นาย ณัติชัย สูติพันธ์วิหาร 5931023821

T1. For normal distribution, $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

Find $\underset{\alpha}{\operatorname{argmax}}[p(y_2, y_1, y_0 | \alpha)]$

$$= \underset{\alpha}{\operatorname{argmax}}[p(y_0)\, p(y_1 | y_0, \alpha)\, p(y_2 | y_1, \alpha)]$$

$$= \underset{\alpha}{\operatorname{argmax}}[\ln(p(y_0)) + \ln(p(y_1 = \alpha y_0 + \omega_0)) + \ln(p(y_2 = \alpha y_1 + \omega_1))]$$

$$= \underset{\alpha}{\operatorname{argmax}}\left[\ln\left(\frac{1}{\sqrt{2\pi\lambda}}\right) - \frac{y_0^2}{2\lambda} + 2\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_1 - \alpha y_0)^2}{2\sigma^2} - \frac{(y_2 - \alpha y_1)^2}{2\sigma^2}\right]$$

Then $0 = -\frac{(y_1 - \alpha y_0)(-y_0)}{\sigma^2} - \frac{(y_2 - \alpha y_1)(-y_1)}{\sigma^2}$

$$0 = y_0 y_1 - \alpha y_0^2 + y_1 y_2 - \alpha y_1^2$$

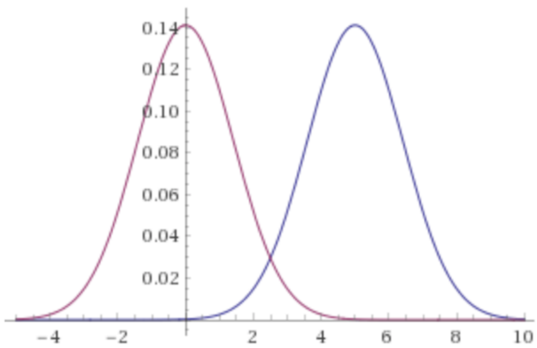
$\therefore\ \alpha = \frac{y_1(y_0 + y_2)}{(y_0^2 + y_1^2)}$ which will maximize $p(y_2, y_1, y_0 | \alpha)$

OT1. Same as **T1.**, then we get

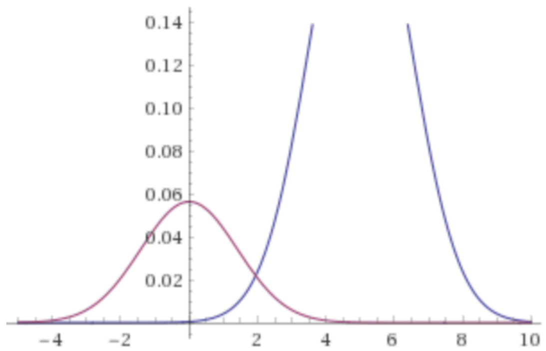
$$0 = \sum_{i=0}^{N} \frac{(y_{i+1} - \alpha y_i)(y_i)}{\sigma^2}$$

$$\therefore\ \alpha = \frac{\sum_{i=0}^{N} y_i y_{i+1}}{\sum_{i=0}^{N} y_i^2}$$

T2. The decision boundary is 2.5.



T3. The decision boundary is about 1.945, which is less than **T2.**



OT2. Find the intersection of 2 graphs, so

$$(0.5)\frac{e^{\frac{-(x-\mu_1)^2}{2\sigma^2}}}{2\sigma^2} = \frac{e^{\frac{-(x-\mu_2)^2}{2\sigma^2}}}{2\sigma^2}(0.5)$$

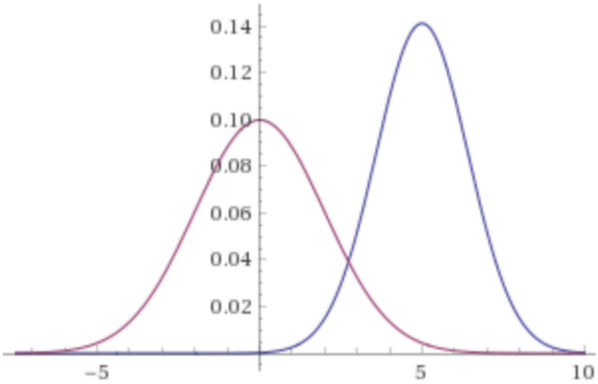
$$(x-\mu_1)^2 = (x-\mu_2)^2$$

$$(2x-\mu_1-\mu_2)(\mu_2-\mu_1) = 0$$

Assume that $\mu_1 \neq \mu_2$

$\therefore\ x = \frac{\mu_1+\mu_2}{2}$ is the decision boundary.

The decision boundary is about 2.736 and 17.265, where x between 2.736 and 17.265 is the happy cat.



T4. Most histograms are not distributed in the normal distribution, so we should not estimate them by a Gaussian method. Additionally, since the data is discrete, and we have too little of them, so we should not estimate them by GMM method.

T5. There are 866 bins with zero count in all features, so I think it is not good to discretize them this way, because, more zero bins we have, more probability that test data will fall in these zero bins.

T6. Age feature should be discretized into 40 bin size. Monthly income feature should be discretized into 100 bin size. Distance from home feature should be discretized into 10 bin size.

T7. Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, PercentSalaryHike, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager should be discretized. The criteria for choosing a feature is whether the number of the unique values of the feature is more than 10 or not. The number 10 comes from the greatest number of all categorical features.

T8. Currently, we discretized all features to be categorical in less than 10 categories, so we should use discrete probability distribution to model the histograms. To find the MLE of each feature, we should find log likelihood instead to prevent floating point underflow. The result is

Feature	Log likelihood
Age	-3785.09
BusinessTravel	-396.33
DailyRate	-7661.18
Department	-391.78
DistanceFromHome	-3178.33
Education	-1242.64
EducationField	-1314.97
EmployeeCount	0.00
EnvironmentSatisfaction	-1119.80
Gender	16.57
HourlyRate	-4471.01
JobInvolvement	-799.95
JobLevel	-1170.43
JobRole	-2013.52
JobSatisfaction	-1125.00
MaritalStatus	-672.75
MonthlyIncome	-10064.77

Feature	Log likelihood
MonthlyRate	-10681.41
NumCompaniesWorked	-2000.50
Over18	0.00
OverTime	131.06
PercentSalaryHike	-2503.55
PerformanceRating	295.04
RelationshipSatisfaction	-1149.69
StandardHours	0.00
StockOptionLevel	-841.15
TotalWorkingYears	-3457.98
TrainingTimesLastYear	-1488.30
WorkLifeBalance	-798.30
YearsAtCompany	-3047.13
YearsInCurrentRole	-2497.60
YearsSinceLastPromotion	-1815.44
YearsWithCurrManager	-2330.21

T9. The prior distribution of the 2 classes is

p(leave) = 16.16%

p(stay) = 83.84%.

T10. By using histogram discretization, the result is

Accuracy: 73.97 %   Precision: 29.73 %

Recall: 47.83 %   F score: 36.67 %

T11. By assuming the features are normally distributed, the result is

Accuracy: 82.19 %   Precision: 45.45 %

Recall: 65.22 %   F score: 53.57 %

T12. By using the random choice baseline, the result is

Accuracy: 47.26 %   Precision: 13.51 %

Recall: 43.48 %   F score: 20.62 %

T13. By assuming the features are normally distributed, the result is

Accuracy: 84.25 %   Precision: 0.00 %

Recall: 0.00 %   F score: 0.00 %

T14. The accuracies between **T11.** and **T13.** are close together, but **T11.** get the most F score, so normal distributed model is the best way to classify attrition from test data.

T15. For discrete distribution,

the best accuracy is 85.62 % at threshold 4.25, and

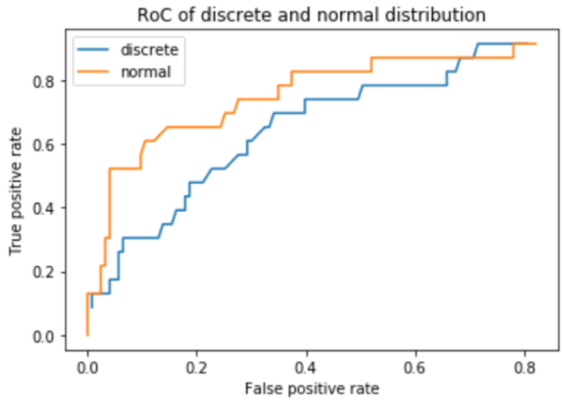
the best F-score is 39.51 % at threshold -1.30.

For normal distribution,

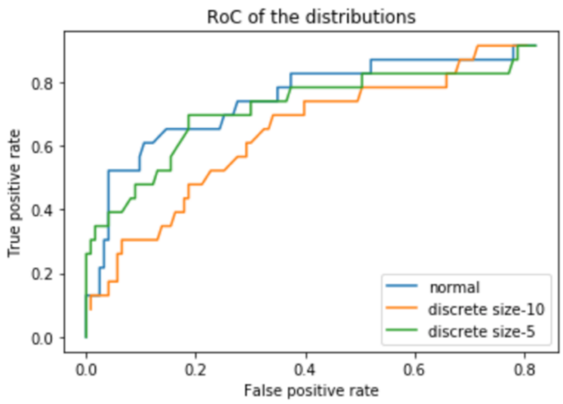
the best accuracy is 89.04 % at threshold 0.75, and

the best F-score is 60.00 % at threshold 0.75.

T16.



T17.



The RoC curves, for the most along the lines, the size-5 bins discretization's graph has the value of true positive rate more than the size-10's one.

In addition, for the size-5 bins discretization,

the best accuracy is 88.36 % at threshold 1.10, and

the best F-score is 51.61 % at threshold -0.70.

Consequently, the size-5's one is better than the size-10's one.

T18. File 'HW2.pdf' and 'HW2.ipynb' summited to mycourseville.

OT3. For discrete distribution, the mean of accuracy rate is 87.12 % with variance 1.43.

For normal distribution, the mean of accuracy rate is 87.67 % with variance 1.84.