Evaluation of email classification methods

The problem statements

This study is about designing a spam filter that can separate ham and spam emails based on various machine learning techniques. The subject line of the emails is analyzed based on the NLP technique of tokenizing sentences. Two methods of Machine learning algorithms are used (1) dimensionality reduction technique PCA is applied to Logistic regression and (2) 1D convolution neural network to train a model on training set and then asses it on a test set. Finally, the models are evaluated on bigdata set of 58000 emails and accuracy and precision is tested.

Dataset

The dataset is based on cleaned Enron corpus, there are a total of 92188 messages belonging to 158 users with an average of 757 messages per user. The dataset has almost an equal distribution of ham and spam emails. In this study 2000 emails are used 1000 ham + 1000 spams. Each email text

Parameter	Example
Body	"Carolina Power & Light and Florida Power Corporation submitted tariff revisions in compliance with a Commission order addressing the energy imbalance provisions that apply in CP&L's zone. \n\nThe proposed revisions include:\n\t- return-in-kind provisions for energy imbalances\n\t- deletion of the separate capacity charge for undersupply of energy outside the deadband for 10+ hours in a month\n\t- a provision that deficient energy will be offset or credited with energy associated with spinning and supplemental reserves.\n\nInterventions/protests are due Aug. 15.\n\nIf you would like further information please contact me.\n\nSusan Scott Lindberg\nx30596"
Date	2001-08-05 18:34:50
From	[(Scott, Susan, Susan.Scott@ENRON.com)]
To	[('Acevedo, Rudy', 'Rudy.Acevedo@ENRON.com'), ('Carson, Mike', 'Mike.Carson@ENRON.com'), ('Comeaux, Keith', 'Keith.Comeaux@ENRON.com'), ('Connor, Joe', 'Joe.Connor@ENRON.com'), ('Fairley, David', 'David.Fairley@ENRON.com'),]
message_id	'<902B8E00B151D44C98CA48BDD4BEA3F5082C01@NAHOU-MSMBX01V.corp.enron.com>'
subject	'CP&L tariff changes (ER01-1807)'

Table 1. Parameters extracted from emails. Subject line is used in this study

is preprocessed with python library mail parser to extract various features. The features extracted are shown in table 1.

Processing on email subjects

The preprocessing of email subject lines includes tokenize the email subject lines. The python library spacy is used to in this case. This method counts the number of times for the occurrences of tokens. The dataframe thus created has > 3500 columns based on the individual tokens identified by the algorithm. In this problem there are more attributes than the number of rows. However, since the dimensionality of the problem is very high its needs to be reduced to lower dimensions by *latent semantic indexing*.

PCA with logistic regression

PCA reduces the dimensionality hence the complexity while maintaining structure (variance) of a dataset. It performs a rotation of the data that maximizes the variance in the new axes.

By combining the advantages of both the PCA and logistic regression, PCA is used to extract feature and reduce the dimensions of process data. Afterwards logistic regression is used as the classifier for spam and ham emails.

The scatter matrix is used to extract feature, and then obtain all the individual characteristics subspace W_i, i = 1, 2, . . ., m. First the eigenvalues and the corresponding eigenvectors is used to generate matrix. Second the eigenvalues are order from largest to smallest, and similarly putting the corresponding eigenvectors in order of largest

to smallest, the optimal projection matrix (X_1, X_2, . . ., X_d) is thus created and is

associated with the d largest generalized eigenvalues. Finally, logistic regression will be used as the classifier of ham and spam emails.

Logistic regression is used classify the emails based on > 3500 attributes. A set of regularization parameter in C is used to reduce overfitting, C=0.1 performs the best on both train-test set.

PCA and logistic regression combination are tested with 3 different samples sizes 16, 160 and 1600. Sample size of 160 have best accuracy and precision values on training and test set.

Convolution neural network:

In the model architecture of the CNN $x_i \in R^k$ corresponds to the k-dimensional tokenizer with i-th word in the sentence. A sentence of length n (padded where necessary) is represented as

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \ldots \oplus \boldsymbol{x}_n$$

In general, let $x_{i:i+j}$ refer to the concatenation of words $x_i, x_{i+1}, \ldots, x_{i+j}$.

A convolution operation involves a filter $w \in R^{hk}$, which is applied to a window of h words to produce a new feature. For example, a feature c_i is generated from a window of words $x_{i:i+h-1}$ by

$$c_i = \int (\boldsymbol{w}.\boldsymbol{x}_{i:i+h-1} + b)$$

Here $b \in R$ is a bias term and f is a non-linear function such as sigmoid. This filter is applied to each possible window of words in the sentence $\{x_{1:h}, x_{2:h+1}, \ldots, x_{n-h+1:n}\}$ to produce a feature map with $c \in R^{n-h+1}$. A max pooling operation is used over the feature map and take the maximum value as the feature corresponding to this filter. These features form the penultimate layer and are passed to a fully connected sigmoid layer whose output is the probability distribution over labels. For regularization dropout is used on the penultimate layer.

Hyperparameters and Training

In the problem rectified linear units is used, filter windows (h) of 2, 5, 160 with 1000 feature maps each, with dropout rate (p) of 0.2 is used with mini-batch size of 32 and epoch length of 30 is used. The metrics that is monitored in every epoch is *accuracy* and *entropy loss*. An early stopping criterion is used based on accuracy. If the accuracy does not improve in 10 epoch then iterations stops.

Method	Parameters	Training Set						Test Set					
		True_pos	False_pos	False_neg	True_neg	Accuracy	Precision	True_pos	False_pos	False_neg	True_neg	Accuracy	Precision
logistic regression	C=0.01	519	19	281	781	0.81	0.65	109	11	91	189	0.75	0.55
logistic regression	C=0.1	608	15	192	785	0.87	0.76	126	23	74	177	0.76	0.63
logistic regression	C=1	759	11	41	789	0.97	0.95	146	32	54	168	0.79	0.73
logistic regression	C=10	800	17	0	783	0.99	1.00	182	91	18	109	0.73	0.91
logistic regression	C=100	800	16	0	784	0.99	1.00	183	100	17	100	0.71	0.92
PCA+ logistic regression	Feature_columns = 16	480	37	320	763	0.78	0.60	111	11	89	189	0.75	0.56
PCA+ logistic regression	Feature_columns = 160	629	45	171	755	0.87	0.79	140	34	60	166	0.77	0.70
PCA+ logistic regression	Feature_columns = 1600	800	17	0	783	0.99	1.00	181	91	19	109	0.73	0.91

Table 2. Metrics from PCA with logistic regression model.

Filter window length (h)	Training Set		Test Set	
	Accuracy	Loss	Accuracy	Loss
2	0.88	0.24	0.79	0.45
50	0.91	0.24	0.74	0.76
160	0.97	0.15	0.78	1.03
1000	0.89	2.27	0.82	4.01

Table 3. Metrics obtained from CNN model.