Predict email open

Manuel Felipe Pineda

Introduction to data science
Universidad Tecnologica de Pereira

November 29, 2016

Overview

Experiments

Each experiment was tested using the cross-validation technique with the k-fold method, with k equals to 10, in order to evaluate the estimators performance. [1] [2]
Each experiment was evaluated based in the f1-score and the accuracy-score, were run in 4 cores and 4 GB of RAM [1]

[1] Intel(R) Core(TM) i5-4300U CPU

Dimensionality reduction

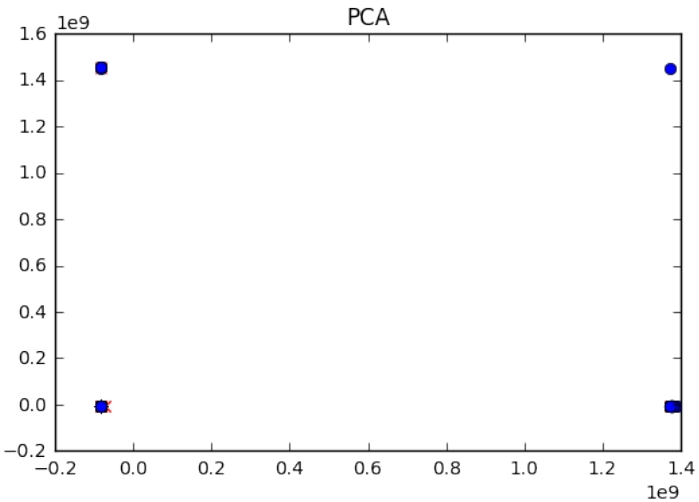

Figure: PCA

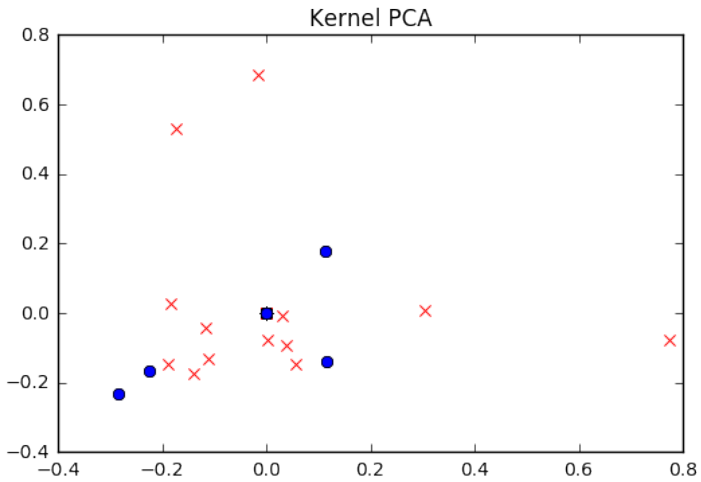

Figure: Kernel - PCA

Estimators I

Table: Comparison between estimators

Method	Accuracy	F1 Score	Time (s)
Linear models			
Perceptron	0.60 (+/- 0.35)	0.37 (+/- 0.21)	15.18
Logistic Regression	0.72 (+/- 0.00)	0.28 (+/- 0.01)	68.48
Stochastic GD	0.68 (+/- 0.23)	0.30 (+/- 0.25)	15.49
SGD log reg as loss	0.60 (+/- 0.35)	0.39 (+/- 0.22)	15.78
Passive Aggressive Classifier	0.56 (+/- 0.38)	0.37 (+/- 0.21)	23.27
Non linear transformation			
SVC	0.67 (+/- 0.01)	0.02 (+/- 0.02)	28.05
NuSVC	0.67 (+/- 0.00)	0.01 (+/- 0.02)	29.39
Random Trees Embbedings	0.72 (+/- 0.00)	0.28 (+/- 0.01)	141.27
Extra Trees Classifier [2]	0.70 (+/- 0.03)	0.41 (+/- 0.06)	1.25

Estimators II

Table: Comparison between estimators

Method	**Accuracy**	**F1 Score**	**Time (s)**
Naive Bayes	0.68 (+/- 0.00)	0.41 (+/- 0.01)	41.89
RBF Sampler (Kernel approx)	0.64 (+/- 0.03)	0.16 (+/- 0.08)	1.54
Manifold Learning			
K Neighbors Classifier	0.61 (+/- 0.04)	0.41 (+/- 0.05)	1.58
Radius Neighbors Classifier	0.67 (+/- 0.00)	0.00 (+/- 0.00)	12.97
ANN			
Multilayer Perceptron	0.33 (+/- 0.00)	0.50 (+/- 0.00)	11.18

[2] This classifier is able to get perfect score using the whole data set

Tuning hyperparameters

The method Exhaustive Grid Search was used to optimize the hyperparameters of the best estimators (f1-score)

- Multilayer Perceptron: 0.5
- K Neighbors: 0.41
- Extra Trees Classifier: 0.51

Conclusions I

The current data is very complex, as result, most of the classifiers get bad performance, even worse that random. However, it is possible to configure some estimators in order to receive better score that pure chance.

This process is very demanding in terms of time and processing because needs to explore a wide range of hyperparameters and the execution becomes exponential in the number of hyperparameters.

Conclusions II

By the way, if we compare the score with respect the official leaderboard for the contest, this solution would result in the place 80 of 500.

References I

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux.
API design for machine learning software: experiences from the scikit-learn project.
In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.