Classifying Income from 1994 Census Data

Tracy Nham
A0994191
tnham@usd.edu

I. INTRODUCTION

The adult dataset, hosted by The Machine Learning Group at UCI, contains census information from 1994. With this data, we are tasked of predicting whether a person makes more than $50K/year. In the following sections, I will analyze the properties of the dataset and use classification algorithms, including logistic regression, Naïve Bayes, and decision trees, to make such predictions.

II. THE ADULT DATASET

The adult dataset is a fairly large set, consisting of 48,842 instances. There are 14 attributes prescribed to each person: {income ('>50K' or '<=50K'), age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country}.

Of those fields, I am interested exploring a subset of the attributes. Their basic statistics are displayed in several tables below.

Attribute	**Values**
Education Level	High School (32%), Some college (22%), Bachelors (16%), Masters (5%), Vocational (4%), 11th (4%), Assoc Academic (3%), 10th (3%), 7-8th (2%), Professional School (2%), 9th (2%), 12th (2%), Doctorate (1%), 5-6th (1%), 1-4th (1%), Preschool (1%)
Relationship	Husband (41%), Not-in-family (26%), Own child (16%), Unmarried (11%), Wife (4%), Other relative (2%)
Race	White (85%), Black (10%), Asian/Pacific Islander (3%), American Indian/Eskimo (1%), Other (1%)
Marital Status	Married-civ-spouse (46%), Never-married (33%), Divorced (14%), Separated (3%), Widowed (2%), Married-AF-spouse (1%), Married-spouse-absent (1%)

Attribute	**Values**
Salary [Label]	<=$50K (76%), >$50K (24%)
Gender	Male (67%), Female (33%)

Attribute	**Mean**	**Median**	**Std Dev**
Age	38.58	37	13.64
Hours worked per week	40.44	40	12.35
Education Number	10.08	10	2.57
Capital Gain	1078	0	7385
Capital Loss	87.3	0	403
Survey Weight	189778	178356	105550

Furthermore, I'm interested in how different attributes correlate with another. Of the individuals who make more than 50K, does their race affect the hours they work?

Based on additional analysis, these two attributes do correlate. Regardless of race, the mean of the hours worked is roughly 44 hrs/wk.

In addition, I wanted to study how marriage may possible affect the label. For each label, I found the percentage of each category from the overall data.

Marital Status Distribution Across Labels

<= 50K	Percentage	> 50K	Percentage
Separated	93	Separated	7
Widowed	91	Widowed	9
Divorced	90	Divorced	10
Married-spouse absent	90	Married-spouse absent	10
Never married	95	Never married	5
Married-AF-spouse	56	Married-AF-spouse	44
Married-civ-spouse	55	Married-civ-spouse	45

From the table, we observe those who make more than $50,000 are most likely to be married than single.

Next, I want to do the same analysis on the level of education.

Highest Level of Education Completed Distribution

<= 50K	Percentage	> 50K	Percentage
Masters	45	Masters	55
Prof-school	25	Prof-school	75
Assoc-voc	74	Assoc-voc	26
Assoc-acdm	74	Assoc-acdm	26
HS-grad	84	HS-grad	16
Bachelors	58	Bachelors	42
Some-college	80	Some-college	20
Doctorate	27	Doctorate	73
Preschool-12th	94	Preschool-12th	6

As expected, those with higher education are more likely to be labeled positive for making over $50,000.

Finally, I want to analyze the relationship between education level and age.

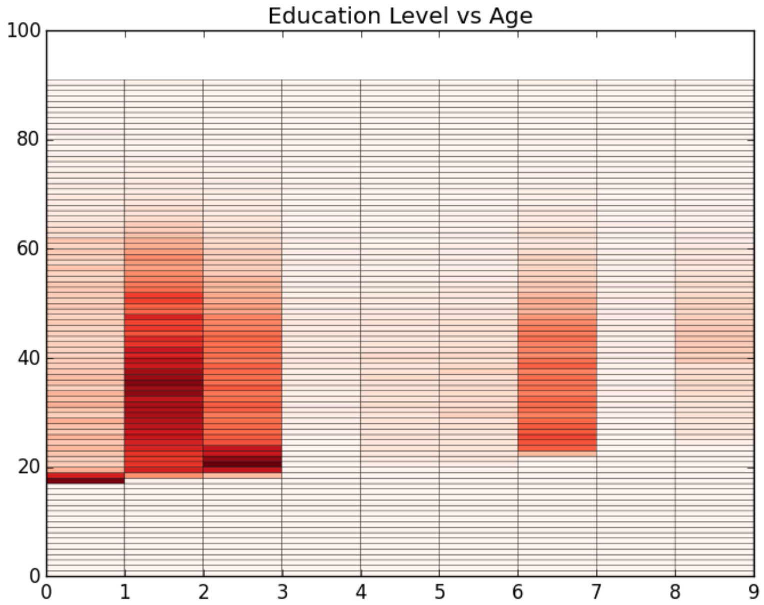

Figure 1. HS not completed, HS grad, Some College, Professional School, Associate's Academic, Associate's Vocational, Bachelor's, Doctorate, Masters

From the figure, we observe that for many of the ‘highest education achieved’ categories, there is a relatively high concentration of younger people in that category. This makes sense for the less than High School completed category since the data contains those young enough to not even be able to complete it yet (17 years old).

The graph shows that a large part of the data falls between the first three categories. The proportion between the data points labeled <= 50K and >50K is a ratio of about 3:1, we can leverage education and age when creating our feature vectors.

III. Predictive Task

Since the dataset was processed and compile to predict the income of an individual, we will also be making these predictions in this study.

IV. METHODS:

Preprocessing

Of those 48,842 samples, 3,620 of them contain missing values. So for consistency, they are removed from the entire set, leaving 45,222 instances. Of these remaining points, about 25% have the label '>50K' and 75% have the label '<=50K'.

Since the data files provided by UCI are already randomly split into a train and test set (2/3 and 1/3 of the original set respectively), all I needed to do was to create a validation set. Using the train_test_split function from the sklearn library, I created a validation set from 1/3 of the points in train. Thus the final sizes of each set are as follows:

Train = 20,207 Validation = 9,954 Test = 15,058

Next, I did some processing on the data:

- I combined the categories where that were less than HS grad into one category.

I'm not sure whether binning ages into ranges will improve our models, but if it does, I will bin them in the following way:

- [>= 18, 18-25, 25-35, 45-55, 55-65, >=65]

Feature Selection

Age:

Age would help the classifier in that those who are younger are less likely to earn more than $50,000 a year. However, I'm not sure if this field is necessary since age is somewhat correlated with education.

Education:

The higher an individual's education the more likely it for them to be labeled positive for making over $50,000. For simplification, the labels under education 'Preschool' to '12th' will be categorized together as we can infer that most people with this level of education is unlikely to make the cutoff amount.

Marital status:

From the analysis in the previous section, marriage is one of the bigger evidence to seeing if an individual has more than $50k income. Each marital status will be assigned a number from 0-6.

Sex:

Biological sex is a huge indicator of income since men tend to earn more money than women.

Evaluation:

Since we are dealing with a binary classification task, to evaluate the performance of the models, I will use the error rate formula:

$$error\ rate(s) = \frac{\#\ of\ mistakes}{size\ of\ the\ set(s)}$$

Models:

I decided on using Naïve Bayes as the baseline model for the predictions because it's known to good performance all around when classifying and because there are no regularization parameters.

As I'm not sure whether or not to include age in our prediction, I will use test the performance of various, but very similar, feature representations.

Naïve Bayes (baseline)

Feature Representations	**Test Error**
4 features where age is in numerical rep	0.216363394873
4 features where age is binned	0.219883118608
3 features, discarded age	0.19398326471

Though the test errors for each representation were all very close, I believe that because age and education level are correlated and Naïve Bayes has an overcounting problem that discarding age altogether had the lowest test error overall.

Logistic Regression:

The next model we will study is Logistic Regression, which solves the overcounting problem in Naïve Bayes. I run logistic regression a total of nine times, experimenting with the feature representations and regularization parameter.

λ = 1

Feature Representations	**Test Error**
4 features where age is a numerical rep	0.194979412937
4 features where age is binned	0.194979412937
3 features, discarded age	0.194979412937

λ = 100

Feature Representations	**Test Error**
4 features where age is a numerical rep	0.195311462346
4 features where age is binned	0.192588657192
3 features, discarded age	0.194979412937

λ = 1000

Feature Representations	**Test Error**
4 features where age is a numerical rep	0.195311462346
4 features where age is binned	0.192588657192
3 features, discarded age	0.194979412937

From the table, it seems that with more regularization, the test errors converge to a particular number. Too little regularization makes the feature almost meaningless when we look at the table where λ = 1. At λ = 100, it seems that we are already overfitting. Similarly for when λ = 1000. But overall, Logistic Regression favors the feature representation that includes age, but age is binned.

Decision Tree

Though we didn't cover the Decision Trees in class, as I was reading the related literature to this dataset, this algorithm was mentioned a number of times. Thus, I decided to examine the performance of this model, again with the three possible feature representations I'm considering.

Feature Representations	**Test Error**
4 features where age is a numerical rep	0.229911010758
4 features where age is binned	0.190994820029
3 features, discarded age	0.190596360738

The decision tree worked better when age was categorized instead of numerical. This is possibly due to the tree overfitting to the training data when it creates so many splitting points when age is numerical. It made a great improvement when age was binned, just about 3%. It did slightly better when age was discarded.

Interesting Finding:

Lastly, I found that using a non-machine learning model outperforms all machine learning models I've tried. The process is, for each sample in the data check:

1. Is the individual White?
2. Is the individual a male?
3. Does individual have at least a Bachelor's degree?

If all three conditions are satisfied, then we predict that the individual's income is greater than $50K.

We get the following results:

Training Error	Validation Error	Testing Error
0.0643341416341	0.0591721920836	0.0646832248639

However, this performance is a result of knowing the properties of the dataset. We know from the exploratory analysis, the data is based toward White males with college degrees having more than $50,000 income.

V. CONCLUSIONS

From the experiments with feature representation, in many cases it's better to represent a field using categories than numerical values, especially when numerical encoding overcomplicates the data. For example, a 21 and 24 year-old would not be expected to make more than $50,000 a year in 1994. According to dollarsigns.com, that would be the equivalent about $80,000 today. To reduce noisiness, it is preferable to group them up if they tend to give the same inference. This can also help speed up the model's training as well. For SVMs for example, reducing the feature space will greatly improve its training and classification time. And as we also saw with different models, it can be better to discard an attribute that positively correlates with another existing attribute, like age and level of education.

It would have been also interesting if I was able to look at the performance of SVMs for this assignment. Unlike Naïve Bayes and Decision Trees, we can apply more penalty to mistakes, which would have been helpful for this particularly dataset because it's imbalanced.

VI. LITERATURE AND OTHER RELATED WORK

The adult dataset was compiled by Barry Becker, who extracted the data from the US 1994 Census database. This dataset was given to Ron Kohavi who used it to study the effectiveness of a new machine learning algorithm he's proposed called the NBTree. According to Kohavi, the NBTree algorthm leverages the "surprising [accuracy]" of Naïve-Bayes and the scalability of Decision Trees. After the completion of Kohavi's paper in 1996, the dataset was donated and now hosted by the Machine Learning Group at UC Irvine.

Since then, the dataset has been referenced in roughly 50 academic papers. In many of these papers, researchers studied the performance of augmenting existing classification machine learning algorithms, such as SVMs and K-NNs by boosting, partitioning, squashing, etc. Due to the dataset's numerous citations, it is clearly popular and well-known amongst the Data Mining and Machine Learning community.

A very similar dataset to the adult dataset is the Census-Income (KDD) Data Set, which was also donated by Kohavi and Terran Lane in 2000. It is a very big set, containing data from both the 1994 and 1995 U.S. Census Bureau's current population survey. It has almost 300,000 sample where each sample contains 40 attributes. Like with the adult dataset, the KDD provides another income classification problem. Currently, the KDD dataset was cited by four papers, each which were specifically looking for very large data. All these papers discuss was to quickly and effectively process large amounts of data, which includes, for example, breaking the data into multiple intervals.

In final words, due to the Kohavi's paper, I learned about Decision Trees, which turned out to be a very effective model when I was comparing it to my own model.