Pattern Recognition
Pisit Wajanasara 5931042721

T1.

T1. Find α that maximizes $p(y_2, y_1, y_0 | \alpha)$

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

Since $y_2 = \alpha y_1 + \omega_1 \Rightarrow p(y_2 | y_1, \alpha) \sim N(\alpha y_1, \sigma^2)$

$y_1 = \alpha y_0 + \omega_0 \Rightarrow p(y_1 | y_0, \alpha) \sim N(\alpha y_0, \sigma^2)$

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

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

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

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

Set to 0 ;
$$0 = (y_2 - \alpha y_1) y_1 + (y_1 - \alpha y_0)(y_0)$$
$$0 = y_2 y_1 - \alpha y_1^2 + y_1 y_0 - \alpha y_0^2$$
$$\alpha (y_1^2 + y_0^2) = y_2 y_1 + y_1 y_0$$

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

OT1.

OT1. Find α that maximize $p(y_{N+1}, y_N, \dots, y_0 | \alpha)$

Since we can write

$y_{N+1} = \alpha y_N + \omega_N$ and $\omega_N \sim N(0, \sigma^2)$

Thus $p(y_{N+1} | y_N, \alpha) \sim N(\alpha y_N, \sigma^2)$

$$p(y_{N+1}, y_N, \dots, y_0 | \alpha) = p(y_{N+1} | y_N, y_{N-1}, \dots, y_0, \alpha)\, p(y_N, y_{N-1}, \dots, y_0 | \alpha)$$

$$= \prod_{i=0}^{N} p(y_{i+1} | y_i, \alpha) \times p(y_0)$$

$$\log p(y_{N+1}, y_N, \dots, y_0 | \alpha) = \sum_{i=0}^{N} \log p(y_{i+1} | y_i, \alpha) + \log p(y_0)$$

$$= \sum_{i=0}^{N} \left[\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_{i+1} - \alpha y_i)^2}{2\sigma^2} \right] + \left[\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{y_0^2}{2\sigma^2} \right]$$

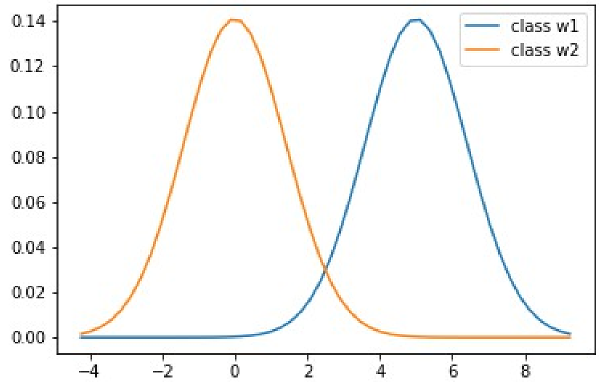
$$\frac{\partial}{\partial \alpha} \log p(y_{N+1}, y_N, \dots, y_0 | \alpha) = \sum_{i=0}^{N} \left[\frac{-2(y_{i+1} - \alpha y_i)(-y_i)}{2\sigma^2} \right]$$

set to 0; $\quad 0 = \sum_{i=0}^{N} \left[y_{i+1} y_i - \alpha y_i^2 \right]$

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

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

T2.



Decision boundary is 2.5

T3.

Decision boundary is 1.945

The boundary shift toward the sad cat distribution.

OT2.

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

$P(x|w_2) = N(\mu_2, \sigma^2)$

$p(w_1) = p(w_2) = 0.5$

The decision boundary is at the point that $p(w_1|x) = p(w_2|x)$

$$\frac{p(x|w_1)p(w_1)}{p(x)} = \frac{p(x|w_2)p(w_2)}{p(x)}$$

$$p(x|w_1)\,p(w_1) = p(x|w_2)\,p(w_2)$$

$$p(x|w_1) = p(x|w_2)$$

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

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

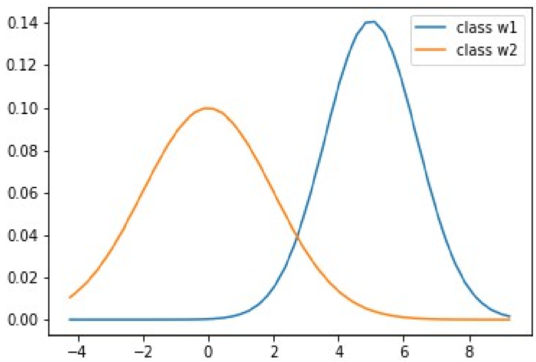
$$x^2 - 2x\mu_1 + \mu_1^2 = x^2 - 2x\mu_2 + \mu_2^2$$

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

$$2x = \mu_2 + \mu_1$$

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

$\therefore$ Therefore decision boundary is at $x = \frac{\mu_2 + \mu_1}{2}$



Decision boundary is 2.7355

T4. From observed histogram, we can use Gaussian to estimate the histogram but not for all column, because some column have a shape that is not Gaussian like. If we use Gaussian to estimate them, they might provide bad result because of this estimation. We can use Gaussian Mixture Model to estimate the histogram because it can handle a case that histogram is not just like a one Gaussian Distribution.

T5. 859 bins have zero count. This is not a good discretization because if we have too many zero bin count this will make we hard to estimate the distribution of the histogram. Moreover, from observation, many bins that have zero count are between other bin that have non-zero count. This show that the distribution we see from histogram is difficultly determined.

T6. (*** See histogram in ipynb file ****)

1) For age, bin size = 10 is most sensible, as you can see in the histogram, this can make the estimation using Gaussian more sensible compared to bin size = 40 and 100.
2) For monthly income, bin size = 40 is most sensible. For the histogram of each bin size we can observe the same distribution shape, so we will also use other data for consideration. We know that min, max and standard deviation of the data are 1009, 19859 and 4738.803810 respectively (You may observe different number in the iPython notebook file because the data will be shuffled each time we run the code.), so bin size = 10 make the binning result too rough for discretization and bin size = 100 make the binning result too detailed with some bins that have zero amount of data.
3) For distance from home, bin size = 10 is most sensible. For other bin size(40, 100), there are many bins that has zero amount of data in it but not in the case which bin size = 10.

T7. The criteria for discretization is

1) The range of data is large (max – min > 30)
2) The data is not categorical data
3) The data is continuous

The columns that follow the criteria are DailyRate, EmployeeCount, HourlyRate, MonthlyRate, TotalWorkingYears and YearsAtCompany.
(*** See discretized features in ipynb file ***)