Tell me your demographic, I'll guess your level of education

Using Census Data To Predict Levels Of Education

†Richard Muniu
‡Department of Computer Science, Swarthmore College

Introduction

Demographic data is often used in Machine Learning to predict people's levels of income. The target label represents their level of income, which is discretized into two classes: above 50,000 per year, or below 50,000 per year.

Instead of using different machine learning techniques to perform and optimize this classical classification problem, we chose to look into how the very same demographic data, including one's income level, can be used to predict a preson's level of education.

Data

Our dataset was the 1994 Census Data, obtained from the UCl Machine Learning Repository. It contained about 48,842 data points, with a spread of continuous and discrete features. This included 32,561 training examples and 16,281 testing ones. We randomly shuffled the data and split it into train and test sets, using the labels {*HS, Some HS, HS grad, Some College, College Grad, Masters, Doctorate*} for multi-class classification and {*No College vs College*} for binary classification.

The data attributes included *Age, Work Class, Income, Marital Status, Occupation, Relationship, Sex, Hours Per Week, Native Country*, and *Education*, which we used as our label.

Methods

We used three different classifiers in this task:

1. SUPPORT VECTOR MACHINES
   - To pre-process the data for this classifier, we used a one-hot encoding for discrete features, leaving continuous features as they were.
   - Unknown values were replaced with a '0' in our encoding scheme.
   - For multi-class classification, we used a one-vs-rest classifier.
   - This method involved hyper-parameter tuning for our LinearSVC.
2. DECISION TREES
   - To pre-process our data, we converted our continuous features to discrete ones, and binarized discrete features. The continuous features were split using arbitrary bin sizes.
   - We used an off-the-shelf implementation of Decision Trees from `scikit-learn`.
3. NAIVE BAYES
   - We created an "unkonwn" value for each unknown feature in the original dataset.
   - Next, we discretized all the continuous features.
   - We then used an off-the-shelf implementation of Decision Trees from `scikit-learn`.

We also used a Most Frequent Class (MFC) classifier as a performance baseline for the aforementioned classifiers. The MFC classifier uses the most frequent label in the data as its classification output (prediction label).

Results

Our MFC classifier had an accuracy score of 36.5% over the 7 possible outputs.

Accuracy Score: 36.5%

Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	0	0	124	0	0	0	0
Some HS	0	0	306	0	0	0	0
HS Grad	0	0	1096	0	0	0	0
Some College	0	0	696	0	0	0	0
College Grad	0	0	551	0	0	0	0
Masters	0	0	192	0	0	0	0
Doctorate	0	0	35	0	0	0	0

Fig. 1: Most Frequent Classifier Accuracy

Our SVM classifier achieved the highest accuracy score, at 44.0%. This was closely followed by the Decision Tree classifier at 41.4% and lastly Naive Bayes, which achieved 35.6% accuracy.

Accuracy Score: 44.0%

Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	54	1	59	8	1	1	0
Some HS	16	2	190	91	5	1	1
HS Grad	25	5	657	335	61	11	2
Some College	7	3	258	339	78	11	0
College Grad	6	0	77	199	223	42	4
Masters	1	0	16	24	107	42	2
Doctorate	0	0	2	2	22	7	2

Fig. 2: SVM accuracy

Accuracy Score: 41.4%

Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	36	22	47	11	8	0	0
Some HS	19	60	144	74	7	2	0
HS Grad	37	83	647	201	117	9	2
Some College	19	46	292	208	111	15	5
College Grad	6	14	132	99	252	42	6
Masters	1	1	23	28	100	32	7
Doctorate	0	0	3	2	16	7	7

Fig. 3: Decision Tree Accuracy

Accuracy Score: 35.4%

Level of Education	No HS	Some HS	HS Grad	Some College	College Grad	Masters	Doctorate
No HS	62	0	50	12	0	0	0
Some HS	22	3	235	45	1	0	0
HS Grad	72	6	848	159	3	8	0
Some College	35	3	515	136	3	3	1
College Grad	36	0	408	86	6	14	1
Masters	15	0	111	58	0	7	1
Doctorate	4	0	21	7	1	2	0

Fig. 4: Naive Bayes Accuracy

Comparison

Our overall comparison of all the classifiers for both binary and multi-class classification is as follows:

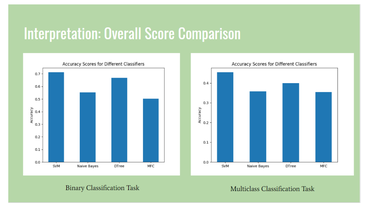

Fig. 5: Overall Score Comparison

Feature Analysis

We used our LinearSVC to compare the relative importance of features, by analyzing the *nxn* matrix of coefficients. The five most important features determining one's level are shown below:

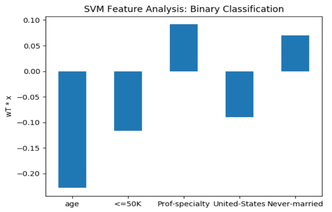

Fig. 6: Overall Score Comparison

Conclusion

Our SVM model produced best results overall, and income, occupation, and age were seen to be the most important features (according to our Tree and SVM analysis).

Future Considerations

We could account for colinear independent variables and view problem as regression rather than classification.

Acknowledgements

I would like to thank Sara Matheson for her constant encouragement and support throughout the project.I'd also like to thank my teammates Kenny Gwon and Raymond Liu for their collaborative efforts. Special thanks to the UCl Repository for making the dataset available..