

MLE

Consider the following very simple model for stock pricing. The price at the end of each day is the price of the previous day multiplied by a fixed, but unknown, rate of return, α , with some noise, w . For a two-day period, we can observe the following sequence

$$y_2 = \alpha y_1 + w_1$$

$$y_1 = \alpha y_0 + w_0$$

where the noises w_0, w_1 are iid with the distribution $N(0, \sigma^2)$, $y_0 \sim N(0, \lambda)$ is independent of the noise sequence. σ^2 and λ are known, while α is unknown.

T1. Find the MLE of the rate of return, α , given the observed price at the end of each day y_2, y_1, y_0 . In other words, compute for the value of α that maximizes $p(y_2, y_1, y_0 | \alpha)$

เพื่อที่ $y_0 \sim N(0, \lambda)$ และ $w_0, w_1 \sim N(0, \sigma^2)$ ดังนั้น $y_1 \sim N(\alpha y_0, \sigma^2)$ และ $y_2 \sim N(\alpha y_1, \sigma^2)$.

$$p(y_2, y_1, y_0 | \alpha) = p(y_2 | y_1, y_0, \alpha) p(y_1 | y_0, \alpha) p(y_0 | \alpha)$$

$$= p(y_2 | y_1, y_0, \alpha) p(y_1 | y_0, \alpha) p(y_0 | \alpha),$$

$$= p(y_2 | y_1, \alpha) p(y_1 | y_0, \alpha) p(y_0 | \alpha). \quad \text{By Markov Process Hint.}$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \exp(-\frac{1}{2} \cdot \frac{(y_2 - \alpha y_1)^2}{\sigma^2}) \frac{1}{\sigma \sqrt{2\pi}} \exp(-\frac{1}{2} \cdot \frac{(y_1 - \alpha y_0)^2}{\sigma^2}) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} \cdot \frac{(y_0 - 0)^2}{\lambda})$$

$$\log p(y_2, y_1, y_0 | \alpha) = -2\log(\sigma \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_2 - \alpha y_1}{\sigma} \right)^2 - \frac{1}{2} \log(\lambda \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_0}{\sqrt{\lambda}} \right)^2$$

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

\therefore เราต้อง $\underset{\alpha}{\operatorname{argmax}} \log p(y_2, y_1, y_0 | \alpha)$ ในการ MLE หา α ให้ได้.

$$\frac{d}{d\alpha} \log p(y_2, y_1, y_0 | \alpha) = -\frac{1}{\sigma^2} \frac{d}{d\alpha} (y_2 - \alpha y_1)(-y_1) - \frac{1}{\sigma^2} \frac{d}{d\alpha} (y_1 - \alpha y_0)(-y_0)$$

$$\frac{d}{d\alpha} \log p(y_2, y_1, y_0 | \alpha) = 0$$

$$+ (y_2 - \alpha y_1)(-y_1) + (y_1 - \alpha y_0)(-y_0) = 0$$

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

$$y_2 y_1 + y_1 y_0 = \alpha (y_1^2 + y_0^2)$$

$$\alpha = \frac{y_2 y_1 + y_1 y_0}{y_1^2 + y_0^2}$$

$$\therefore \text{MLE ของ } \alpha \text{ คือ } p(y_2, y_1, y_0 | \alpha) \text{ ที่ } \frac{y_2 y_1 + y_1 y_0}{y_1^2 + y_0^2}$$

OT1. Consider the general case, where

$$y_{n+1} = \alpha y_n + w_n, n = 0, 1, 2, \dots$$

Find the MLE given the observed price y_{N+1}, y_N, \dots, y_0

กำหนดให้ T_1 คือ

$$\frac{d}{d\alpha} \log (P(Y_n, Y_{n-1}, \dots, Y_2, Y_1, Y_0 | \alpha)) = \frac{1}{2\sigma^2} (Y_n - \alpha Y_{n-1}) (-Y_{n-1}) - \dots - \frac{1}{2\sigma^2} (Y_1 - \alpha Y_0) (Y_0)$$

$$\frac{d}{d\alpha} \log (P(Y_n, Y_{n-1}, \dots, Y_2, Y_1, Y_0 | \alpha)) = 0$$

$$(Y_n - \alpha Y_{n-1})(Y_{n-1}) + \dots + (Y_1 - \alpha Y_0) Y_0 = 0$$

$$Y_n Y_{n-1} + \dots + Y_1 Y_0 = \alpha [(Y_{n-1})^2 + \dots + Y_0^2]$$

$$\alpha = \frac{Y_n Y_{n-1} + \dots + Y_1 Y_0}{(Y_{n-1})^2 + \dots + Y_0^2}$$

Simple Bayes Classifier

A student in Pattern Recognition course had finally built the ultimate classifier for cat emotions. He used one input features: the amount of food the cat ate that day, x (Being a good student he already normalized x to standard Normal). He proposed the following likelihood probabilities for class 1 (happy cat) and 2 (sad cat)

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

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

* Notation
 p : pdf
 P : pmf
 F : cdf

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.

$$\text{Posterior } P(w_1|x) = \frac{P(x|w_1)P(w_1)}{P(x)} \quad \text{and} \quad P(w_2|x) = \frac{P(x|w_2)P(w_2)}{P(x)}.$$

แล้ว Prior probability ของทั้งสองคลาสเท่ากัน ที่ $P(w_1) = P(w_2) = 0.5$

ลงมือ ทำให้ในกรณี x ใน Class 1 ผลลัพธ์ $P(w_1|x) > P(w_2|x)$

ซึ่ง $P(w_1|x) > P(w_2|x)$

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

$$\frac{P(x|w_1)}{P(x|w_2)} > \frac{P(w_2)}{P(w_1)} = 1.$$

แล้วจึงสามารถหา Decision Boundary โดยที่ x ที่ทำให้ $\frac{P(x|w_1)}{P(x|w_2)} > 1$ นั่น.

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

$$\frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} > 1$$

$$\exp\left(\frac{1}{2}\left(\frac{x-5}{\sqrt{2}}\right)^2 - \frac{1}{2}\left(\frac{x}{\sqrt{2}}\right)^2\right) > 1$$

$$\frac{-10x+25}{4} > 0$$

$$x > 2.5.$$

∴ เวดจ์ Decision Boundary ที่ $x=2.5$ ถ้า $x > 2.5$ จะสูงกว่าใน Class 1

ถ้า $x < 2.5$ จะสูงกว่าใน Class 2.

in Posterior value.

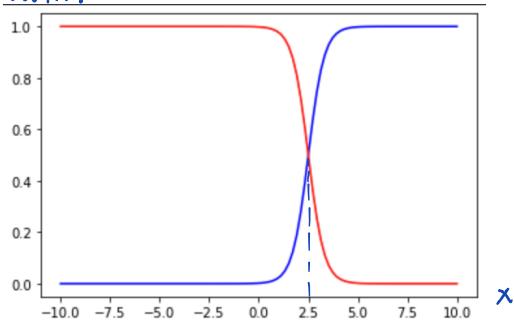
$$P(w=w_1|x) = \frac{p(x|w_1)p(w_1)}{p(x)}$$

$$= \frac{p(x|w_2)p(w_2)}{p(x)}$$

$$= \frac{p(x|w_1)p(w_1)}{p(x|w_1)p(w_1) + p(x|w_2)p(w_2)} \quad \because p(x) = \sum_{w \in \{w_1, w_2\}} p(x,w) = \sum p(x|w)p(w).$$

$$= \frac{p(x|w_1)}{p(x|w_1) + p(x|w_2)}. \quad \text{ผลลัพธ์คือ} \rightarrow P(w=w_2|x).$$

$$P(w|x)$$



จุดนี้ $\begin{cases} w_1 \\ w_2 \end{cases}$

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

โจทย์การนับว่าเดิมก่อน T3. ให้ $P(w_1) = 0.8 \quad P(w_2) = 0.2$

การนับนี้ x 屬于 Class 1 ให้ $P(w_1|x) > P(w_2|x)$

ซึ่ง $P(w_1|x) > P(w_2|x)$

$$\frac{p(x|w_1)p(w_1)}{p(x)} > \frac{p(x|w_2)p(w_2)}{p(x)} \quad -x^2+10x-25 > x^2$$

$$\frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} = \frac{0.2}{0.8} = \frac{1}{4}$$

โดยใช้รากที่ 4 ของนี้

$$\frac{p(x|w_1)}{p(x|w_2)} > \frac{1}{4} \quad \exp\left(\frac{1}{2}\left(\frac{x-5}{\sqrt{2}}\right)^2 + \frac{1}{2}\left(\frac{x}{\sqrt{2}}\right)^2\right) > \frac{1}{4} \quad 25 - 2\log 4 < 10x \quad *: > \frac{25 - 4\log 4}{10}$$

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right) > \frac{1}{4} \quad -\frac{-10x+25}{4} > -\log 4$$

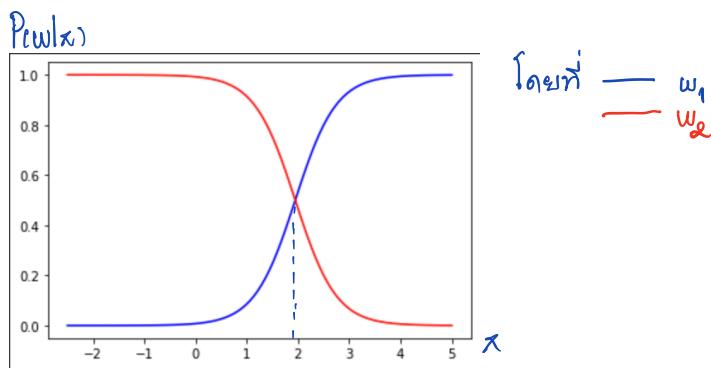
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) > \frac{1}{4} \quad -\frac{10x+25}{4} < \log 4$$

$$\approx 1.945$$

\therefore เรากำหนด Decision Boundary ลงที่ $x = 1.945$ ถ้า $x \geq 1.945$ จะส่งผู้มาเข้า Class 1
ถ้า $x \leq 1.945$. จะส่งผู้มาเข้า Class 2.

in Posterior Value

$$\begin{aligned} P(w=w_1|x) &= \frac{p(x|w_1)p(w_1)}{p(x|w_1)p(w_1) + p(x|w_2)p(w_2)} \\ &= \frac{p(x|w_1)(0.8)}{p(x|w_1)(0.8) + p(x|w_2)(0.2)} \\ &= \frac{4p(x|w_1)}{4p(x|w_1) + p(x|w_2)}. \quad \text{น้ำ: } P(w=w_2|x) = \frac{p(x|w_2)}{4p(x|w_1) + p(x|w_2)}. \end{aligned}$$



OT2. For the ordinary case of $P(x|w_1) = N(\mu_1, \sigma^2)$, $P(x|w_2) = N(\mu_2, \sigma^2)$, $p(w_1) = p(w_2) = 0.5$, prove that the decision boundary is at $x = \frac{\mu_1 + \mu_2}{2}$

If the student changed his model to

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

$$P(x|w_2) = N(0, 4)$$

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

เราสามารถหาค่า x ที่โดยใช้แนวตัดขวางที่ T_2 และ T_3 นี้

ถ้า $p(w_1) = p(w_2) = 0.5$ นั่นก็ เรากำหนดทาง Decision Boundary ได้

$$\text{โดย } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_1)}{P(w_2)} = 1.$$

$$\frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right)}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2\right)} > 1$$

$$\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right) > \exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2\right)$$

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

$$2x\mu_1 + \mu_1^2 > x^2 - 2x\mu_2 + \mu_2^2$$

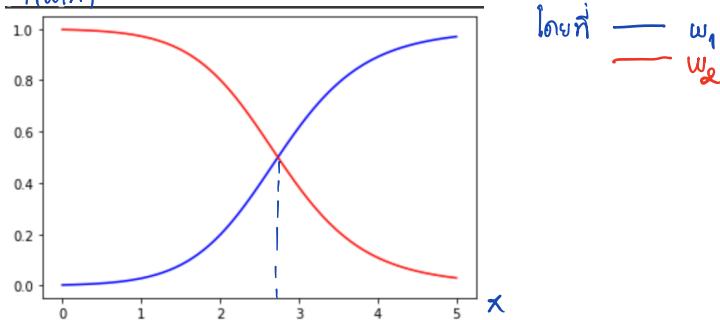
$$2x(\mu_2 - \mu_1) > \mu_2^2 - \mu_1^2$$

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

\therefore Decision boundary คือ $\frac{\mu_2 + \mu_1}{2}$

ສ່ວນອົງກົມທີ່ຈົດຕະຫຼາດ.

$P(w|x)$



ມະ Posterior value.

$$P(w=w_1|x) = \frac{p(x|w_1)}{p(x|w_1) + p(x|w_2)} \quad \text{ແມ່ນກໍລຳຍາຄລື່ງກັນໃນ } P(w=w_2|x). \quad \text{ຈາກນີ້ T2.}$$

ແລ້ວມະ Decision Boundary ມີ $\frac{p(x|w_1)}{p(x|w_2)} > 1$ ໄວ້ກີ່ເກີນຕອງກັບ T2 ເພີ້ນກັນ.

$$\frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{x^2}{2}\right)} > 1$$

$$\exp\left(-\frac{(x-5)^2}{2} + \frac{x^2}{2}\right) > \frac{1}{\sqrt{2}}$$

$$-\frac{1}{2}\left(\frac{x-5}{\sqrt{2}}\right)^2 + \frac{1}{2}\left(\frac{x}{\sqrt{2}}\right)^2 > -\frac{1}{2}\log 2$$

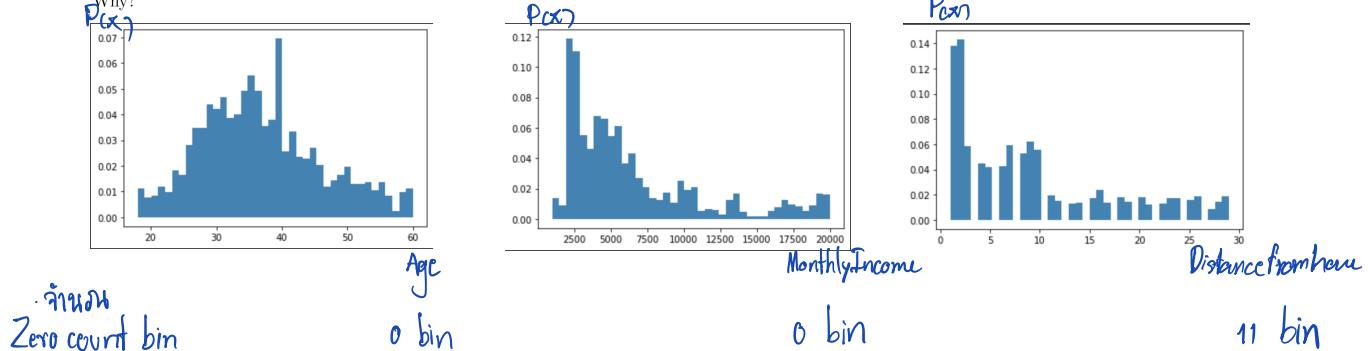
$$\frac{2(x^2 - 10x + 25) - x^2}{4} < -\frac{1}{2}\log 2$$

$$x^2 - 20x + 50 - 4\log 2 < 0$$

$$x = \frac{20 \pm \sqrt{400 - 4(1)(50 - 4\log 2)}}{2(1)} \approx 2.73.$$

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

Why?

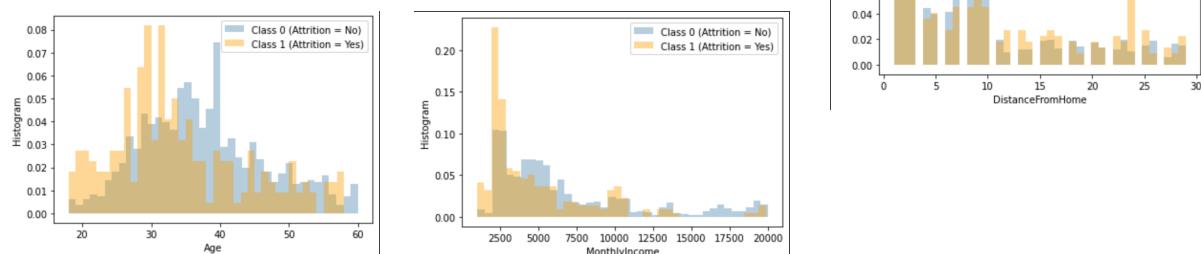


ในกรณี Discretization นี้ ถือว่าเป็นการ Discretization ก่อตัวในรูปแบบ เนื่องจากมีบล็อกที่ไม่ต่อเนื่องกัน เช่น Bin 1 ไม่ติดกับ Bin 2 ของ Feature Distance From Home ทำสองครั้งที่ Count zero อยู่ด้วย ซึ่งอาจส่งผลต่อประสิทธิภาพ

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

เราสามารถใช้ Gaussian ในการประมาณ Histogram ได้เช่นเดียวกับ Feature Age อย่างเดียว เมื่อจากลักษณะของ Feature นี้เป็นแบบต่อเนื่อง: Bin 1 ของ Feature Age ลักษณะ: เป็นตัวเลขคู่ ส่วนอีกสอง Feature ลักษณะ: ตัวเลขคี่ ไม่สามารถคาดคะเนได้

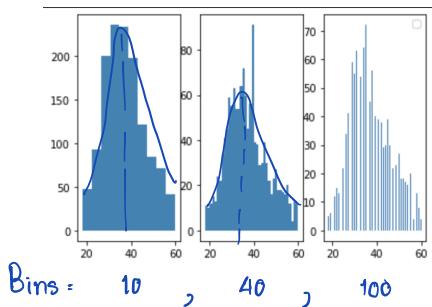
ลองใช้ Gaussian Mixture Model ดูให้ดู . โดยให้ทั้ง 2 Class 0 และ 1 สามารถสอดคล้องกันได้ด้วย
บน Feature Age เราสามารถเห็นได้ว่าสอดคล้องกันโดยทั่วไปของ Gaussian. สำหรับ Feature Monthly Income จะพบว่ามี Gaussian ที่เพ้นท์มาที่ทางขวาของเส้นตัวเลข. และ Feature DistanceFromHome ลักษณะ: หางทางขวา: ความถี่ลดลงเรื่อยๆ แต่ Gaussian Mixture Model



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?

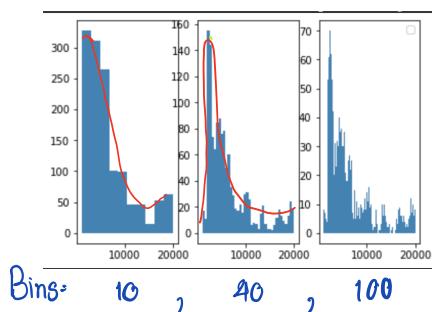
ภารกิจลับๆ ก้าวที่ต้องการ Plot histogram โดยใช้ชี้สีตามที่กำหนดให้

ใน Feature Age



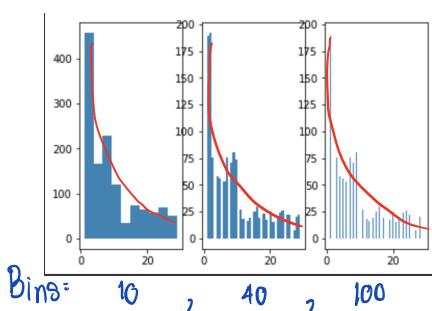
ปรับสี Bins ที่เลือกสุดท้าย 40 นี่จะมากถ้าเป็น Bins=10
ต่ำไปนิดเด้อคือชี้สีปุ่มน้ำเงินกระโดดตัวเดียวที่กลาง
โดยดูจาก Hist ที่รากนั้นมากที่สุดแล้วมีรูปสามเหลี่ยมห่างจาก
รากเท่านั้นให้ใช้ Bins=40 และถ้าเป็น Bins=100
จะ Hist ที่ไม่ถูก Count เยอะเกินไปหรือซึ่งต่างจาก Bins=40.

ใน Feature MonthlyIncome



ปรับสี Bins ที่เลือกสุดท้าย 40 นี่จะมากถ้าเป็น Bins=10
จะบ่งบอกว่าการกระจายตัวไม่ต่อเนื่องกันก่อให้เกิดความไม่สม่ำเสมอ
หากหน่วยการ分布อย่างต่อเนื่องกันต่อๆ ไปถ้าเป็น Bins=100
จะบ่งบอกว่า Histogram ที่เกินไปโดยดูจาก distribution นี้ทาง
Histogram ที่คล้ายเป็นภูเขาจะลดลงเรื่อยๆ

ใน Feature Distance From Home .

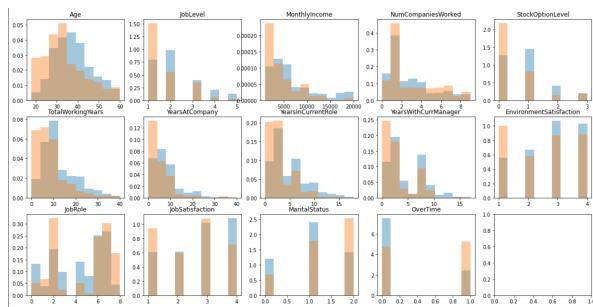


ปรับสี Bins ที่เลือกสุดท้าย 10 นี่จะมากถ้าเป็น Bins=10
บ่งบอกว่าการ分布ที่รัดเข้าไว้กันแน่น Bins ตัวเดียว
โดยการบินบนน้ำทึบเรียงฟังก์ชันที่ไม่ถูกนับซ้ำกันอีกต่อไป.

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 ที่จะนำ回去 Discretized ที่มีลักษณะเป็น Numerical feature คือ:
 เป็น Continuous feature แล้วจะต้องทำการแปลงเป็น Histogram ก่อนแล้ว Bin
 ใหม่ที่ไม่ถูก Count เลย. ทุก Feature ที่เลือกมาหรือไม่เลือก
 จะเป็น Feature ที่สามารถแบ่งเป็น Class & class ได้ค่อนข้างดี เช่น วัย (ex | class1)
 และ เพศ (class2) ห่างกันมากกว่า 5%
 โดยที่เลือกมีดังนี้

```
1 num_feature = ['Age', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'StockOptionLevel', 'TotalWorkingYears',
2 YearsAtCompany', 'YearsInCurrentRole', 'YearsWithCurrManager']
3 cat_feature = ['EnvironmentSatisfaction', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'Overtime']
```



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.

ເນັ້ນສຳພາດ Model histogram ໂດຍໃຊ້ Multinomial Distribution ເຊື່ອງຈາກເຫຼົາສາມາກມອງວ່າ

ມີ Bin ສຶ່ງກລົງທີ່ມີຄະດີ ເປັນກຳນົດ ແລ້ວໂຄກສໍານົງຈະຕັກໃນແຕ່ລະກລົງກີ່ມີໄມ່ເກົ່າກົ່ນ.

ກຳທັນດູ້ທີ່ ຖືສຶ່ງ ຕ້ານອນ Hist ໃນແຕ່ລະ Bin ທີ່

p_i ສຶ່ງໂຄກສໍານົງຈະຕັກໃນ Bin ທີ່;

k ສຶ່ງ ຕ້ານອນ Bin ທັງກຳນົດ.

ທະ: m ສຶ່ງ ດ້ວຍກຳນົດກັ່ງທຸມດຸ.

ເຊື່ອງຈາກສ້າງທາລານ p_i ຕ້ານອນ $k=1$ ຕ່ອງ. ເຮົາຈະກວານ p_i ປັບອອກທີ່ເກີນທັນກຳທີ່ໄປໆຢູ່ p_1 .

\therefore ແລະລັງຈາກ p_1 ຊັ້ນກົງ p_i ທີ່ແລ້ວໜີ $p_i = 1 - \sum_{i=2}^k p_i$

ສັງເກົ່າສາມາກຮະນຸ MLE ໄດ້ຕົກນີ້ $L = P(n_1, n_2, \dots, n_k | p_1, p_2, \dots, p_k)$

$= P(n_1, n_2, \dots, n_k | p_1, p_2, \dots, p_k)$.

ກັນນິ້ນເຮົາ: ຖ້າ $\underset{p_1, p_2, \dots, p_k}{\operatorname{argmax}} P(n_1, n_2, \dots, n_k | p_1, p_2, \dots, p_k)$

$$= \underset{p_1, p_2, \dots, p_k}{\operatorname{argmax}} \frac{m!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

$$= \underset{p_1, p_2, \dots, p_k}{\operatorname{argmax}} \log \frac{m!}{n_1! n_2! \dots n_k!} + n_1 \log p_1 + n_2 \log p_2 + \dots + n_k \log p_k.$$

$$\text{ພິຈາລະນາກີ່ } p_i \quad \frac{\partial L}{\partial p_i} \cdot \frac{\partial}{\partial p_i} \log \frac{m!}{n_1! n_2! \dots n_k!} + n_1 \log (1 - \sum_{i=2}^k p_i) + \dots + n_i \log p_i + \dots + n_k \log p_k$$

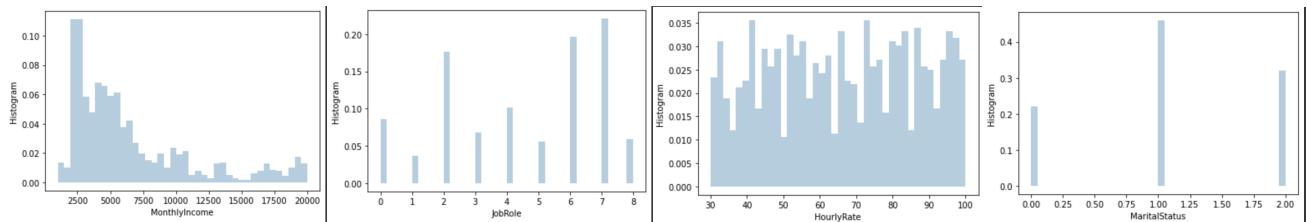
$$0 = \frac{n_1}{1 - \sum_{i=2}^k p_i} + \frac{n_i}{p_i}$$

$$\frac{n_1}{1 - \sum_{i=2}^k p_i} = \frac{n_i}{p_i} \Rightarrow \frac{n_1}{p_1} = \frac{n_i}{p_i} + \dots + \frac{n_k}{p_k}$$

$$n_i = t p_i \quad \text{ນາ: } n_1 = t p_1$$

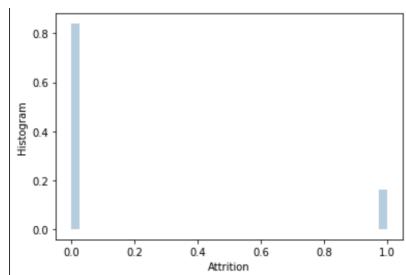
$$\therefore \frac{n_i}{p_i} = m$$

$$\frac{n_1 + \dots + n_k}{p_1 + \dots + p_k} = t = \frac{m}{1} \quad \text{ແລດວາກ່າວ } p_i = \frac{n_i}{\sum n_i} \quad \text{ຈະກຳທີ່ລືກທີ່ likelihood ກົ່ານັກກົ່າສົດ.}$$



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

Binomial



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

- กำหนด $P(x_i|attrition)$ ที่เท่ากับ 0 ถ้า x_i ค่าเสียหาย ก่าหนึ่ง.
- กรณีนักการ Model ข้อมูลจาก Histogram ไม่เป็นการ Model ตาม Distribution ดั้งเดิม ก็ต้องปรับเปลี่ยน.

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.

Implementation อยู่ใน Code ข้อ T11.

```

1 feature = num_feature+cat_feature
2 model = Naive_bayes_hist_classifier(40)
3 model.fit(X_train[feature], y_train)

1 Metric.compute(X_test[feature], y_test, model, verbal = True)
primative [6, 8, 119, 14] TP, FP, TN, FN
accuracy 0.8503401360544217
precision 0.42857142857142855
recall 0.3
f1 0.3529411764705882

```

```
scipy.stats.norm(mean, std).pdf(feature_value)
```

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

```
1 feature = num_feature+cat_feature
2 model = Naive_bayes_normal_classifier()
3 model.fit(X_train[feature], y_train)

1 Meric.compute(X_test[feature], y_test, model, verbal = True)

primitive [4, 0, 127, 16] TP,FP,TN,FN
accuracy 0.891156462585034
precision 1.0
recall 0.2
f1 0.33333333333333337
```

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.

```
1 model = Random_guess_classifier()
2 model.predict(X_train[feature])

array([0, 0, 0, ..., 1, 1, 1])

1 Meric.compute(X_test[feature], y_test, model, verbal=True)

primitive [10, 65, 62, 10] TP,FP,TN,FN
accuracy 0.4897959183673469
precision 0.13333333333333333
recall 0.5
f1 0.2105263157894737
```

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.

```
1 model = Majority_class_classifier()

1 Meric.compute(X_test[feature], y_test, model, verbal=True)

primitive [0, 0, 127, 20] TP,FP,TN,FN
accuracy 0.8639455782312925
precision nan
recall 0.0
f1 nan
```

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

Baseline ទី១ T13 នៃសរុបការប្រើប្រាស់នៅ Model នេះនៅក្នុងការតាមរយៈមុខ
Test set ដែលវានៅអីនេះតែតាមរាងនៅមុននេះនៅក្នុង F1 score នៅក្នុង Model
មួយនៅមួយនៅការណា . នៅក្នុង Baseline ទី១ T14 ដែលវាបានត្រូវជូន
Accuracy នៅក្នុងតាមរយៈមុខនៅក្នុង Imbalanced data នៅក្នុង Class ០ ដែលត្រូវ
ចូលរួមនៅក្នុង Accuracy នៃក្នុងក្នុង Baseline ទី២។ ការណា៖ ប្រើប្រាស់នៅក្នុងលទ្ធផល

T16. Use the following threshold values

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

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

```
1 max_acc = 0
2 pos_F1 = -1
3 max_F1 = 0
4 pos_F1 = -1

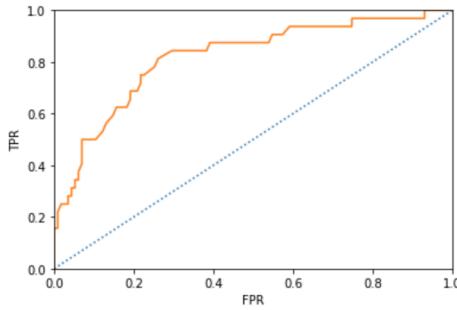
5 for i in t:
6     output = Meric.compute(X_test[feature], y_test, model, verbal=False, threshold=i)
7     if output['accuracy'] > max_acc:
8         max_acc = output['accuracy']
9         pos_acc = i
10    if output['F1'] > max_F1:
11        max_F1 = output['F1']
12        pos_F1 = i
13
14 max_acc, pos_acc, max_F1, pos_F1

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in double_scalars
app.launch_new_instance()
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:14: RuntimeWarning: invalid value encountered in long_scalars
0.8163265306122449,
1.4999999999942504,
0.7687074829931972
-0.9000000000056332)
```

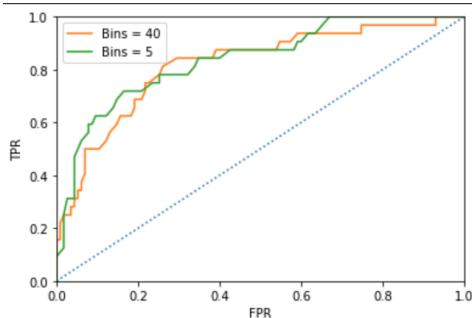
```
1 Meric.compute(X_test[feature], y_test, model, verbal=True, threshold=1.4999999999942304)
2 primitive [5, 4, 115, 23]
3 accuracy 0.8163265306122449
4 precision 0.5555555555555556
5 recall 0.7857142857142858
6 f1 0.2702702702702703

1 Meric.compute(X_test[feature], y_test, model, verbal=True, threshold=-0.9000000000056332)
2 primitive [12, 18, 101, 16]
3 accuracy 0.7687074829931972
4 precision 0.4
5 recall 0.42857142857142855
6 f1 0.4137931034482759
```

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.



ROC curve ນັບກົງສອງແກ່ມາດຈະແນນຕີນັ້ນັ້ນ ກາຮສຽງວ່າ
ວິທີທີ່ທີ່ກ່າວ່າດີນີ້ມີໂດຍກໍ່ນຳການທັງກວດໃນປະເລືດ
ຍົດໄມ້ເຕີລີໃນ Precision ແລ້ວ Recall ແຕ່ກ່າວສຶກສົກການ
ຂອງກັ້ນສະຫຼຸບຕົວລະ Model ຖ້າເສັ່ນກີ່ນີ້ມີຄວາມ.

Done!

T19. Submit your code (.py or .ipynb) on mycourseville.

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.

```
1 accuracies = []
2
3 for i in range(10):
4     X_train, X_test, y_train, y_test = train_test_split(X, y)
5     model = Naive_bayes_hist_classifier(40)
6     model.fit(X_train[feature], y_train)
7     output = Meric.compute(X_test[feature], y_test, model, verbal=False, threshold=-2.45)
8     accuracies.append(output['accuracy'])
9
10 accuracies = np.array(accuracies)
11 accuracies.mean(), accuracies.std()**2, accuracies
```

```
(0.70136054217687,
 0.0008649173955296407,
 array([0.69387755, 0.70068027, 0.7414966 , 0.69387755, 0.68707483,
       0.69387755, 0.71428571, 0.7414966 , 0.63265306, 0.71428571]))
```