CSCI3320 project report – What makes people happy

1155092191 CHEN Ran
1155092208 Wang Zhi Cong

2 Raw Data Preprocessing

2.2 Data Discretization and Analysis

Some statistic analysis of data:

The mean YOB of the train data is 1979.

The gender ration is 59:41, which means there are six males in every ten interviewees.

Most of people's income is between \$50,000~\$74,999, which is the third level of our classification (the higher level means higher income). It sounds reasonable.

The number of different parties: Democrat 721, Republican 947, Independent 909, Libertarian 332, Other 205. The Republican and Independent almost accounted for half of the interviewees.

Besides some obvious features, there are many features like Q12XXXX, It seems like either-or questions. We are not sure whether they are influential to our prediction result. So, we first try to delete all the Q12XXXX:

Train Accuracy (scikit method): 57.6454668471

Train Accuracy (own method): 50.2092050209205

It seems that only with a very few features on person information cannot do correct prediction. We cannot get the conclusion on a person whether happy or not only depends on his/her's income, YOB, party…

Next, we choose all the Q12XXXX:

Train Accuracy (own method): 66.52719665271967

Train Accuracy (scikit method): 70.9066305819

Compare to the predict with all features, the accuracy is quite close. Which means, even we do not know interviewees' personal data, we still can know whether they are happy or not through some questions!

And of course, the accuracy will increase if we add more and more features into our model.

2.3 Missing Data Filling

Some confusion before we start do classification tree:

1. When there are missing data on the tree's root. For example, more than 10% gender is unknown on training data. What we need to do?
   - Delete them.
   - Filling them with mean/mode/median.

   If we delete them, we will lose some important information on training set. That will influence our model and result.

   If we retain them, the 'fuzzy processing' on missing data will do unknown effect on the result. Because if the root got wrong, the entropy will get higher and higher on classification.
2. Features are not independent to each other.

   The connection among the Party/Income/Edu/Household are 'delicate', we can hardly do mathematic modeling on such information and do predict on missing ones.

Reasons for the way we group the users:

We have spent a lot of time on trying different classification tree:

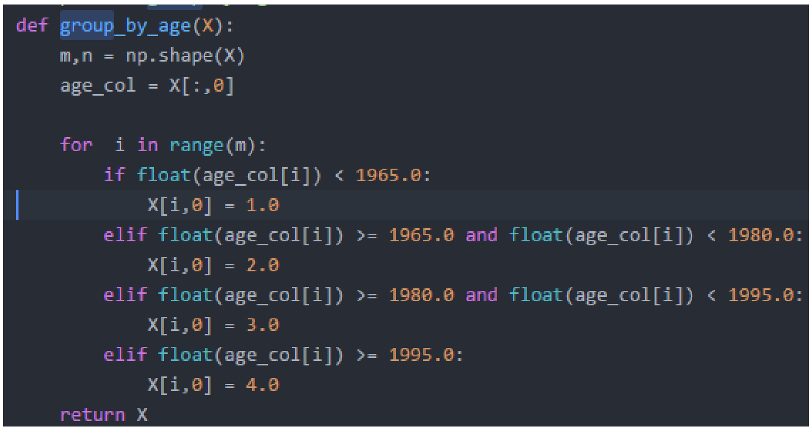
1. Gender → Income: 2 * 6 = 12 classes, do missing filling on NaN data at each level.
2. Gender → Party: 2 * 5 =10 classes, do missing filling on NaN data at each level.
3. Gender → Party → Income: 2 * 5 *6 = 30 classes, do missing filling on NaN data at each level.

But the result with our LR model shows that the tree seems has no effect on the prediction. The promotion is lower than 1%.

After we observe the biplot, we find out that the age and Income are orthogonal to each other. It implies there is less connection between YOB and Income.

YOB → Income model:

Detail: The YOB data are so disperse, which is not good to do classification. And people at different age has different understanding on happy. So we do the work to group them:



```
def group_by_age(X):
    m,n = np.shape(X)
    age_col = X[:,0]

    for  i in range(m):
        if float(age_col[i]) < 1965.0:
            X[i,0] = 1.0
        elif float(age_col[i]) >= 1965.0 and float(age_col[i]) < 1980.0:
            X[i,0] = 2.0
        elif float(age_col[i]) >= 1980.0 and float(age_col[i]) < 1995.0:
            X[i,0] = 3.0
        elif float(age_col[i]) >= 1995.0:
            X[i,0] = 4.0
    return X
```

Then we classify the data into four classes. We fill the missing gender with ratio 6:4 (statistic analysis).

On the second level, we fill the missing income with mean value and do the Income classification.

Now, the tree is complete. We fill the rest of missing data with mean/mode/median.

The result with our LR model shows that the tree seems has a little effect on the prediction. The promotions are about 3%. We are not sure whether this result is good or not. I think it should be much higher in ‘actual combat’.

3 Classification

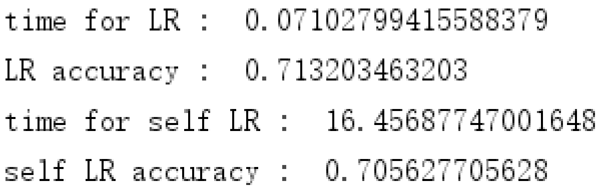
3.1 Train Classifiers in Scikit-Learn

3.1.1 Logistic Regression

Comparison:

a) Running time: the model in sklearn is much faster than our model. It takes about more than 1 minute to run 100 iterations for our model to reach the same accuracy as the model in sklearn.

b) Accuracy: the accuracies of both models are almost the same with tiny difference.



```
time for LR :  0.07102799415588379
LR accuracy :  0.713203463203
time for self LR :  16.45687747001648
self LR accuracy :  0.705627705628
```

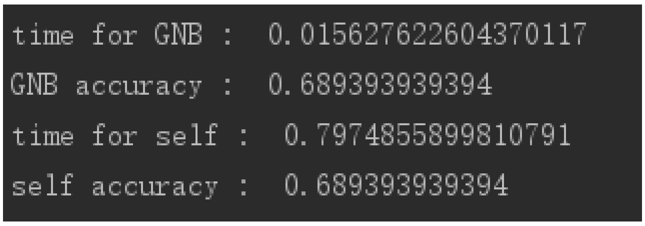
3.1.2 Naïve Bayes

We choose GaussianNB as our classifier. We suppose the distribution of each of the feature follows the Gaussian distribution. We think this may be better to assume the distribution as Gaussian than the other two in this training set. And the cross-validation test result also confirms this result.

Comparison:

a) Running time: The model in sklearn is still faster than our model.

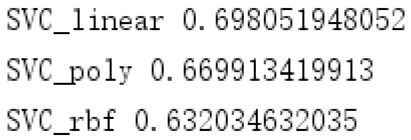
b) Accuracy: The two models are same.



```
time for GNB :  0.015627622604370117
GNB accuracy :  0.689393939394
time for self :  0.7974855899810791
self accuracy :  0.689393939394
```

3.1.3 SVM

We choose kernel function ‘linear’. There are only two classes, happy and unhappy. Also, the question part of the data has great percentage of the whole data. So maybe the linear kernel will have a better performance. The cross-validation result shows that linear performs best.



```
SVC_linear 0.698051948052
SVC_poly 0.669913419913
SVC_rbf 0.632034632035
```

3.3 Write A Report

Q: What are the characteristics of each of the four classifiers?

LR: Easy to realize. It can be used to classify and also predict possibility. It has good performance on linearly separable data.

NB: In this classifier, the features should be assumed to be independent. Also, the distribution of each feature is assumed to follow some certain distribution, like Gaussian or Bernoulli. Naïve Bayes converges quickly, a small set of training set can do well. But it cannot deal with the dependency between features.

SVM: This classifier is mainly for 2-class condition. It is better for small scale of data and will cost a lot of time for large dataset. The effectiveness of SVM depends on the selection of kernel, the kernel's parameters, and soft margin parameter C. SVM is supervised learning model, so it requires full labeling of input data. And SVM cannot calibrate class membership probabilities.

RF: Easy to explain and understanding. We can deal with dependency between features easily. It is nonparametric method, we do not to worry about the outlier points problems. The combination of decision trees can generalize the errors and avoid the overfitting when there is not too much noise.

Q: Different classification models can be used in different scenarios. How do you choose classification models for different classification problems? Please provide some examples.

We choose model according to the characters of both the problems and the models.

For small dataset and two-class classification, like the happiness prediction we have done, we will choose SVM because SVM may be better in this situation.

For datasets with independent features, like to predict the ball’s type according to size, color and other independent features., maybe it is better to use NB classifier.

LR is always used to predict the probability, so problems like the probability that the customer will buy this product can be solved by LR.

RF is flexible, so it may perform well for several situations. But it is essential to set the proper parameters for RF algorithm and this will cost much time.

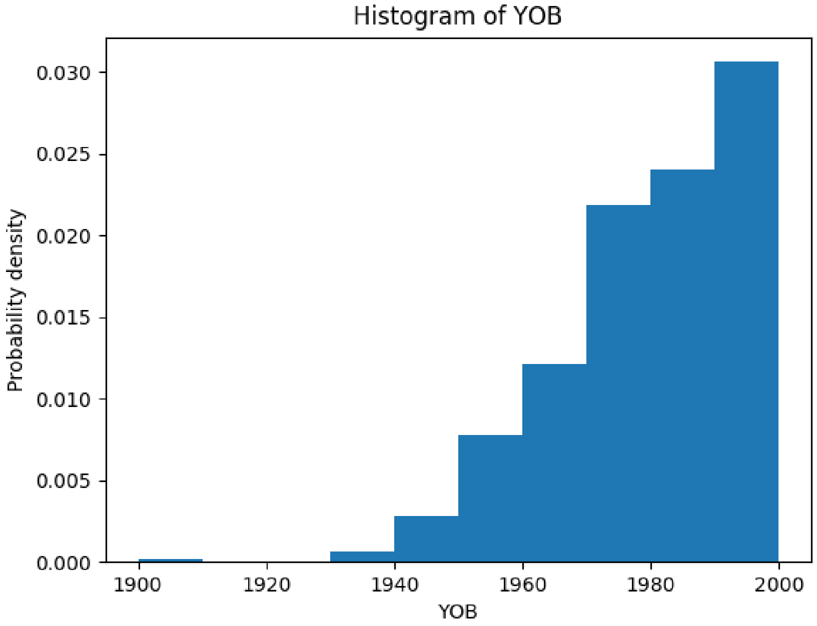
Q: How do the cross-validation techniques help in avoiding overfitting?

Cross-validation does not completely overcome the overfitting problem in model selection, it just reduces it. By using cross-validation, we can see the performance by splitting the dataset into trainset and testset. In this project, it mainly helps us to choose the better model from several similar models, like Gaussian, Multinomial and Bernoulli in Naïve Bayes. Maybe the prediction of the trainset has better accuracy but through the cross-validation, we can find the overfit model and then tune or select another one.

4 Visualization

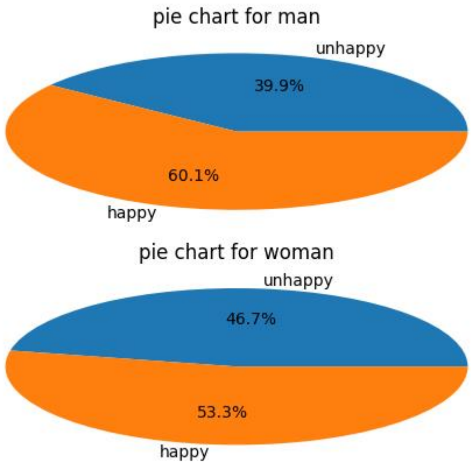
4.1 Basic Visualization Methods

4.1.1 Histogram of YOB



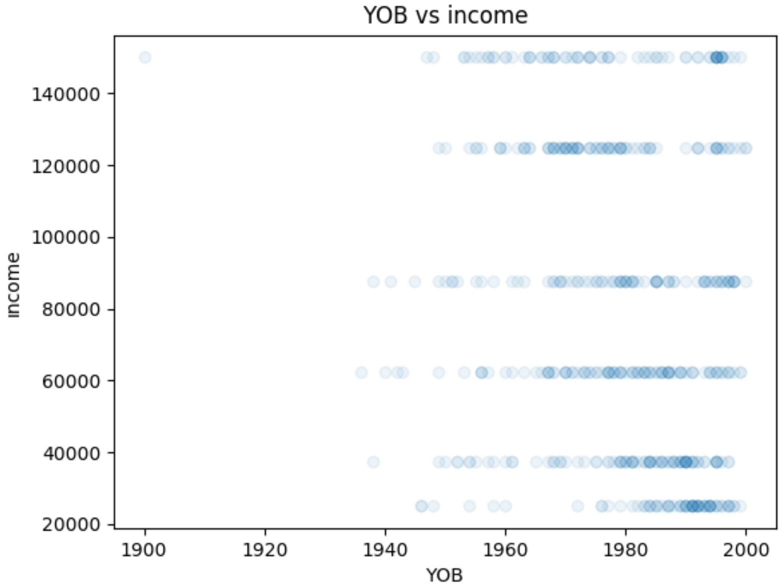
In this image, we find that the YOB distribution increases as the YOB increasing, which means most of the interviewees are younger.

4.1.2 Pie chart for the fraction of happy men/women



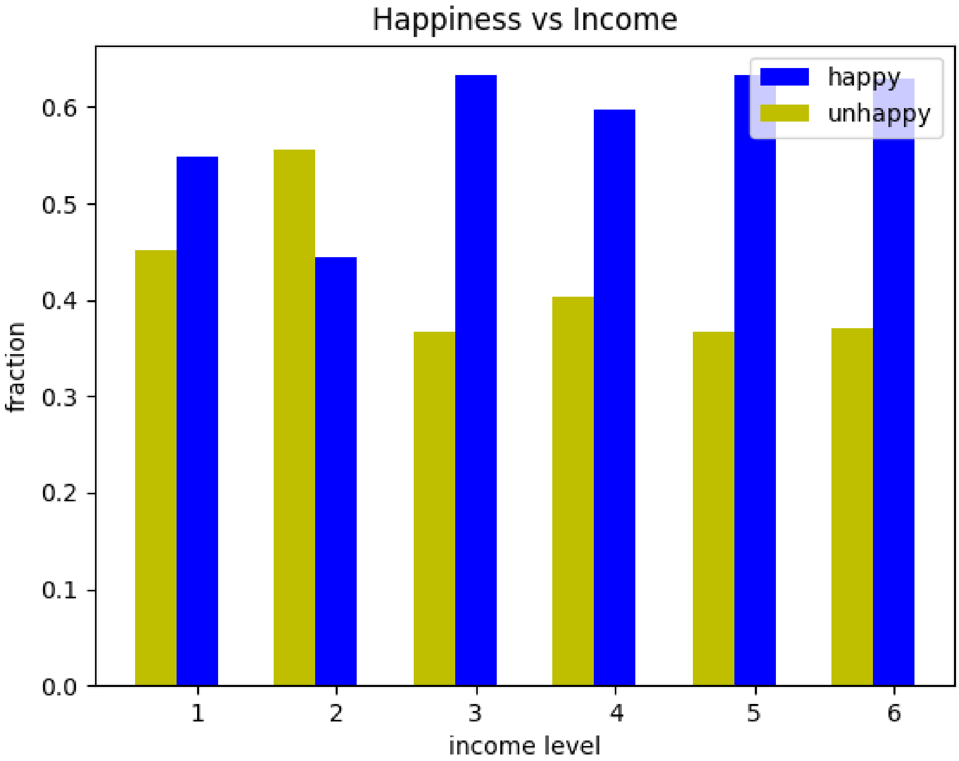
In this image, we find that the percentage of the happy man is far higher than the percentage of happy women, which means men tend to be happier than women.

4.1.3 Scatter plot of YOB and income



In this image, the darker area shows that more people with certain YOB has such income, which show the income situations of people with different YOB.

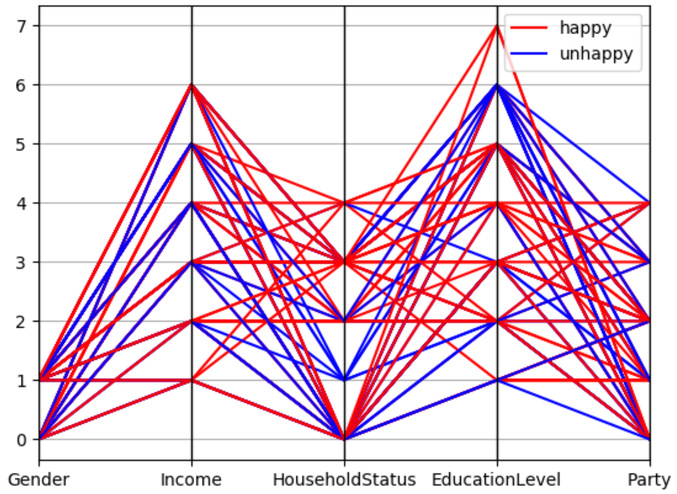
4.1.4 Bar chart of income and happiness



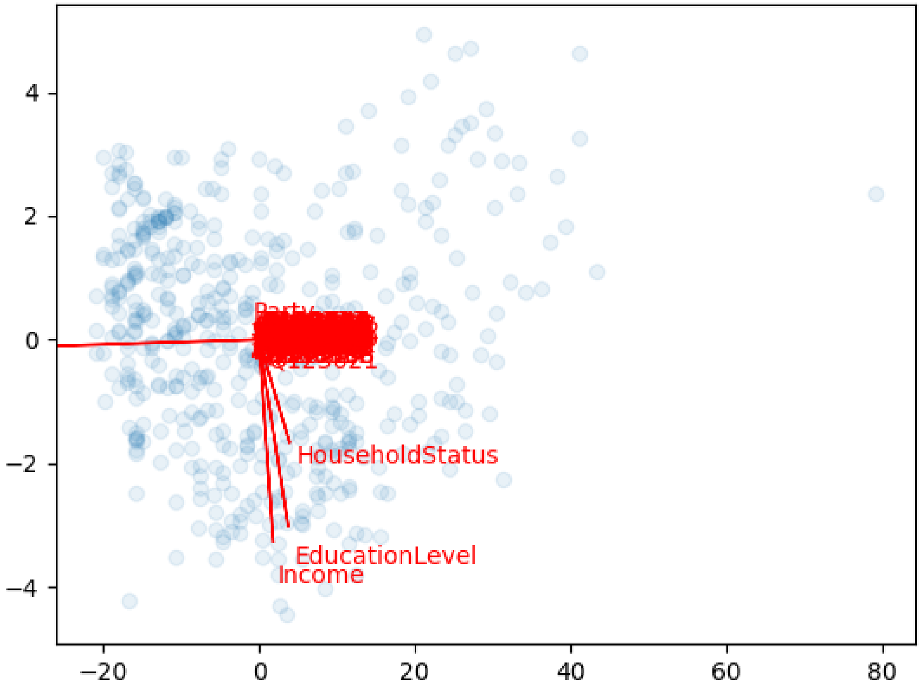
This image shows that except people with highest income and lowest income, more people are happy.

4.2 Visualizing High-dimensional Data

4.2.1 Parallel Coordinates Plot



4.2.2 PCA and biplot



The vector pointed to the left represent YOB.

Q1 What's the physical meaning the vector corresponded to each variable? Explain it in one sentence.

The vectors show the information of features and points are the samples on the twodimension space and the arrows from the origin are to reinforce the idea that these features can be projected to approximate the original data.

Q2 What are the factors closely related to happiness according to this biplot? Write down your answer and use one more sentence to explain why.

I think they are YOB, HouseholdStatue, EducationLevel and Income because the vectors of those factors are more significant on the image which may be more closely related to happiness.

4.3 Visualizing Classification Result

4.3.1 Visualize SVM with kernel rbf