91.3%

F-score: 94.5%

Validation Score: 98.2%

Index	pred:ham	pred:spam
true:ham	951	3
true:spam	14	147

Multinomial Naive Bayes

Accuracy: 97.0%

Precision: 100%

Recall: 78.9%

F-score: 88.2%

Validation Score: 96.3%

Index	pred:ham	pred:spam
true:ham	954	0
true:spam	34	127

Support Vector Machine

{'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}

Accuracy: 93.5%

Precision: 98.7%

Recall: 94.4%

F-score: 96.5%

Validation Score: 98.4%

Index	pred:ham	pred:spam
true:ham	952	2
true:spam	9	152

Summary of Models

- **Highest accuracy**: logistic regression (98.5%)
- **Highest precision** (lowest false positive rate): random forest, multinomial Naive Bayes (100%)
- **Highest recall** (lowest false negative rate): support vector machine (94.4%)

Final Thoughts

- Model performance is surprising
- Accuracy is not everything
    - Low false positive rate might be more important?
- More models?
- Topic modeling?
- Email spam?

Thank you!