EECE 5644 Assignment 2 Jingcheng Wang

Repository: https://github.com/tomwang777/2025-Fall-EECE-5644-Machine-Learning/tree/main/Assignment%202

Question 1

N_train = [50, 500, 5000];
N_val = 10000;

Part 1

Min-P(error) (Bayes) ≈ 0.2710

Question 1.

Probability density function. $p(x) = P(L=0)\,p(x|L=0) + P(L=1)\,p(x|L=1)$ $P(L=0)=0.6$. $P(L=1)=0.4$

$= 0.6\,p(x|L=0) + 0.4\,p(x|L=1)$

Class class-conditional pdfs. $p(x|L=0) = 0.5g(x|m_{01},C_{01}) + 0.5g(x|m_{02},C_{02})$

$p(x|L=1) = 0.5g(x|m_{11},C_{11}) + 0.5g(x|m_{12},C_{12})$

Part 1. Bayes classifier. ✓

Bayes Decision Rule. Decide $L=1$ if $P(L=1|x) > P(L=0|x)$.

$$P(L=i|x) = \frac{P(L=i)p(x|L=i)}{p(x)}$$

∵ Let $\Lambda(x) = \frac{p(x|L=1)P(L=1)}{p(x|L=0)P(L=0)}$.

∴ $\Lambda(x) > 1$. $L=1$

$\Lambda(x) < 1$. $L=0$.

That is. $\Lambda(x) = \frac{0.4[0.5g(x|m_{11},C_{11}) + 0.5g(x|m_{12},C_{12})]}{0.6[0.5g(x|m_{01},C_{01}) + 0.5g(x|m_{02},C_{02})]}$.

Decision Boundary.

$p(x|L=1)P(L=1) = p(x|L=0)P(L=0)$.

$\Lambda(x) = 1$.

ROC Curve. $L=1 \rightarrow \Lambda(x) > \tau$.

Min-P(error). $\hat{L}(x) = \underset{i\in\{0,1\}}{\arg\max}\, P(L=i|x)$

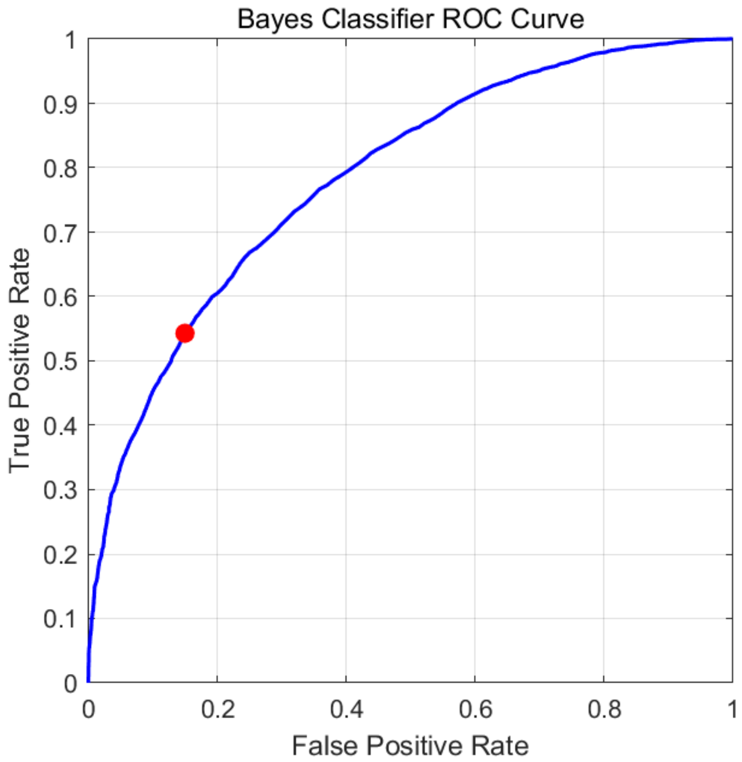
$\hat{L}(x) = \underset{i\in\{0,1\}}{\arg\max}\, P(L=i)p(x|L=i)$.

$$P(L=1|x) = \frac{P(L=1)p(x|L=1)}{p(x)} = \frac{P(L=1)p(x|L=1)}{P(L=0)p(x|L=0) + P(L=1)p(x|L=1)} > 0.5.$$

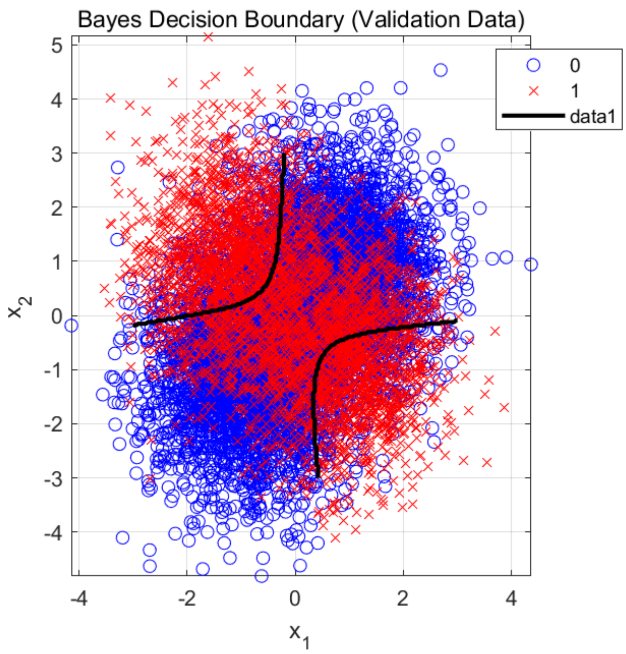
Given validation set. $\{(x_n, L_n)\}_{n=1}^{N}$.

$$\hat{P}_{error} = \frac{1}{N}\sum_{n=1}^{N} \{\hat{L}_n \neq L_n\}.$$

$\simeq 0.2710$.



Bayes Classifier ROC Curve
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1



Bayes Decision Boundary (Validation Data)
0
1
data1
x_2
x_1
5
4
3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4

Part 2

No.

Date

Part 2.

$\sigma(t) = \frac{1}{1+e^{-t}}$.

Logistic-linear model. $h(x;w) = \frac{1}{1+e^{-w^T z(x)}}$ $\quad z(x) = [1, x_1, x_2]^T$.

$= \sigma(w_0 + w_1 x_1 + w_2 x_2)$.

Goal function. (Negative log-likelihood). $J(w) = -\sum_{n=1}^{N} [y_n \log h(x_n; w) + (1-y_n)\log(1-h(x_n;w))]$.

Optimizing parameter $w^* = \arg\min_w J(w)$.

Logistic-quadratic model. $h(x;w) = \frac{1}{1+e^{-w^T z(x)}}$ $\quad z(x) = [1, x_1, x_2, x_1^2, x_1x_2, x_2^2]^T$.

(The same as above.)

$= \sigma(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2)$.

Training size = 50

Linear logistic error = 0.461 | Quadratic logistic error = 0.373

Training size = 500

Linear logistic error = 0.358 | Quadratic logistic error = 0.281

Training size = 5000

Linear logistic error = 0.389 | Quadratic logistic error = 0.342

Validation error (linear) = 0.389

Discussion

As the number of training samples increases, the classifier training becomes better and more precise and the minimum error probability decreases. The classification effect of the logistic quadratic model is better than that of the logistic linear model because it focuses on more quadratic combinations. Compared with the theoretical Bayes optimal classifier in Part 1, they basically achieve similar classification results, indicating that the model of this experiment has almost achieved the best theoretically feasible case.

* I found that the generalization error rate does not decrease monotonically with the number of samples. This may be due to random sampling or insufficient regularization of the model. After repeated tests, I still cannot get a monotonically decreasing result.

Question 2

Question 2.

Scalar-real y : 2-dimensional real vector $x = [x_1, x_2]^T$.

$$y = c(x, w) + r.$$

$$c(x, w) = w[1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3]^T.$$

$$r \sim \mathcal{N}(0, \sigma_v^2).$$

$$p(y|x, w) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{1}{2\sigma_v^2}(y - w^T c(x))^2}$$

ML Estimator $\hat{w}_{ML}$. Maximize $\ln p(D|w)$.

For N i.i.d. $D = \{x_n, y_n\}_{n=1}^{N}$.

$$\ln p(D|w) = \sum_{n=1}^{N} \ln p(y_n|x_n, w)$$

$$\ln p(D|w) = \sum_{n=1}^{N} \left[-\frac{1}{2}\ln(2\pi\sigma_v^2) - \frac{1}{2\sigma_v^2}(y_n - w^T c(x_n))^2 \right].$$

$$\hat{w}_{ML} = \arg\min_w \sum_{n=1}^{N} (y_n - w^T c(x_n))^2.$$

Let Z become $N \times d$ Design Matrix

y become $N \times 1$ output vector.

$$\hat{w}_{ML} = \arg\min_w \|y - Zw\|^2. \quad \nabla_w \|y - Zw\|^2 = 0.$$

$$= \underbrace{(Z^T Z)^{-1} Z^T y}.$$

MAP Estimator $\hat{w}_{MAP}$. Maximize $p(w|D) = p(D|w) \times p(w)$.

$$p(w|D) \propto p(D|w)\, p(w).$$

$$p(w) = \mathcal{N}(0, rI)$$

$$\ln p(w) = -\frac{d}{2}\ln(2\pi r) - \frac{1}{2r} w^T w.$$

Goal function. $\hat{w}_{MAP} = \arg\min_w \{ -\ln p(D|w) - \ln p(w) \}$.

Remove the constants: $= \arg\min_w \left\{ \sum_{n=1}^{N} \frac{1}{2\sigma_v^2}(y_n - w^T c(x_n))^2 + \frac{1}{2r} w^T w \right\}$.

$\lambda = \frac{\sigma_v^2}{r}$.

$$= \arg\min_w \{ \|y - Zw\|^2 + \lambda \|w\|^2 \}$$

$$= \underbrace{(Z^T Z + \lambda I)^{-1} Z^T y}.$$

ML Model Validation MSE: 4.1891

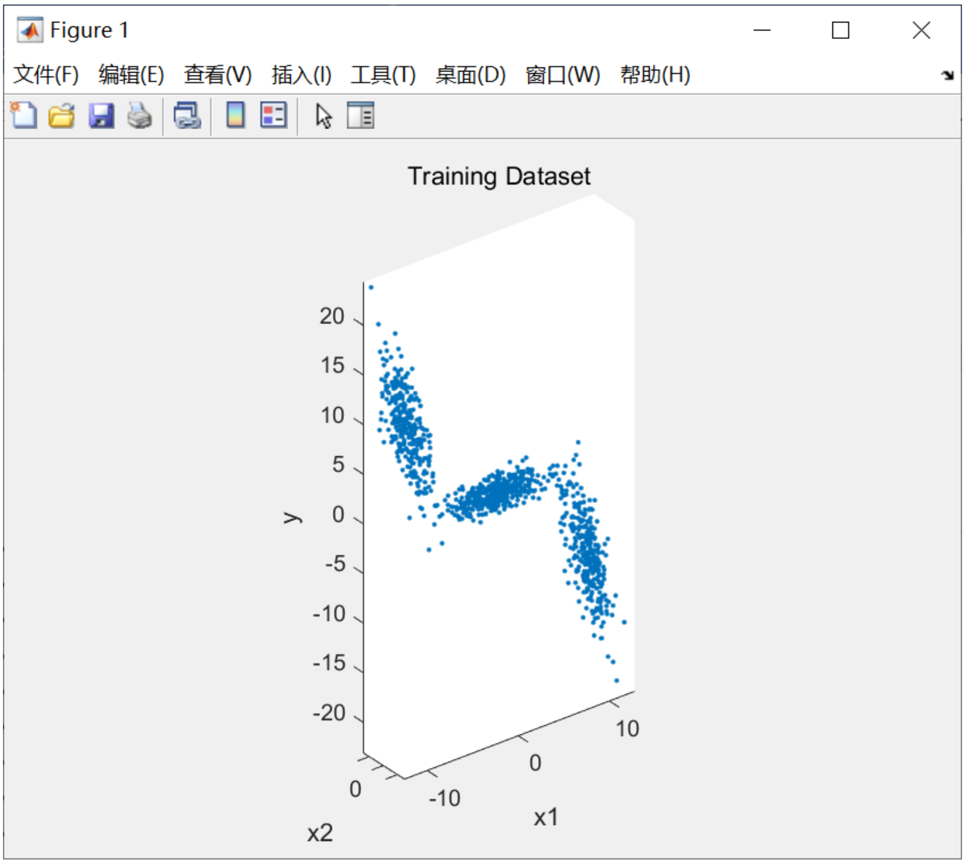
MAP Model Validation MSE: 4.1963

MAP Model Validation MSE: 4.1939

MAP Model Validation MSE: 4.1913

MAP Model Validation MSE: 4.1888

MAP Model Validation MSE: 4.1869
MAP Model Validation MSE: 4.1857
MAP Model Validation MSE: 4.1850
MAP Model Validation MSE: 4.1848
MAP Model Validation MSE: 4.1849
MAP Model Validation MSE: 4.1852
MAP Model Validation MSE: 4.1857
MAP Model Validation MSE: 4.1862
MAP Model Validation MSE: 4.1867
MAP Model Validation MSE: 4.1872
MAP Model Validation MSE: 4.1876
MAP Model Validation MSE: 4.1880
MAP Model Validation MSE: 4.1883
MAP Model Validation MSE: 4.1885
MAP Model Validation MSE: 4.1887
MAP Model Validation MSE: 4.1888
MAP Model Validation MSE: 4.1889
MAP Model Validation MSE: 4.1889
MAP Model Validation MSE: 4.1890
MAP Model Validation MSE: 4.1890
MAP Model Validation MSE: 4.1890
MAP Model Validation MSE: 4.1890
MAP Model Validation MSE: 4.1891



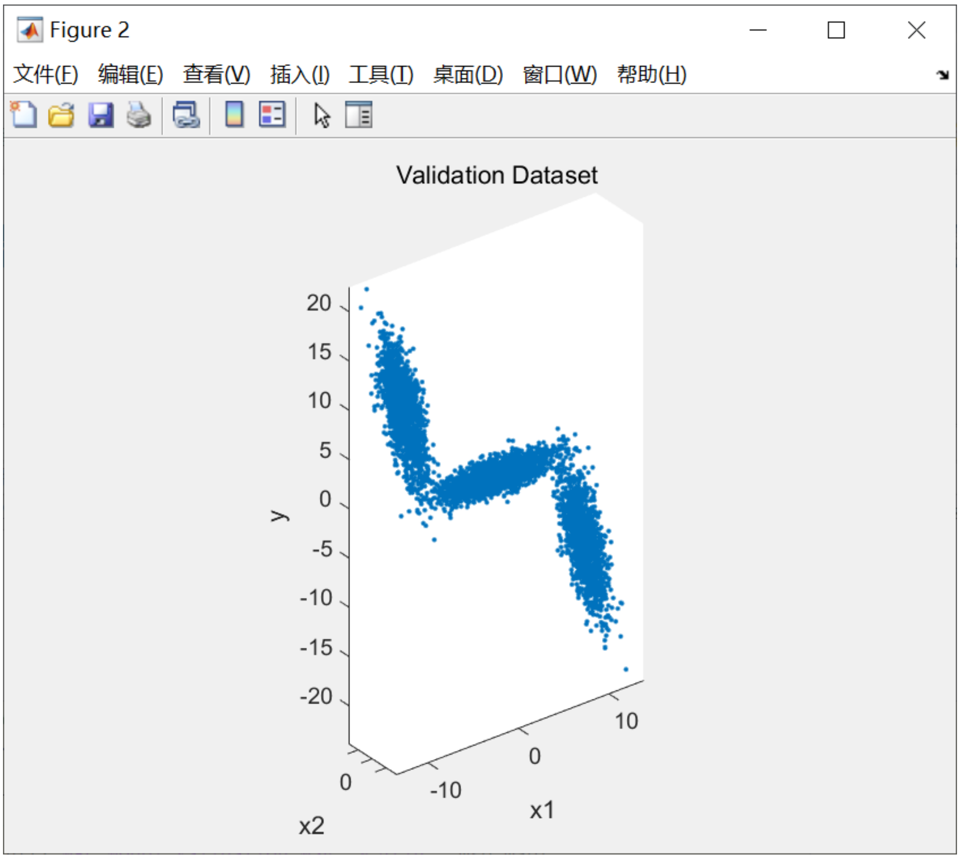


Figure 2
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
Validation Dataset
y
x1
x2

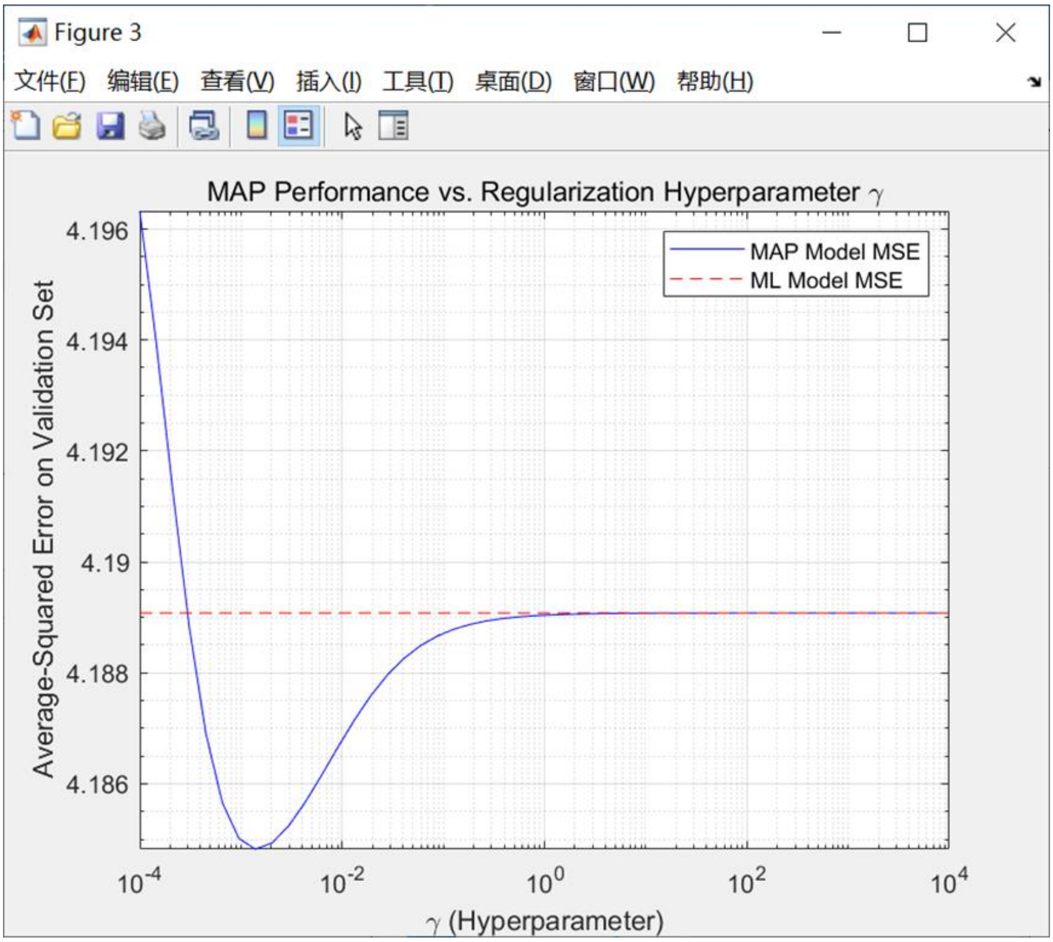


Figure 3
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
MAP Performance vs. Regularization Hyperparameter γ
MAP Model MSE
ML Model MSE
Average-Squared Error on Validation Set
γ (Hyperparameter)

As γ changes, the mean squared error of the MAP training model on the validation set first decreases to a lowest point, then increases, and finally remains constant. This indicates that after the regularization reaches a certain level, the performance of the model will no longer be affected by the regularization parameter. However, the mean squared error of the ML training model remains unchanged and is equal to the final value of the MAP model. This indicates that the ML model is not affected by regularization, and regularization of the MAP model is intended to give it the same performance as the ML model.

Question 3

No.

Date

Question 3.

The true position of a vehicle: $X_T = [x_T, y_T]^T$.

$$X_{MAP} = \arg\max_x p(x|r). \quad r = [r_1, \dots, r_K]^T \text{ for evaluating distance.}$$

According to Bayes Rule:

$$p(x|r) \propto p(r|x)p(x).$$

$$X_{MAP} = \arg\max_x \{\ln(p(r|x)) + \ln(p(x))\}.$$

$$X_{MAP} = \arg\min_x \{-\ln(p(r|x)) - \ln(p(x))\}.$$

$$r_i = d_{T_i} + n_i. \quad d_{T_i} = ||x - X_i||.$$

$n_i \sim N(0, \sigma_i^2)$ Gaussian Noise with 0 mean.

$$r_i \sim N(d_{T_i}, \sigma_i^2)$$

$$p(r_i|x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(r_i - d_{T_i})^2}$$

$$p(r|x) = \prod_{i=1}^{K} p(r_i|x).$$

$$-\ln(p(r|x)) = -\sum_{i=1}^{K} \ln(p(r_i|x)).$$

$$-\ln(p(r|x)) = -\sum_{i=1}^{K} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \frac{1}{2\sigma_i^2}(r_i - d_{T_i})^2\right].$$

Prior function $\ln(p(x))$.

$$p(x) = (2\pi\sigma_x\sigma_y)^{-1} e^{-\frac{1}{2}[x\ y]\begin{bmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x \\ y\end{bmatrix}}$$

$$-\ln(p(x)) = -\ln((2\pi\sigma_x\sigma_y)^{-1}) + \frac{1}{2}[x\ y]\begin{bmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x \\ y\end{bmatrix}$$

MAP Goal function $\underbrace{g(x)}_{\text{minimize}} = -\ln(p(r|x)) - \ln(p(x)).$

Remove the ~~constraints~~ constants.

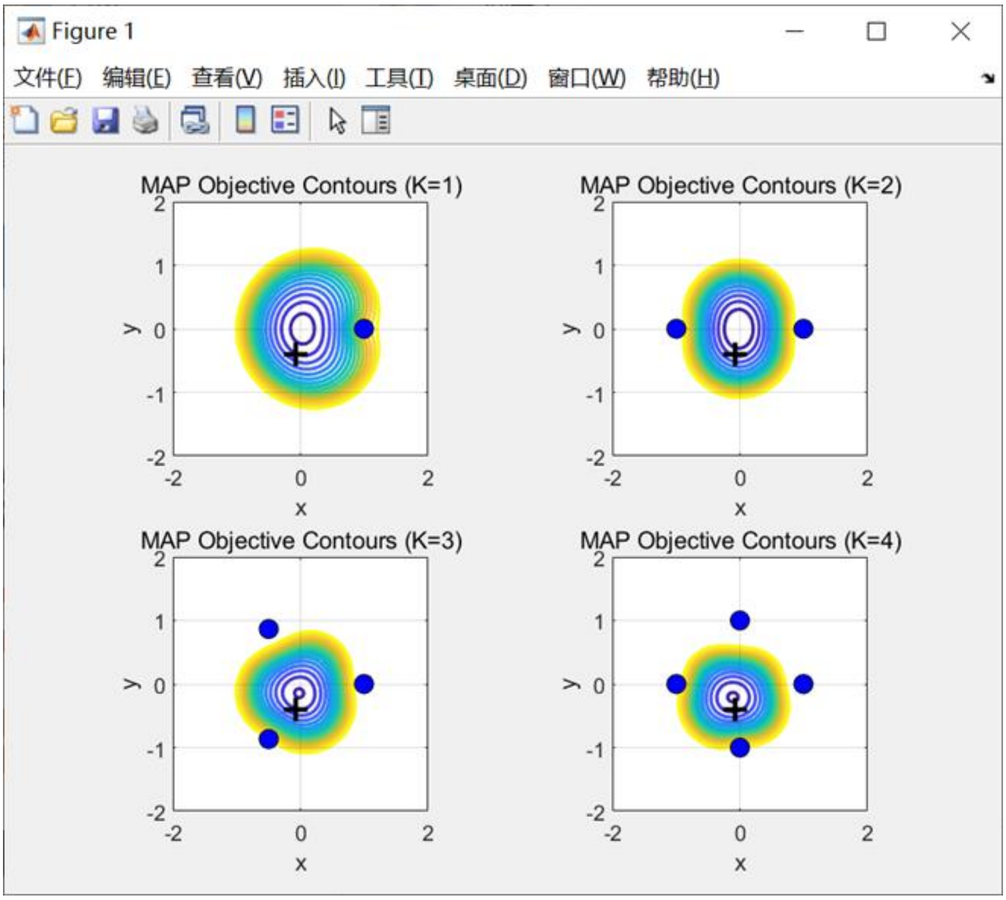
$$X_{MAP} = \arg\min_x g_{simplified}(x) = \arg\min_x \left\{\sum_{i=1}^{K} \frac{(r_i - ||x - X_i||)^2}{\sigma_i^2} + [x\ y]\begin{bmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x \\ y\end{bmatrix}\right\}$$

For Evaluation. X_T.

X_i. K (number).

$\sigma_i^2 = 0.09$

$\sigma_x = \sigma_y \approx 0.25$. $\quad r_i = \underbrace{||X_T - X_i|| + n_i}_{\geq 0}. \quad n_i \sim N(0, 0.09).$



Regardless of the K value, the MAP estimate and the true position will have a certain estimation error and will not completely coincide. However, as the K value increases, the MAP estimate is getting closer to the true position. As the contour lines become closer and closer, the variance decreases, the estimated posterior probability distribution becomes sharper, and the MAP estimate becomes more certain.

Question 4

Question 4.

The estimation risk $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$.

when classified into w_i, $i \in \{1, \dots, c\}$:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$$

$$= \underbrace{\lambda(\alpha_i|w_i)}_{\substack{0 \\ j=i}} P(w_i|x) + \sum_{j\neq i, j=1}^{c} \underbrace{\lambda(\alpha_i|w_j)}_{\substack{\lambda_s \\ j\neq i}} P(w_j|x)$$

$\sum_{j=1}^{c} P(w_j|x) = 1$.
$$= \lambda_s \sum_{j\neq i, j=1}^{c} P(w_j|x)$$

$\sum_{j\neq i, j=1}^{c} P(w_j|x) = 1 - P(w_i|x)$.
$$= \lambda_s (1 - P(w_i|x)).$$

when rejection happens. α_{c+1}.

$$R(\alpha_{c+1}|x) = \sum_{j=1}^{c} \underbrace{\lambda(\alpha_{c+1}|w_j)}_{\lambda_r} P(w_j|x)$$

$$= \lambda_r \sum_{j=1}^{c} P(w_j|x) \qquad \sum_{j=1}^{c} P(w_j|x) = 1.$$

$$= \lambda_r.$$

$$\alpha = \underset{i\in\{1,\dots,c\}}{\arg\min}\, R(\alpha_i|x) \qquad i^* = \underset{j\in\{1,\dots,c\}}{\arg\max}\, P(w_j|x).$$

$$= \underset{i\in\{1,\dots,c\}}{\arg\min}\, \lambda_s(1-P(w_i|x)). \quad \lambda_s > 0.$$

$R(\alpha_i|x) \leq R(\alpha_{c+1}|x)$ then classifies.

$\lambda_s(1-P(w_i|x)) \leq \lambda_r$.

$1 - P(w_i|x) \leq \frac{\lambda_r}{\lambda_s}$.

$P(w_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$. otherwise reject.

① $\lambda_r = 0$.

$P(w_i|x) \geq 1$.

~~MAP Rule~~.

Always reject.

except $P(w_i|x) = 1$.

② $\frac{\lambda_r}{\lambda_s} > 1$. $\lambda_r > \lambda_s$.

$P(w_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s} < 0$.

MAP Rule.

Always classify.

never reject.

Question 5

Question 5.

Likelihood function. $p(D|\Theta) = \prod_{n=1}^{N} p(z_n)$

$p(z) = \prod_{k=1}^{K} \theta_k^{z_k}$.

only when $z_k = 1$ θ_k is considered.

$$= \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \theta_k^{z_{nk}} \right) = \prod_{k=1}^{K} \theta_k^{N_k}.$$

$$N_k = \sum_{n=1}^{N} z_{nk} \quad z_{nk} = 1.$$

For mL estimator. $\ln(p(D|\Theta)) = \sum_{k=1}^{K} N_k \ln(\theta_k)$.

Introduce Lagrange multiplier λ.

$$L(\Theta, \lambda) = \sum_{k=1}^{K} N_k \ln(\theta_k) + \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right)$$

$$\frac{\partial L}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0.$$

$$\theta_k = \frac{N_k}{\lambda}.$$

$$\sum_{k=1}^{K} \theta_k = 1.$$

$$\frac{1}{\lambda} \sum_{k=1}^{K} N_k = 1.$$

$$\underbrace{}_{N}$$

$$N = \lambda.$$

$$\therefore \hat{\theta}_{k,mL} = \frac{N_k}{N}.$$

For mAP estimator. Dirichlet distribution.

$$p(\Theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}. \quad \alpha = [\alpha_1, \dots, \alpha_k]^T$$

According to Bayes Rule.

$$p(\Theta|D) \propto p(D|\Theta) p(\Theta|\alpha) \qquad B(\alpha) \text{ is a constant.}$$

$$p(\Theta|D) \propto \left(\prod_{k=1}^{K} \theta_k^{N_k}\right)\left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right)$$

$$p(\Theta|D) \propto \prod_{k=1}^{k} \theta_k^{N_k + \alpha_k - 1}.$$

$$\ln p(\Theta|D) \propto \sum_{k=1}^{K} (N_k + \alpha_k - 1) \ln(\theta_k).$$

Just like mL. use $N_k + \alpha_k - 1$ to replace N_k.

we can get

$$\hat{\theta}_{k,mAP} = \frac{N_k + \alpha_k - 1}{\sum_{k=1}^{K} (N_k + \alpha_k - 1)}.$$

Citation

1. Course recording
2. Course notes
3. Course codes provided on Canvas
4. Discussion with classmates
5. Generative AI models