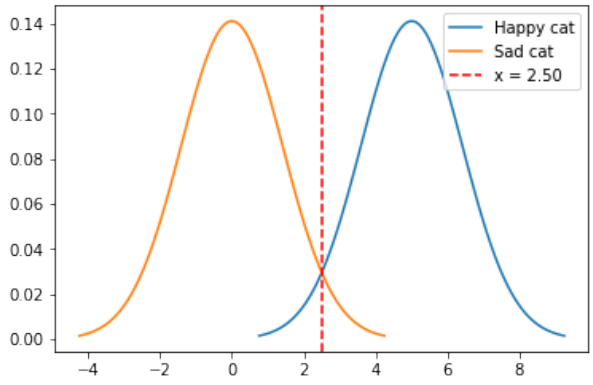
Nisaruj Rattanaaram 6031033521

Homework 2 MLE and Naive Bayes

Simple Bayes Classifier

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.



T2. $P(x|w_1) = N(5,2)$

$P(x|w_2) = N(0,2)$

$$\frac{P(x|w_1)}{P(x|w_2)} \square \frac{P(w_2)}{P(w_1)}$$

∵ prior เท่ากัน ∴ $P(w_2)/P(w_1) = 1$

$$\frac{P(x|w_1)}{P(x|w_2)} \square 1$$

PDF $\quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\frac{1}{\sqrt{2\pi\cdot 2}} e^{-\frac{(x-5)^2}{2\cdot 2}} \cdot \sqrt{2\pi\cdot 2}\, e^{\frac{(x-0)^2}{2\cdot 2}} \square 1$$

$$e^{\frac{-(x-5)^2+x^2}{4}} \square 1$$

take ln : $\frac{-(x-5)^2+x^2}{4} \square 0$

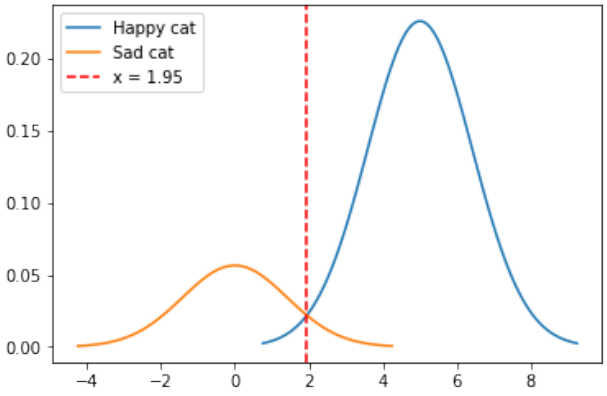
$(x+x-5)(x-x+5) \square 0$

$2x-5 \square 0$

$x \square 5/2$

Boundary อยู่ที่ x = 2.5

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



T3. $P(W_1) = 0.8$ ให้ $P(W_2) = 0.2$

จาก LRT : $\frac{P(x|W_1)}{P(x|W_2)} \square \frac{P(W_2)}{P(W_1)}$ $\frac{0.2}{0.8} = \frac{1}{4}$

แทน PDF ของ $N(\mu, \sigma^2)$ $\frac{1}{\sqrt{2\pi \cdot 2}} \cdot e^{-\frac{(x-5)^2}{2\cdot 2}} \cdot \sqrt{2\pi \cdot 2} \cdot e^{\frac{(x-0)^2}{2\cdot 2}} \square 2^{-2}$

take ln : $\frac{x^2 - (x-5)^2}{4} \square -2 \ln 2$

Solve x ให้ boundary อยู่ที่ $X \approx 1.9455$

OT2. For the ordinary case of $P(x|w1) = N(\mu1,\sigma^2)$, $P(x|w2) = N(\mu2,\sigma^2)$, $p(w1) = p(w2) = 0.5$, prove that the decision boundary is at $x = \mu1+\mu2 / 2$

OT2. $P(x|w_1) = N(\mu_1, \sigma^2)$
$P(x|w_2) = N(\mu_2, \sigma^2)$

จาก LRT :

$$\frac{P(x|w_1)}{P(x|w_2)} \square \frac{P(w_2)}{P(w_1)}$$

$\because P(w_2) = P(w_1) = 0.5$
$\therefore P(w_2) / P(w_1) = 1$

PDF ของ $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

จะได้ $\frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}} \square 1$

$e^{\frac{-(x-\mu_1)^2 + (x-\mu_2)^2}{2\sigma^2}} \square 1$

take ln : $(x-\mu_2)^2 - (x-\mu_1)^2 \cdot \frac{1}{2\sigma^2} \square 0$

$(x-\mu_2 - x + \mu_1)(x-\mu_2 + x - \mu_1) \square 0$

$2x - (\mu_1 + \mu_2) \square 0$

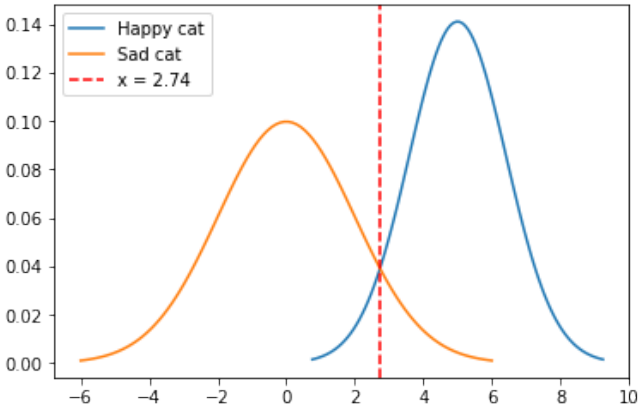
$x \square \frac{\mu_1 + \mu_2}{2}$

$\therefore$ สรุปได้ว่า decision boundary อยู่ที่ $x = \frac{\mu_1 + \mu_2}{2}$ ▭

If the student changed his model to

P(x|w1) = N(5,2)
P(x|w2) = N(0,4)

• Plot the posteriors values of the two classes on the same axis. What is the decision boundary for this classifier? Assume equal prior probabilities.



OT 2.2 $P(x|w_1) = N(5,2)$
$P(x|w_2) = N(0,4)$

จาก LRT และ $P(w_1) = P(w_2)$ ได้ว่า

$$\frac{P(x|w_1)}{P(x|w_2)} \square 1$$

PDF ของ $N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\frac{\frac{1}{\sqrt{2\pi(2)}} e^{-\frac{(x-5)^2}{2\cdot 2}}}{\frac{1}{\sqrt{2\pi(4)}} e^{-\frac{(x-0)^2}{2\cdot 4}}} \square 1$$

$$e^{\frac{-2(x-5)^2 + x^2}{8}} \square 1/\sqrt{2}$$

take ln : $\frac{-2(x-5)^2 + x^2}{8} \square \ln\left(\frac{1}{\sqrt{2}}\right)$

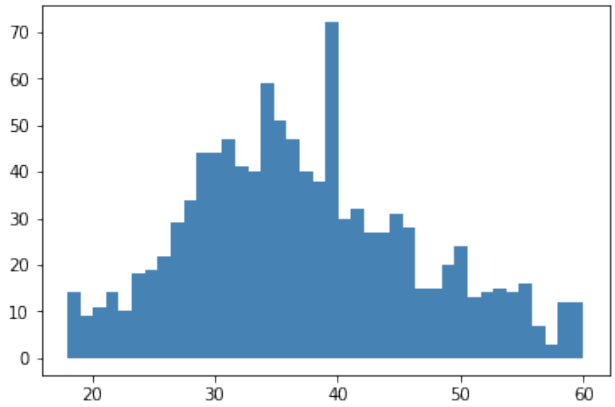
$-2(x-5)^2 + x^2 \square 8\ln(1/\sqrt{2})$

Solve X ได้ boundary คือ $X \approx 2.7355$

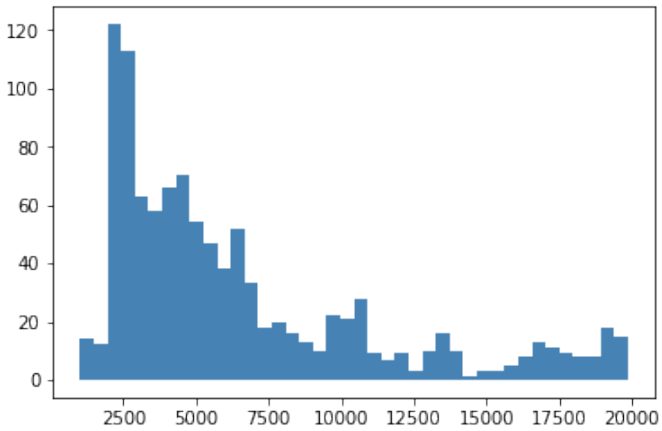
Employee Attrition Prediction

Histogram discretization

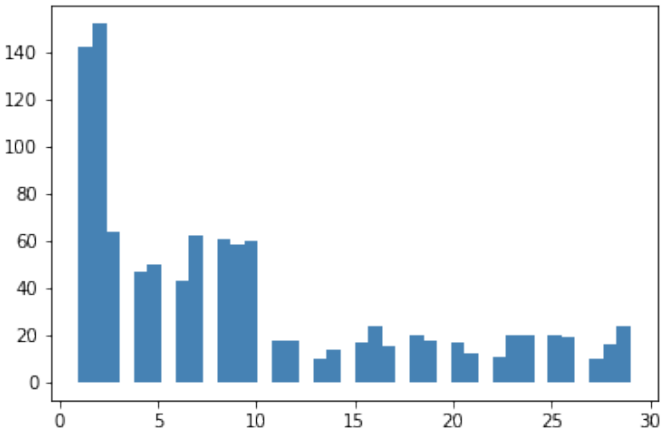
T4. Observe the histogram for Age, MonthlyIncome and DistanceFromHome. How many bins have zero counts? Do you think this is a good discretization? Why?



Age (# bins = 0)



MonthlyIncome (# bins = 0)



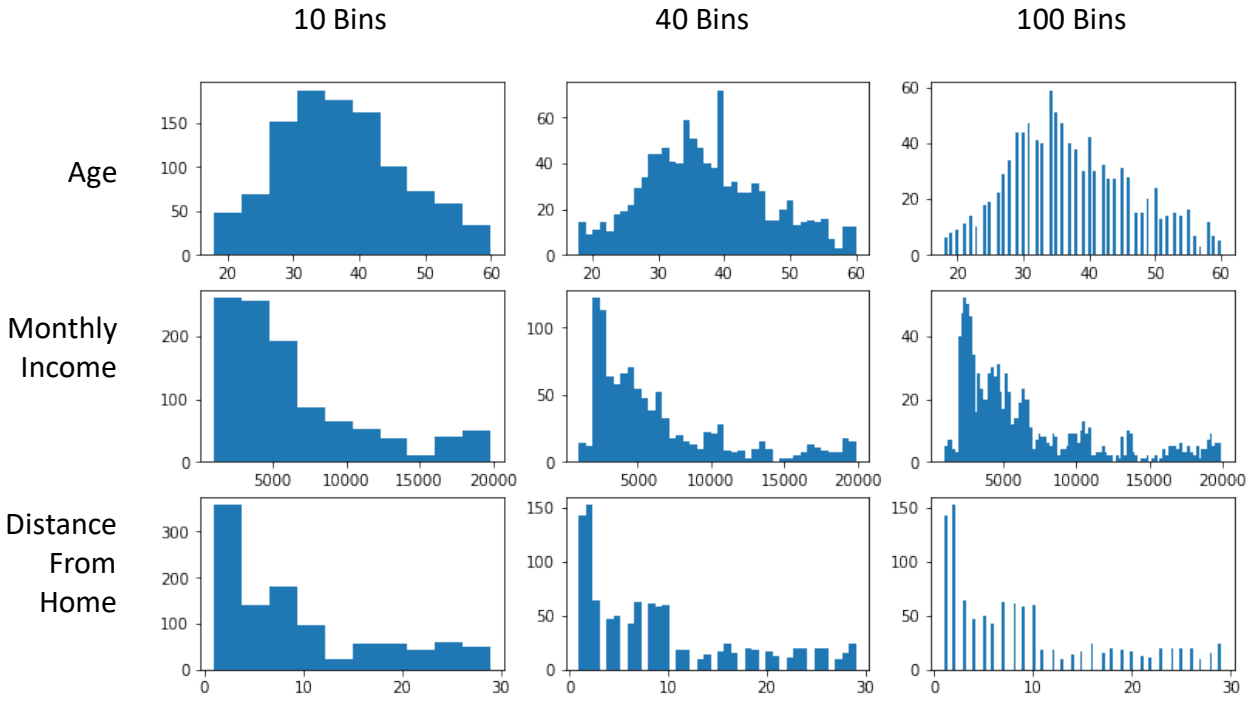
DistanceFromHome (# bins = 11)

ใน feature Age และ MonthlyIncome สามารถทำ discretization ได้ดี ในขณะที่ DistanceFromHome ทำได้ไม่ดีเท่าที่ควร เนื่องจากมีหลาย bin ที่ไม่มีข้อมูล ทำให้กราฟที่ได้ไม่ต่อเนื่อง

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

ในบาง feature สามารถใช้ Gaussian distribution ได้ แต่ในบางครั้งอาจจะได้ผลไม่ดีเพราะ model ของเราอาจจะไม่ใกล้เคียง normal แต่ถ้าใช้ GMM ซึ่งสามารถมี normal ได้หลายลูกได้ทำให้มีความ flexible มากกว่า ทั้งนี้ก็ต้องระวังการเกิด overfitting ด้วย

T6. 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?

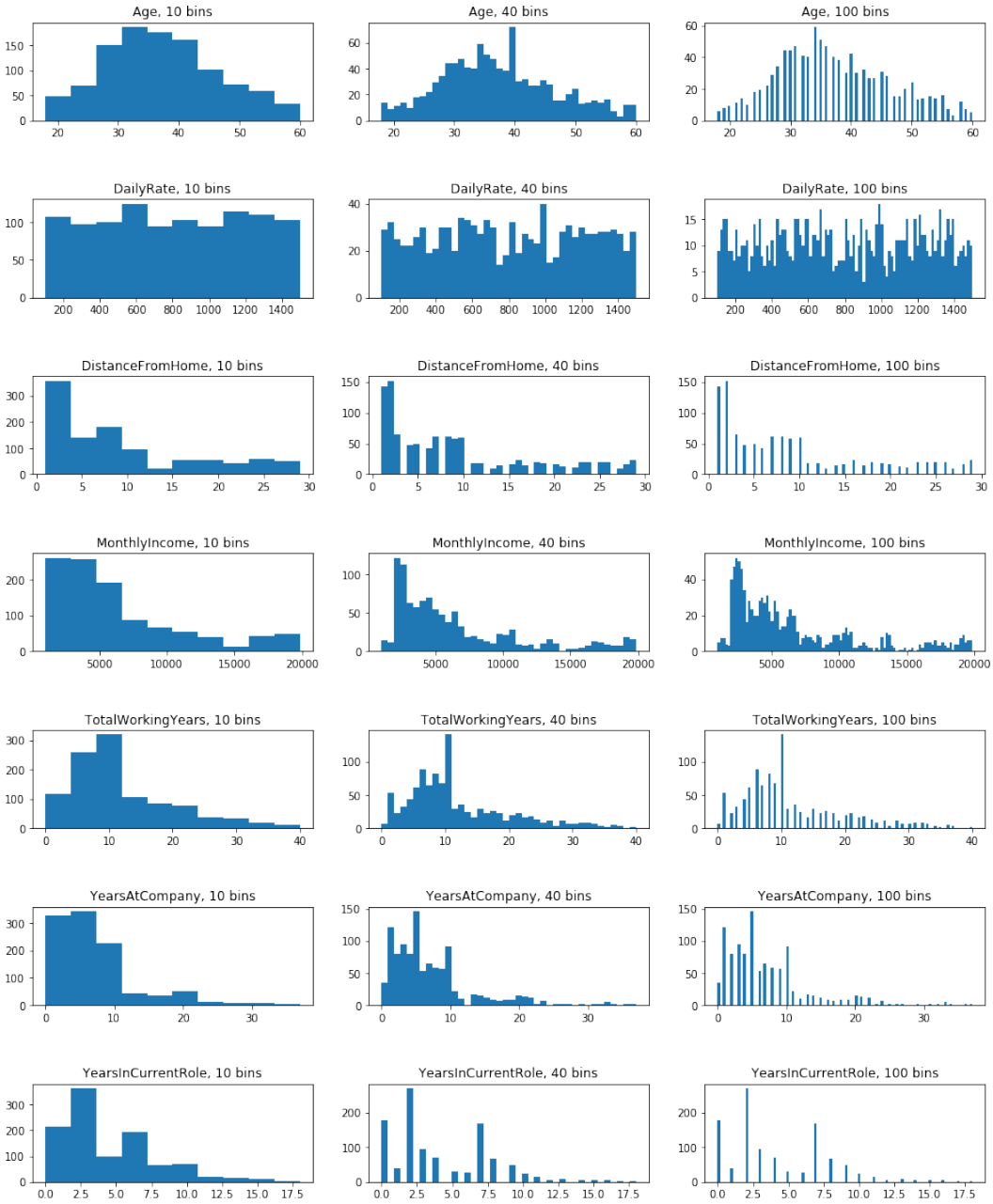


Age: 40 bins, **MonthlyIncome**: 100 bins, **DistanceFromHome**: 10 bins

เราสามารถเลือกจำนวน bin ได้จากกราฟที่มีความถี่ของช่วงที่เยอะ แต่จะต้องไม่ให้มีช่องว่างที่ bin_count = 0 มากเกินไป

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.

Feature ที่มักนำมา discretize ได้มักจะเป็นข้อมูลเชิงปริมาณและเมื่อทำการ discretize แล้วควรจะได้ค่าที่มีความต่อเนื่องกัน ได้แก่ Age DailyRate DistanceFromHome MonthlyIncome TotalWorkingYears YearsAtCompany YearsInCurrentRole

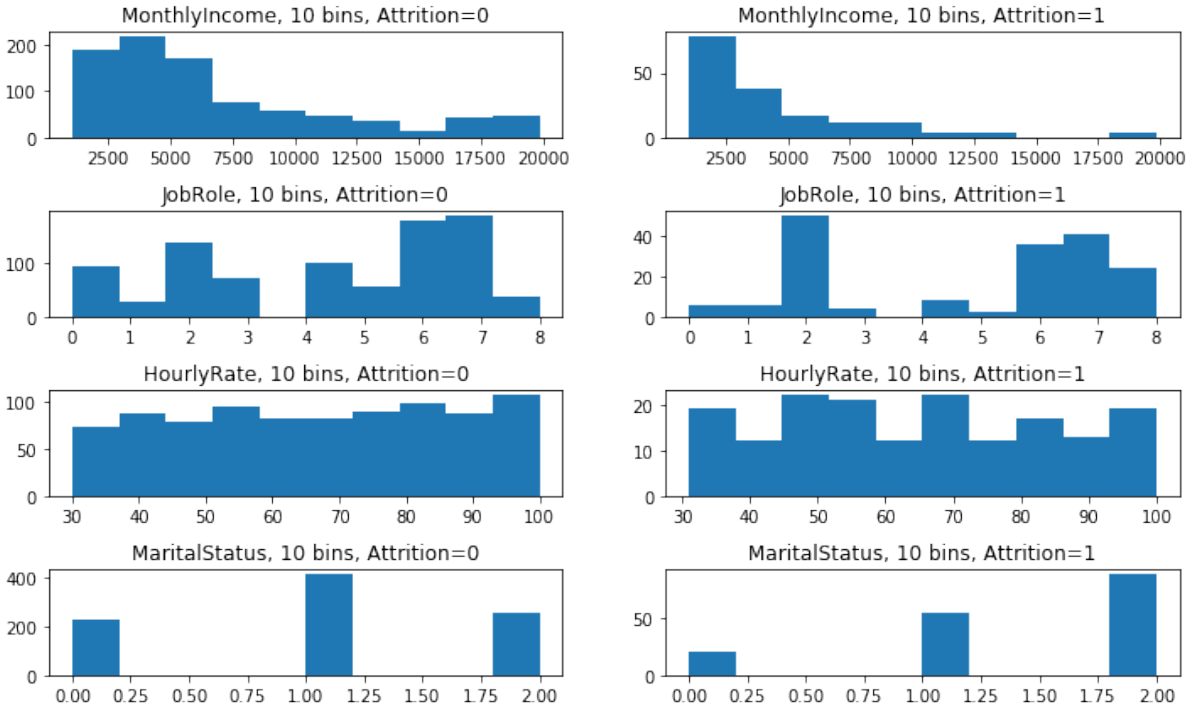


The MLE for the likelihood distribution of discretized histograms

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.

Binomial Distribution

$MLE = \frac{k}{N}$ โดยให้ k เป็น จำนวนข้อมูลใน bin ที่ feature นั้นตกอยู่ และ N คือจำนวนข้อมูลของ $class_i$



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

$$P(leave) = \frac{n(leave)}{N} = 0.162$$

$$P(stay) = \frac{n(stay)}{N} = 0.838$$

Naive Bayes classification

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.

เมื่อ P(xi | attrition) มีค่าเป็นศูนย์ ให้เราเปลี่ยนเป็น 1 แทน เพื่อป้องกันไม่ให้ผลคูณเป็นศูนย์ หรืออีกวิธีสามารถทำได้โดยใช้ Laplace smoothing

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, Recall, and F score for detecting attrition. See Lecture 1 for the definitions of each metric.

Discretized Naïve Bayes		
	Detected Yes	Detected No
Actual Yes	8	15
Actual No	9	114

$$Precision = \frac{n(True\ Positive)}{n(Detected\ Yes)} = 0.471$$

$$Recall = \frac{n(True\ Positive)}{n(Actual\ Yes)} = 0.348$$

$$F\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 0.4$$

Probability density function

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

PDF Naïve Bayes		
	Detected Yes	Detected No
Actual Yes	4	19
Actual No	4	119

Precision = 0.5

Recall = 0.174

F score = 0.258

Baseline comparison

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.

Random choice baseline		
	Detected Yes	Detected No
Actual Yes	14	9
Actual No	55	68

Precision = 0.203

Recall = 0.609

F score = 0.304

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.

Majority rule baseline		
	Detected Yes	Detected No
Actual Yes	0	23
Actual No	0	123

Precision = NaN

Recall = 0

F score = NaN

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

สำหรับ PDF Naïve Bayes มี F-score ใกล้เคียงกับของ Random baseline โดยเฉลี่ย การทำ classifier ด้วยวิธีนี้จึงให้ความแม่นยำไม่ต่างจากการสุ่ม

ในขณะที่ Discretized Naïve Bayes มี F-score = 0.4 ซึ่งมากกว่า Random baseline จึงอาจสรุปได้ว่าการทำ classifier ด้วยวิธีนี้สามารถจำแนกได้ดีกว่า

* ส่วน Majority rule baseline ไม่สามารถวัดได้เพราะ Precision = 0/0

Threshold finding

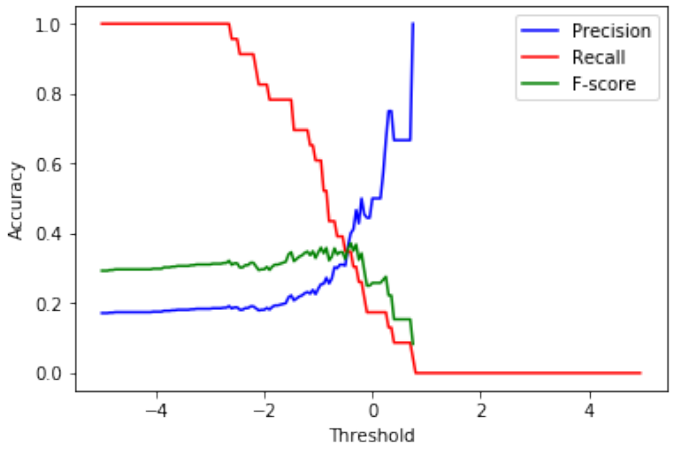
T16. Use the following threshold values `t = np.arange(-5,5,0.05)` find the best accuracy, and F score (and the corresponding thresholds)

หากเรา maximize F-score จะพบว่า threshold = -0.4 ให้ค่า F-score สูงสุด

Precision = 0.4

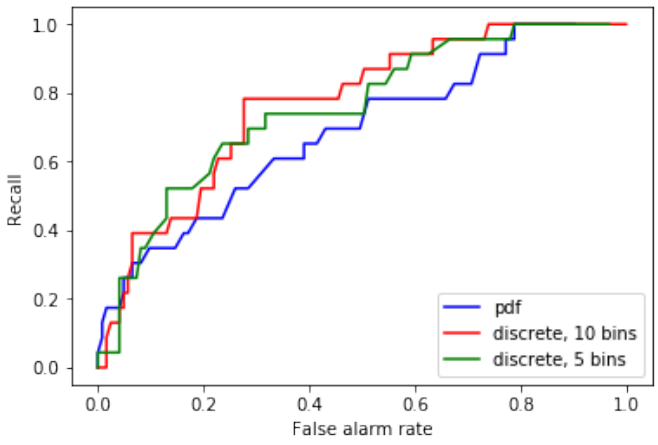
Recall = 0.348

F score = 0.372



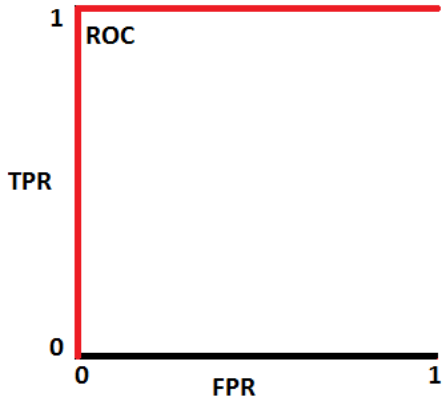
Receiver Operating Characteristic (RoC) curve

T17. Plot the RoC of your classifier.

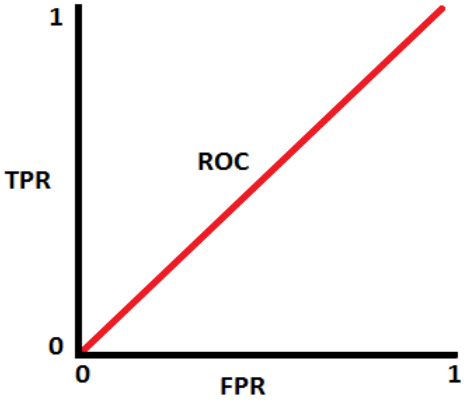


T18. Change the number of discretization bins to 5. What happens to the RoC curve? Which discretization is better? The number of discretization bins can be considered as a hyperparameter, and must be chosen by comparing the final performance.

การตรวจสอบประสิทธิภาพของ classifier จาก RoC curve สามารถทำได้โดยการดูว่ากราฟนั้นชิดมุมซ้ายบนมากเพียงใด หากเป็นดังรูปที่ 1 จะหมายถึง classifier ในอุดมคติ นั่นคือสามารถจำแนกได้อย่างไม่ผิดพลาดเลย แต่หาก curve นั้นมีความเข้าใกล้เส้น y = x มากดังรูปที่ 2 จะหมายความว่า classifier ของเราแทบไม่มีสามารถในจำแนกประเภทได้เลย (หากกราฟ RoC มีลักษณะตรงข้ามกับรูปที่ 1 จะแสดงให้เห็นว่า classifier ของเรามีความสามารถในการตอบสวนทางกับความเป็นจริง)



รูปที่ 1 Ideal Situation



รูปที่ 2 Worst Situation

จาก RoC ในข้อ 17 จะเห็นได้ว่า curve ของ 5 bins จะเข้าใกล้เส้นตรง y = x มากกว่า 10 bins นั่นหมายความว่า 10 bins discretization นั้นสามารถจำแนกประเภทได้ดีกว่า 5 bins

T19. Submit your predictions and code on mycourseville.

https://colab.research.google.com/drive/10HdbdybJtSUkfLrHjvZpHsilOVWYCp_M

(Optional) Classifier Variance

OT3. Shuffle the database, and create new test and train sets. Redo the entire training and evaluation process 10 times (each time with a new training and test set). Calculate the mean and variance of the accuracy rate.

F-score Mean = 0.283

F-score Variance = 0.008