Rafael Aguayo
A10075971

Assignment 2. Feature Selection and Evaluation for Adult Data Set

Abstract

For this project, I am testing the accuracy of several classifiers and seeing which features give better performance. I will be comparing different feature sets and see how they compare in terms of classification accuracy. I will be using Linear SVM with several different penalty parameters as well as logistic regression with the L1 and L2 penalty parameters.

2. The Dataset

The Adult data set comes from a 1994 Census database which was stored on the UC Irvine machine learning repository. It contains 14 features ranging from 8 categorical and 6 numerical (continuous) features. The features:
age: continuous,
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt (based on demographic data): continuous,
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous,
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous,
capital-loss: continuous,
hours-per-week: continuous,
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
The prediction task associated with this task is to determine whether a person makes over 50K a year or less.

This dataset contains over 48842 data samples, including those samples with unknowns. The data is split into 32,652 training samples and 16281 testing samples. After we remove the samples with missing data, we have 30,162 training samples and 15,060 testing samples. (For the purpose of this experiment, we are going to be working on the data excluding the samples with unknown values.) In the training set there is 22,654 data samples that make <=50K. There are 7,508 samples that make >50K. In the testing set there are 11360 samples that make <=50K.

There are 3,700 samples that make >50K. There is about a 25% (>50K) and 75% (<=50K) split in the training and testing data individually. This shows consistency between the two data sets.

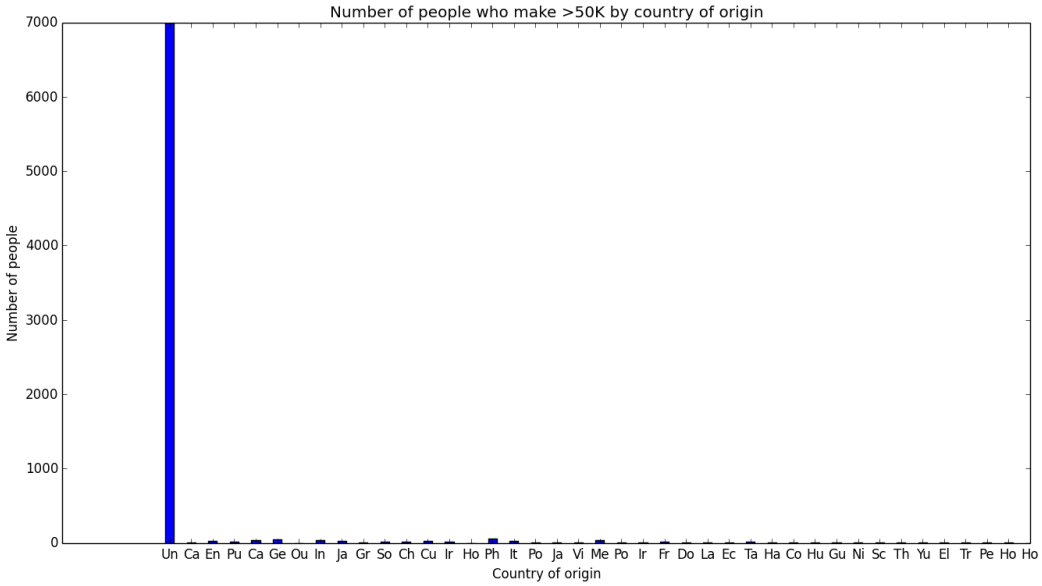

Figure 1.

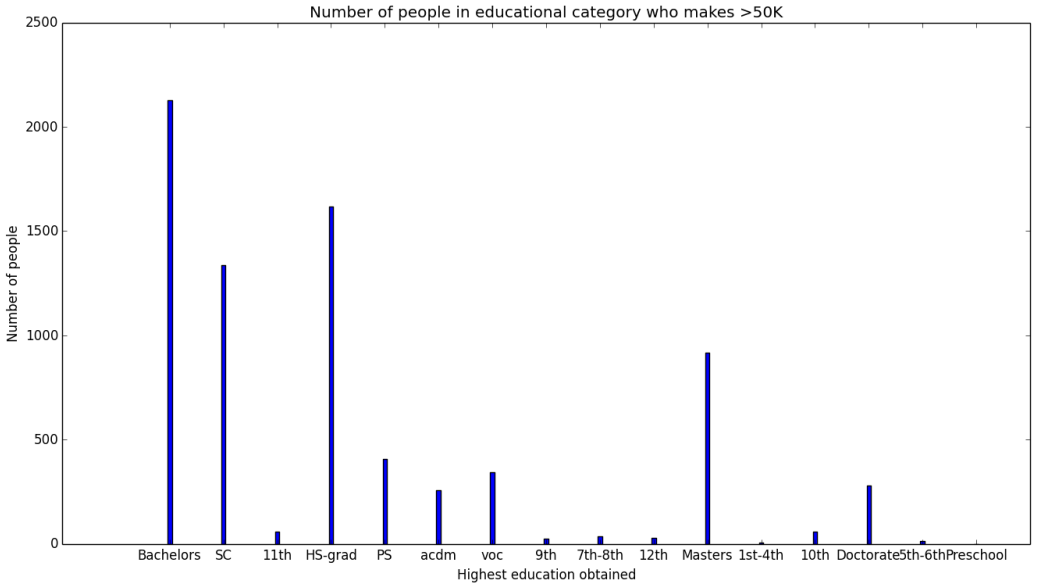

Figure 2.

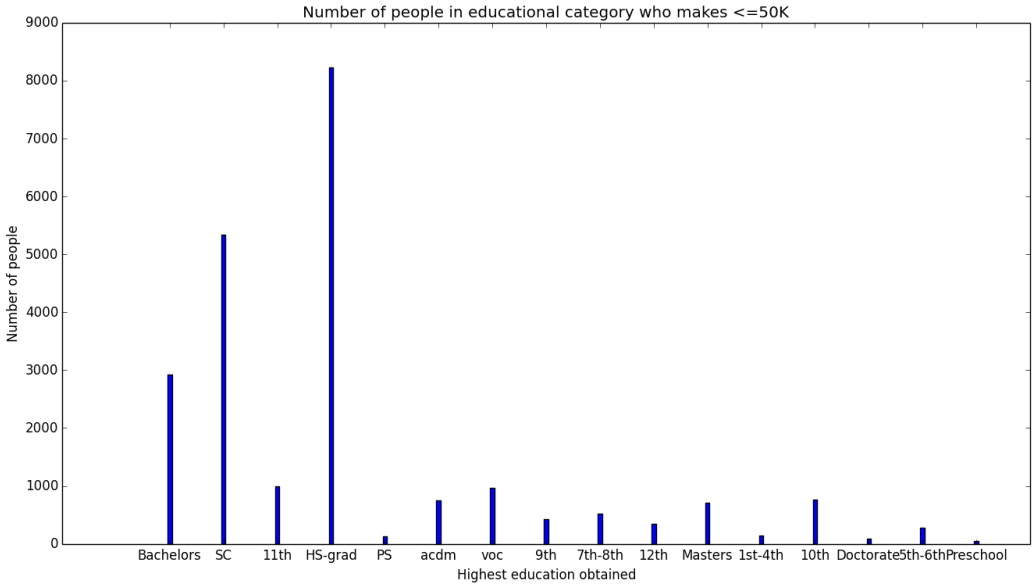

Figure 3.

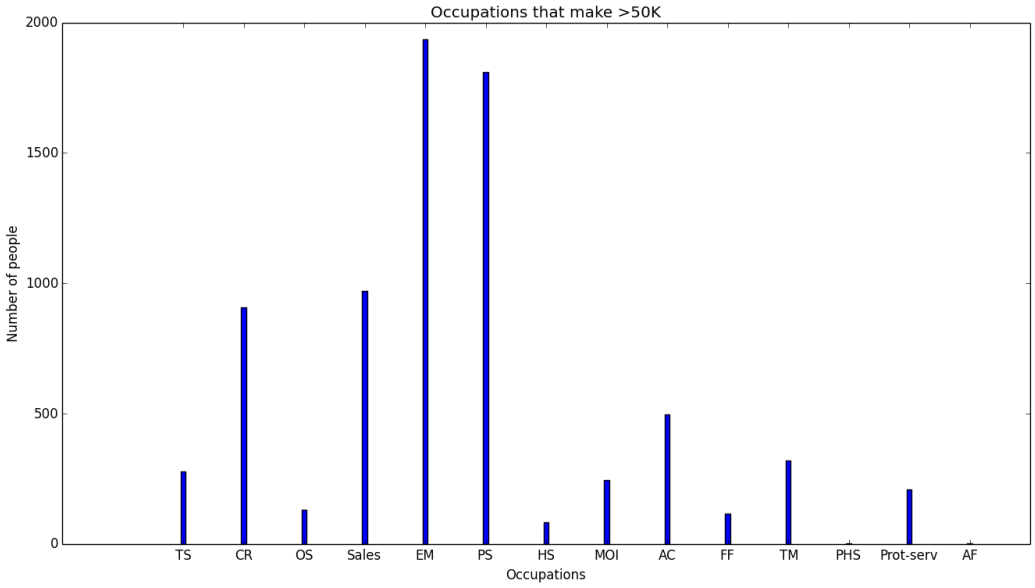

Figure 4.

(Statistics based on training and test data for numerical attributes)

	Mean	Median	Max	Min	Std. Dev.
Age	38.58	37	90	17	13.64
Final Weight	189778	178356	12280	1485000	105550
Edu-Num	10.08	10	16	1	2.57
Cap-Gain	1078	0	100000	0	7358
Cap-Loss	87.3	0	4356	0	403
Hours/week	40.44	40	99	1	12.35

We see in Figure 1 that most people's country of origin is the US, so I presume that we will be able to remove the country of origin from our feature set, thus reducing the dimensionality of our feature space, while still maintaining a good classifier. In Figure 2 and Figure 3 we see that the distribution of data makes sense. High school graduates are the dominant category for people who make <=50K. People with Bachelor degrees are the dominant category for people who make >50K. This distribution of data would help train the classifiers to know what educational backgrounds income range would typically be. Finally, when looking at Figure 4 we see that the majority of people who make >50K are in exec-managerial positions or have professional specialties. The other categorical variables also tell us something important about the data, so it seems that they will all be relevant in helping to make the best classifier.

I also suspect that features with high standard deviation would enable a stronger classifier such as final weight and cap-gain, and possibly hours/week, cap-loss, and age. Even after computing the statistics of mean, median, etc, I do not believe these will be helpful features for the classifier. The reason being is that there is more information where someone who is 18 and makes <=50K and someone who is 50 that makes >50K that could help the classifier.

3. Predictive Task

In the predictive task we are trying to evaluate the classification accuracy on whether a person makes >50K or <=50K for several classifiers with different sets of features given to us. For example, from the evaluation of the basic statistics of the dataset we see that the majority of people in the dataset have their country of origin from the United States. We would rerun the classifiers with the last feature of country of origin removed and see if there is a difference in the accuracy score. A major reason behind doing this is because we cannot use PCA to reduce the dimensionality of categorical variables. Although, some might argue that you would be able to use PCA. My justification for not using PCA is that this method tries to look at the covariance between features, and we do not have such statistics with categorical variables. We will do an experiment with the continuous (integer) features in the dataset where we do apply PCA and see the results from that.

In terms of the features, we shall leave the continuous variables as is. As for the categorical features, we will convert them into a one-hot encoding (binary encoding)[2] where we use values 0 or 1 to indicate the absence or presence of each category. There is also a target based encoding but this has a drawback of dependency to the distribution of the target. By using the binary encoding for each categorical variable we expand our feature space for each data point to be 105 dimensions long. The necessity of each dimension is debatable, but that is why we will experiment by removing certain sets of binary categorical variables.

$$Feature\ space \rightarrow [x_1, x_2, \dots, x_{105}]$$

One of the models that we are using is LinearSVM. Since we are doing a binary classification problem, LinearSVM would try to find a single split in the feature space that would classify as making more or less than 50K. For this reason, I found that we would not need to use other kernels such as radial (although it may still be interesting to see the results). Another reason for using this model is that it works well with a combination of categorical and continuous data. We will use penalty parameters C=1000,100,10, and 1 and see the different results we get from this.

We will also use logistic regression because this works well for binary problems. Logistic regression converts a real valued expression into a probability.

$$p_\theta(y_i|X_i)\epsilon\ [0,1]$$

By applying regression to our feature space, we come up with a model:

$$Linear\ Regression\ Model \rightarrow [x_1\theta_1, x_2\theta_2, \dots, x_{105}\theta_{105}]$$

In logistic regression we apply this model to a sigmoid function to attain:

$$\sigma(t) = \frac{1}{1+e^{-t}}\ , t = [x_1\theta_1, x_2\theta_2, \dots, x_{105}\theta_{105}]$$

$$\sigma([x_1\theta_1, x_2\theta_2, \dots, x_{105}\theta_{105}]) = \frac{1}{1+e^{[x_1\theta_1, x_2\theta_2, \dots, x_{105}\theta_{105}]}}.$$

In order to evaluate the models and feature selections, we will compare their accuracies compared to other methods used to classify the data. Other papers have used other methods such as nearest neighbor, Naïve Bayes Tree in which we will be comparing to.

4. Related Work

There has been a lot of related work on this dataset. Like I have mentioned before, this data set comes from the UC Irvine machine-learning repository. This repository holds donated datasets where other people can download them to perform their own analysis on them using their own desired techniques.

In terms of the adult dataset, most people have used this dataset to test different variations of classifiers to see if they can get a boost in accuracy or a boost in the time taken to solve these problems. One such paper also employed the binary encoding for the adult data set but they used a training set of 4000 samples and a testing set of 35,000 samples [3]. There are not a lot of papers on what I am trying to do with this dataset, which is feature selection and evaluation. I would say the most similar work involves predicting a borrower's chance of defaulting on credit loans. In this sense, attributes are important such as whether a borrower owns a car or a house to determine their credit reliability. In another paper using the German Credit Data set, they deem certain features "selectable" based on statistics computed from the data [4]. These features could differ depending on the model used, in which this case they were using neural networks. Following algorithms were run with the following error rates from several different papers, all after removal of unknowns and using the original train/test split [1].

	Algorithm	Error
1	C4.5	15.54
2	C4.5-auto	14.46

3	C4.5 rules	14.94
4	Voted ID3 (0.6)	15.64
5	Voted ID3 (0.8)	16.47
6	T2	16.84
7	1R	19.54
8	NBTree	14.10
9	CN2	16.00
10	HOODG	14.82
11	FSS Naive Bayes	14.05
12	IDTM (Decision table)	14.46
13	Naive-Bayes	16.12
14	Nearest-neighbor (1)	21.42
15	Nearest-neighbor (3)	20.35
16	OC1	15.04

Figure 5.

5. Results and Conclusion

In terms of feature selection, my results were somewhat expected. The accuracy scores seemed to drop for the logistic regression as we decreased the number of features used overall. The accuracy scores were somewhat scattered for the different features used for Linear SVC. For the Linear SVC, the accuracy results were generally within 5% of each other. As for the logistic regressor, the accuracy results were generally within 10% of each other. Considering this is a binary classification problem, we could see why we would get such close results. Since there are only two classes to predict, even a guessing method might come up with better than mediocre results. In this data set, since there is about a 75/25 data split in the testing set for predicting whether someone makes over 50K or not, one could simply mark all instances as <=50K and they could achieve a 75% accuracy. Although this is true, when making a prediction on this type of data there is no penalty for a false negative. For example, on the German credit data set, there is a big penalty for a false negative where you give a person a loan where they would most likely default on the loan payment. Approaching a model where we would need to reduce the number of false negatives would be a more interesting study, but that is reserved for future work.

The lowest error for the logistic regressor model was 0.1516%, with the highest error being 0.2632%. The lowest error for SVC was 0.2077%, with the highest error being 0.2295%. The logistic regressor model performed really well in comparison to the error rates that others received in Figure 5. The linear SVC also performed decently in comparison to the error rates in Figure 5. These results show that these two classifiers are really viable options in terms of making binary classifications. Also, feature selection is a very important aspect in terms of getting a lower error prediction rate. Even though our feature space was not extremely large (105 dimensions), it shows that having more dimensions allows us to have a better classifier and make better predictions. Although, it could be the case that having too many dimensions will cause us to attain a worse accuracy score.

Another interesting discovery was made by looking only at the integer attributes within the data and applying PCA to the feature space. We see that most of the variance is explained after the first dimension in Figure 6. The difference between having 1 dimension, and having 4 dimensions does not change the accuracy score by much in this experiment.

(Average over 3 trials)

Linear SVC	C=1000	C=100	C=10	C=1
All Features	0.7787	0.7913	0.7807	0.7756
Excl. Native	0.7737	0.7882	0.7705	0.7871
Excl. native, marital status	0.7727	0.7810	0.7923	0.7733
Excl. education	0.7742	0.7806	0.7841	0.7767
Integer Attributes Only	0.7766	0.7815	0.7788	0.7761
Only age attribute	0.7543	0.7531	0.75418	0.7543

Logistic Regression	L1 penalty	L2 penalty
All Features	0.8483	0.7924
Excl. native	0.8479	0.7929
Excl. native, marital status	0.8476	0.7922
Excl. education	0.8474	0.7931
Integer attributes only	0.8068	0.7924
Only age attribute	0.7368	0.7368

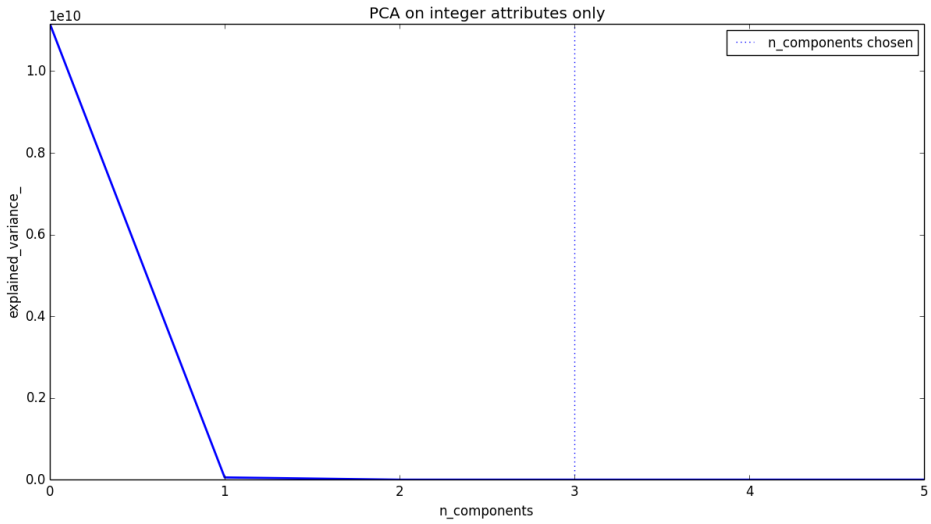

Figure 6.

References

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[2] http://www.saedsayad.com/encoding.htm
[3] R. Caruana and A. Niculescu-Mizil. An empirical evaluation of supervised learning for ROC area. First Workshop of ROC Analysis in AI (ROCAI '04), 2004.
[4] P. O'Dea, J. Griffith, and C. O'Riordan. Combining feature selection and neural networks for solving classification problems. Irish Conference on Artifical Intelligence & Cognitive Science, 2001.