Spam classification using SVM and naive Bayes model

Yang Li

Classifier Evaluation Metric

Goal

Propose a classifier model for spam email classification

Evaluation Metric

To evaluate a classifier model, we could compare the accuracy, the complexity of the algorithm and the time it takes to make predictions[1]. The most important evaluation metric is accuracy. There are various classification models(SVM[2], naive Bayes[3]), To improve the accuracy, this project will study the influence of learning rate and different features extractor[4] to the SVM model accuracy.

Data Set and Feature Extractor

Data

- The dataset is downloaded from the website of "http://www.dt.fee.unicamp.br/~tiago/smsspamcollection".
- The data is splitted into a training set(4457 messages) and a test set(558 messages).Each message is a paragraph of email labeling either as "spam" or "ham" in the beginning. Two examples of (x,y) are shown below

Label (y)	Message (x)
Ham	My phone
Spam	Thanks for your subscription to Ringtone UK your mobile will be charged £5/month Please confirm by replying YES or NO.

Feature Extractor

- A unigram model that mapped each word to the number of occurrences of that word in the message was developed.
- For example, the first message x in table above will be mapped into feature $\phi(x) = \{My:1,phone:1\}$ and label "ham" will be mapped to y=-1(i.e. not spam). The "spam" label corresponds to y=1.

Model and Algorithm

Loss Function

- The loss function for SVM is the hinge loss:

$$Loss_{hinge}(x, y, w) = \max\{0, 1 - w\phi(x)y\}$$

Stochastic Gradient Descent

- The loss function is the hinge loss:

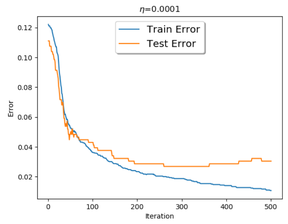
$$w \leftarrow w - \eta \nabla_w Loss_{hinge}(x, y, w)$$

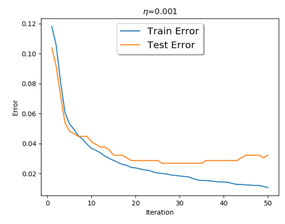
$$\nabla_w Loss_{hinge}(x, y, w) = \begin{cases} 0, w\phi(x)y \geq 1 \\ -\phi(x)y, w\phi(x)y < 1 \end{cases}$$

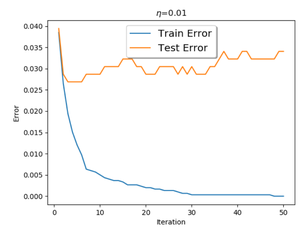
Results

The learning rate η .

η	Train Set Error	Test Set Error
0.01	0	0.0287
0.001	0.011	0.0268
0.0001	0.011	0.0305







- When $\eta = 0.01$, the error is bigger than the other two cases.
- When $\eta = 0.0001$,the error converges very slowly(around 200 iterations).
- The optimized learning rate is $\eta = 0.01$, with better accuracy and run time

Using specific spam words as features

	Train Set Error	Test Set Error
Original features	0.011	0.0268
Specific spam features	0.011	0.0197

- The model with specific spam words features requires less time to converge, and the accuracy is better.
- The frequent spam words were given a bigger counts to improve the accuracy of predicting spam emails. Specifically, they are ['claim', 'won','prize','tone','urgent!', 'babe', 'Text','text','Txt', 'texts', 'sexy', 'subscribed' ','Call', 'call', 'win', '@', 'link']

Future work

- Try other features that may improve the model predictions
- Use bi-gram model instead of the unigram model
- Use other data set to test current model(such as Enron Email Dataset)
- Compare the SVM algorithm with the naive Bayes model

References

[1] Koller, Daphne, and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
[2] Caruana, Godwin, Maozhen Li, and Man Qi. "A MapReduce based parallel SVM for large scale spam filtering." *2011 eighth international conference on fuzzy systems and knowledge discovery (fskd)*. Vol. 4. IEEE, 2011.
[3] Pantel, Patrick, and Dekang Lin. "Spamcop: A spam classification & organization program." *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*. 1998.

Stanford University