

# STAT 3690 Lecture Note

## Part IX: Classification

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2023/Mar/25 12:54:24

---

### Classification

- Predictive task in which the response takes values across  $K$  discrete categories (i.e., not continuous)
  - Having training data with known class labels
  - Predict one subject's label  $Y$  according to  $p$ -vector  $\mathbf{X}$
  - Binary classification:  $K = 2$
  - E.g.
    - \* Given a scanned handwritten digit, determine what digit was written.
    - \* Predict the region of Italy in which a sample of olive oil was made, according to its chemical composition.

---

### Bayes classifier

- Classification according to posteriors

$$\Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{\ell=1}^K f_{\ell}(\mathbf{x})\pi_{\ell}}, \quad k = 1, \dots, K$$

- $f_k(\mathbf{x})$ : the probability density/mass function of  $\mathbf{X}$  conditioning on Class  $k$
  - $\pi_k = \Pr(Y = k)$ : prior of Class  $k$
- Bayes classifier

$$h(\mathbf{x}) = \arg \max_{k=1, \dots, K} \Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \arg \max_{k=1, \dots, K} f_k(\mathbf{x})\pi_k$$

---

### Linear discriminant analysis (LDA, from the perspective of Bayes classifier)

- Assuming  $f_k(\mathbf{x}) = \text{density of } \text{MVN}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$
- LDA classifier (population version)

$$h(\mathbf{x}) = \arg \max_{k=1, \dots, K} f_k(\mathbf{x})\pi_k = \arg \max_{k=1, \dots, K} \delta_k(\mathbf{x})$$

- Discriminant functions  $\delta_k(\mathbf{x}) = \mathbf{x}^{\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^{\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln \pi_k$ 
    - \* Linear functions with respect to  $\mathbf{x}$

$$\begin{aligned}
& \max_{k=1, \dots, K} f_k(\mathbf{x}) \pi_k \\
&= \max_{k=1, \dots, K} (2\pi)^{-\frac{p}{2}} \left\{ \det(\Sigma_k) \right\}^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right\} \pi_k \\
&= \max_{k=1, \dots, K} \exp \left\{ -\frac{1}{2} (\mathbf{x}^\top \Sigma_k^{-1} \mathbf{x} - 2 \mathbf{x}^\top \Sigma_k^{-1} \mu_k + \mu_k^\top \Sigma_k^{-1} \mu_k) \right\} \pi_k \\
&= \max_{k=1, \dots, K} \exp \left( \mathbf{x}^\top \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma_k^{-1} \mu_k \right) \pi_k \\
&= \max_{k=1, \dots, K} \left( \mathbf{x}^\top \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma_k^{-1} \mu_k + \ln \pi_k \right)
\end{aligned}$$

- LDA classifier (empirical version)
  - Training data:  $\mathbf{x}_i \in \mathbb{R}^p$  and  $y_i \in \{1, \dots, K\}$ ,  $i = 1, \dots, n$ 
    - \*  $n_k$ : the number of training observations in class  $k$ ,  $k = 1, \dots, K$
  - Estimation for  $\mu_k$ ,  $\Sigma$  and  $\pi_k$ 
    - \*  $\hat{\pi}_k = n_k/n$
    - \*  $\hat{\mu}_k = n_k^{-1} \sum_{i=1}^n \mathbf{x}_i \cdot \mathbf{1}(y_i = k)$
    - \*  $\hat{\Sigma} = (n-1)^{-1} \sum_{k=1}^K \sum_{i=1}^n (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^\top \cdot \mathbf{1}(y_i = k)$
  - Empirical LDA classifier

$$\hat{h}(\mathbf{x}) = \arg \max_{k=1, \dots, K} \hat{\delta}_k(\mathbf{x})$$

$$* \hat{\delta}_k(\mathbf{x}) = \mathbf{x}^\top \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^\top \hat{\Sigma}^{-1} \hat{\mu}_k + \ln \hat{\pi}_k$$

- 
- Example 9.1 (Fisher's or Anderson's iris data)
    - 50 flowers from each of 3 species of iris: setosa, versicolor, and virginica.
    - Measurements in centimeters of the variables sepal length and width and petal length and width.
- 

## Quadratic discriminant analysis (QDA, from the perspective of Bayes classifier)

- Assuming  $f_k(\mathbf{x}) = \text{density of } \text{MVN}_p(\mu_k, \Sigma_k)$
- QDA classifier (population version)

$$h(\mathbf{x}) = \arg \max_{k=1, \dots, K} f_k(\mathbf{x}) \pi_k = \arg \max_{k=1, \dots, K} \delta_k(\mathbf{x})$$

- Discriminant functions  $\delta_k(\mathbf{x}) = -\mathbf{x}^\top \Sigma_k^{-1} \mathbf{x} + 2 \mathbf{x}^\top \Sigma_k^{-1} \mu_k - \mu_k^\top \Sigma_k^{-1} \mu_k + 2 \ln \pi_k - \ln \det \Sigma_k$ 
  - \* Quadratic functions with respect to  $\mathbf{x}$

$$\begin{aligned}
& \max_{k=1, \dots, K} f_k(\mathbf{x}) \pi_k \\
&= \max_{k=1, \dots, K} (2\pi)^{-\frac{p}{2}} \left\{ \det(\Sigma_k) \right\}^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x} - \mu_k) \right\} \pi_k \\
&= \max_{k=1, \dots, K} \left\{ \det(\Sigma_k) \right\}^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}^\top \Sigma_k^{-1} \mathbf{x} - 2 \mathbf{x}^\top \Sigma_k^{-1} \mu_k + \mu_k^\top \Sigma_k^{-1} \mu_k) \right\} \pi_k \\
&= \max_{k=1, \dots, K} \left( -\mathbf{x}^\top \Sigma_k^{-1} \mathbf{x} + 2 \mathbf{x}^\top \Sigma_k^{-1} \mu_k - \mu_k^\top \Sigma_k^{-1} \mu_k + 2 \ln \pi_k - \ln \det(\Sigma_k) \right)
\end{aligned}$$

- QDA classifier (empirical version)
  - Training data:  $\mathbf{x}_i \in \mathbb{R}^p$  and  $y_i \in \{1, \dots, K\}$ ,  $i = 1, \dots, n$ 
    - \*  $n_k$ : the number of training observations in class  $k$ ,  $k = 1, \dots, K$
  - Estimation for  $\mu_k$ ,  $\Sigma$  and  $\pi_k$ 
    - \*  $\hat{\pi}_k = n_k/n$
    - \*  $\hat{\mu}_k = n_k^{-1} \sum_{i=1}^n \mathbf{x}_i \cdot \mathbf{1}(y_i = k)$

$$\begin{aligned}
& * \hat{\Sigma}_k = (n_k - 1)^{-1} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^\top \cdot \mathbf{1}(y_i = k) \\
& - \text{Empirical classifier} \\
& \hat{h}(\mathbf{x}) = \arg \max_{k=1, \dots, K} \hat{\delta}_k(\mathbf{x}) \\
& * \hat{\delta}_k(\mathbf{x}) = -\mathbf{x}^\top \hat{\Sigma}_k^{-1} \mathbf{x} + 2\mathbf{x}^\top \hat{\Sigma}_k^{-1} \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_k^\top \hat{\Sigma}_k^{-1} \hat{\boldsymbol{\mu}}_k + 2 \ln \hat{\pi}_k - \ln \det \hat{\Sigma}_k
\end{aligned}$$

- Example 9.2 (iris data, con'd)

## Misclassification/error rate

- Population:  $\Pr(Y \neq h(\mathbf{X}))$ 
  - $h(\cdot)$ : the classifier to be evaluated
- Apparent estimation
  - Implementation
    1. Fit a classifier according to training data
    2. Apply the fitted classifier to training data as well
    3. Estimate the error rate by the misclassification proportion
  - Comments
    - \* Training and testing with identical data points
    - \* Severe underestimation likely
- Parametric estimation
  - Implementation
    1. Express  $\Pr(Y \neq h(\mathbf{X}))$  in terms of unknown parameters
    2. Plug in estimates of unknown parameters
  - Comment
    - \* Able to derive the analytical form of  $\Pr(Y \neq h(\mathbf{X}))$  in rare cases
    - \* Underestimation likely

For  $K=2$  and  $\mathbf{X}|Y=k \sim MVN(\boldsymbol{\mu}_k, \Sigma)$ ,  
 the error rate of LDA classifier  $h(\mathbf{X})$

$$\begin{aligned}
& = \Pr(Y \neq h(\mathbf{X})) \\
& = \Pr(Y=1, h(\mathbf{X})=2) + \Pr(Y=2, h(\mathbf{X})=1) \\
& = \Pr(Y=1, \hat{\delta}_1(\mathbf{X}) < \hat{\delta}_2(\mathbf{X})) + \Pr(Y=2, \hat{\delta}_1(\mathbf{X}) > \hat{\delta}_2(\mathbf{X})) \\
& = \pi_1 \Pr(\hat{\delta}_1(\mathbf{X}) < \hat{\delta}_2(\mathbf{X}) | Y=1) + \pi_2 \Pr(\hat{\delta}_1(\mathbf{X}) > \hat{\delta}_2(\mathbf{X}) | Y=2)
\end{aligned}$$

Let  $U = \hat{\delta}_1(\mathbf{X}) - \hat{\delta}_2(\mathbf{X}) = \mathbf{X}^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^\top \Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \Sigma^{-1}\boldsymbol{\mu}_2 + \ln(\pi_1/\pi_2)$ , then

$$\begin{aligned}
E(U|Y=1) & = \boldsymbol{\mu}_1^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^\top \Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \Sigma^{-1}\boldsymbol{\mu}_2 + \ln(\pi_1/\pi_2) \\
& = \frac{1}{2}\boldsymbol{\mu}_1^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^\top \Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \Sigma^{-1}\boldsymbol{\mu}_2 + \ln(\pi_1/\pi_2) \\
& = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln(\pi_1/\pi_2) \\
E(U|Y=2) & = -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln(\pi_2/\pi_1) \\
\text{var}(U|Y=1) & = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
\text{var}(U|Y=2) & = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{aligned}$$

Further,

$$\begin{aligned}
\frac{U - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \ln(\pi_1/\pi_2)}{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}} \bigg| Y=1 & \sim N(0, 1) \\
\frac{U + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \ln(\pi_2/\pi_1)}{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}} \bigg| Y=2 & \sim N(0, 1)
\end{aligned}$$

So,  $\Pr(Y \neq h(\mathbf{X}))$

$$\begin{aligned}
& = \pi_1 \Pr(U < 0 | Y=1) + \pi_2 \Pr(U > 0 | Y=2) \\
& = \pi_1 \Phi\left(\frac{-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \ln(\pi_1/\pi_2)}{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}}\right) \\
& \quad + \pi_2 \Phi\left(\frac{-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln(\pi_2/\pi_1)}{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}}\right),
\end{aligned}$$

where  $\Phi(\cdot)$  is the standard normal cdf.

- Estimation via  $M$ -fold cross validation (CV)
  - Implementation

1. The dataset is randomly partitioned into  $M$  chunks.
  2. Train one classifier upon each combination of  $M - 1$  chunks.
  3. Apply each classifier to the corresponding remaining chunk and compute the empirical error rate.
  4. Estimate the population error rate by averaging these  $M$  empirical error rates.
- Comment
    - \* Leave-one-out CV  $\Leftrightarrow n$ -fold CV
  - Estimation via  $M \times L$ -fold CV
    - Implementation
      1. Repeat the four steps of  $M$ -fold CV  $L$  times.
      2. Average all the  $ML$  resulting empirical error rates.
    - Comment
      - \*  $M \times 1$ -fold CV  $\Leftrightarrow M$ -fold CV
- 

## A joint application of LDA/QDA & PCA

- Revisit the dataset of handwritten digits Part 7: `mnist` is a list with two components: `train` and `test`. Each of these is a list with two components: images and labels.
    - The `images` component is a matrix with each row for one image consisting of  $28 \times 28 = 784$  entries (pixels). Their value are integers between 0 and 255 representing grey scale.
    - The `labels` components is a vector representing the digit shown in the image.
    - Uninvertible  $\mathbf{S}_k$  because of the shared blank on canvas
- 

## Alternative methods for classification in the view of regression

- (Multinomial) logistic regression
- $k$ -nearest neighbors ( $k$ -NN)
- Tree-based
  - Decision tree/classification and regression tree (CART)
  - Random forest