T1. Find the MLE of the rate of return, α, given the observed price at the end of each day y2, y1, y0. In other words, compute for the value of α that maximizes p(y2, y1, y0|α)

Chart, scatter chart

Description automatically generated

T2. Plot the posteriors values of the two classes on the same axis. Using the likelihood ratio test, what is the decision boundary for this classifier? Assume equal prior probabilities.

A picture containing chart

Description automatically generated

T3. What happen to the decision boundary if the cat is happy with a prior of 0.8?

Chart, scatter chart

Description automatically generated with medium confidence

T4. Observe the histogram for Age, MonthlyIncome and DistanceFromHome. How many bins have zero counts? Do you think this is a good discretization? Why?

Text

Description automatically generatedGraphical user interface, histogram

Description automatically generated

In my opinion, this is almost a good discretization except for the distancefromhome feature that made this discretization underperformed.

T5. Can we use a Gaussian to estimate this histogram? Why? What about a Gaussian Mixture Model (GMM)?

To my mind, it is possible to use a Gaussian to estimate the histogram. This could be done by using the MLE process. First, we need to find the parameter of the Gaussian which are mean and standard deviation. From the data, it is quite straightforward to find these two values. Moreover, in order to use GMM, we need to define the number of the gaussian we want to fit with the data. In this case, only one Gaussian might be enough to describe the distribution of the data. As a result, it may be excessive to apply GMM.

T6. Now plot the histogram according to the method described above (with 10, 40, and 100 bins) and show 3 plots for Age, MonthlyIncome, and DistanceFromHome. Which bin size is most sensible for each features? Why?

Chart, bar chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedGraphical user interface, chart, application, histogram

Description automatically generated

From the above figures, the 40 and 100 bins might not be appropriate since some of the bins contain the value of zero which is hard to infer about the real distribution of the data. As a result, the 10-bin is the best choice for binning.

T7. For the rest of the features, which one should be discretized? What are the criteria for choosing whether we should discretize a feature or not? Answer this and discretize those features into 10 bins each. In other words, figure out the bin edge for each feature, then use digitize() to convert the features to discrete values.

Basically, we want to inference the real distribution of the data. Consequently, we don’t want any of the bin being zero. Hence, we will consider the interval and ratio data types to be transform by binning method. The features that we will transform are

['Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'].

T8. What kind of distribution should we use to model histograms? (Answer a distribution name) What is the MLE for the likelihood distribution? (Describe how to do the MLE). Plot the likelihood distributions of MonthlyIncome, JobRole, HourlyRate, and MaritalStatus for different Attrition values.

We will use Gaussian distribution to model the histograms. The parameters of the Gaussian distribution are mean and variance. To find the these attributes, we must consider the pdf of the Gaussian function. Then, we take the derivative respected to each parameter and set the whole derivative term to zero. Finally, we can obtain the formula to calculate for the mean and variance.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

T9. What is the prior distribution of the two classes?

Bernoulli distribution because there are only two possible class which are 0 and 1.

T10. If we use the current Naive Bayes with our current Maximum Likelihood Estimates, we will find that some P(xi |attrition) will be zero and will result in the entire product term to be zero. Propose a method to fix this problem.

Instead using zero, we will use the very low value for this case. We will apply 0.001 in spite of zero.

T11. Implement your Naive Bayes classifier. Use the learned distributions to classify the test set. Don’t forget to allow your classifier to handle missing values in the test set. Report the overall Accuracy. Then, report the Precision, 6 Recall, and F score for detecting attrition. See Lecture 1 for the definitions of each metric.

Record the distribution of training data in each feature and each class.

Graphical user interface, text, application

Description automatically generated

Calculate the prior terms.

Text

Description automatically generated

Bin the test dataset as the same method as the training set. We also use the train\_set to find the binning range.

Text

Description automatically generated

Fit the model to the test dataset.

Text

Description automatically generated with medium confidence

T12. Use the learned distributions to classify the test set. Report the results using the same metric as the previous question.

Record the mean and standard deviation of the features that were originally be binned.

Graphical user interface, text, application, email

Description automatically generated

Apply the new model to the test dataset.

Text

Description automatically generated

T13. The random choice baseline is the accuracy if you make a random guess for each test sample. Give random guess (50% leaving, and 50% staying) to the test samples. Report the overall Accuracy. Then, report the Precision, Recall, and F score for attrition prediction using the random choice baseline.

Text, letter

Description automatically generated

T14. The majority rule is the accuracy if you use the most frequent class from the training set as the classification decision. Report the overall Accuracy. Then, report the Precision, Recall, and F score for attrition prediction using the majority rule baseline.

Text

Description automatically generated

T15. Compare the two baselines with your Naive Bayes classifier.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-score |
| model from T11 | 0.8013 | 0.3928 | 0.4782 | 0.4313 |
| model from T12 | 0.8424 | 0.5000 | 0.0434 | 0.0800 |
| Random baseline | 0.5000 | 0.1323 | 0.3913 | 0.1978 |
| Majority baseline | 0.8424 | nan | 0.0 | nan |

The models seem to perform better than the baselines.

T16. Use the following threshold values

t = np.arange(-5,5,0.05)

find the best accuracy, and F score (and the corresponding thresholds)

Text

Description automatically generated

T17. Plot the RoC of your classifier.

Chart, line chart

Description automatically generated

T18. Change the number of discretization bins to 5. What happens to the RoC curve? Which discretization is better? The number of discretization bins can be considered as a hyperparameter, and must be chosen by comparing the final performance.

Chart, line chart

Description automatically generated

Considering the area under RoC, I believe that the 5-bucket is better than 10-bucket.