Mail Classifier

Programming Club Summer Project 2017

8th July,2017

Team members :

- Aman Gupta
- Kshitiz Tyagi
- Arnav Garg

Mentor: Varun Khare

1 Aim

This is a project aiming to classify mails depending on parameters set by the user.
People receive dozens of mails each day and depending on time they have preferences as to what kind of mails they want to view. The goal of this application is to solve this problem.
The application has been trained by tagging a sufficient amount of previous emails. Then the program is trained on these emails, after which given a new email, it can predict which category it belongs to. The application has been implemented in Python.

2 Study Resources

- Machine Learning by Andrew Ng- Coursera Course
- Introduction to Machine Learning - Udacity Course
- Learn Python the Hard Way - Online Ebook

3 Timeline

1. Week 1 (20th-26th May)

   - Learnt basic commands in python from the book 'Learn Python the Hard Way'- printing,input methods, escape sequences, python format characters, unpacking variables,reading and writing files,if and else statements,loops etc
   - Learnt more advanced concepts of python such as functions, lists, dictionaries, modules, classes and objects.

2. Week 2 (27th May-2nd June)

   - Started with the Coursera course of Machine Learning by Andrew Ng.Got a basic idea of what machine learning is and its two types-Supervised learning and Unsupervised learning.
   - Learnt about Linear regression- linear regression with one variable,cost function,gradient descent,multivariate linear regression,feature scaling,mean-normalization,learning rate.
   - Learnt about Polynomial regression and how to compute parameters analytically.Studied hypothesis representation and interpretation of hypothesis output.Learned about decision boundary-linear and non linear.
   - Studied the logistic regression model. Learnt about advanced optimization algorithms.Got the hang of regularization.

3. Week 3 (3rd-9th June)

   - Learnt about Neural Networks-multiclass classification using neural networks,back propagation algorithm, random initialization,training neural networks.
   - Learnt how to evaluate learning algorithms and how to select models by dividing data into training/cross validation/test sets ,diagnosing bias vs variance and learning curves.
   - Studied about the error analysis of learning algorithm - handling skewed data,error metrics for skewed classes,Precision/Recall ,F score.
   - Studied Support Vector Machines (SVMs)-large margin classification.Learnt about kernels and similarity functions.

4. Week 4 (10th-16th June)

   - Learnt about Unsupervised learning - clustering,K-means algorithm,random initialization,choosing number of clusters and elbow method.
   - Read about Principal Component Analysis (PCA)-PCA algorithm,reconstruction from compressed representation,choosing number of principal components.

- Studied about Anomaly Detection using density estimation and Gaussian distribution.
- Started the Udacity course 'Introduction to Machine Learning'.
- Learnt to use scikit-learn - machine learning library for Python.
- Learnt about the Naive Bayes classification algorithm. Did a mini project involving the classification of emails by the authors, learning how to run this classifier.

5. Week 5 (17th-23rd June)

   - Learnt about two more classification algorithms- SVMs and Decision Trees.Learnt how to implement them on the email classification problem using scikit-learn.Compared accuracy and training and testing time of different algorithms.
   - Worked on the Enron data set (provided by the course) and learnt various things involving dealing with large datasets and getting useful information from it.
   - Learnt about Regressions to model continuous data and its implementation.Studied about R-squared metric for regression,multivariate regression,regression score.
   - Studied about Outliers in a dataset-learnt algorithms for identifying and removing them.

6. Week 6 (24th-30th June)

   - Learnt about feature scaling - it's need and Max/Min scaler in Sklearn.
   - Learnt about Text learning -How to use text data in our machine learning algorithm.Learnt about bag of words in sklearn and various text learning techniques like stemming,removing stopwords,parsing,weighing by term-frequency,cleaning away signature words and using the TfIdf vectorizer.
   - Learnt about feature selection- dealing with high bias and variance,applying regularization and Lasso regression, improving accuracy on the test data by removing overfitting.
   - Learnt about data dimensionality and reducing the number of dimensions with principal component analysis (PCA).

7. Week 7 (1st-7th July)

   - Started coding up our mail classifier to classify our webmail emails into three classes -'Club Mails,Internship Mails,Nominations Mails.
   - Took a dataset of 30 mails from each class, parsed all the text, and trained our SVM classifier on this data.
   - Optimized the parameters of our Svm classifier according to the training accuracy.

- Tested our classifier on new mails.

4 GitHub Repository

https://github.com/amangupta87/Mail-Classifier