STA 314: Statistical Methods for Machine Learning I

Lecture - Discriminant Analysis

Xin Bing

Department of Statistical Sciences
University of Toronto

Discriminant Analysis

- Logistic regression directly parametrizes

$$\mathbb{P}(Y = k \mid X = \mathbf{x}), \qquad \forall k \in C.$$

- By contrast, **Discriminant Analysis** parametrizes the distribution of

$$X \mid Y = k, \qquad \forall k \in C.$$

Normal distributions are oftentimes used.

Discriminant Analysis

What does parametrizing $X \mid Y = k$ buy us?

- By Bayes' theorem,

$$\mathbb{P}(Y = k \mid X = \mathbf{x}) = \frac{\mathbb{P}(X = \mathbf{x} \mid Y = k)\mathbb{P}(Y = k)}{\mathbb{P}(X = \mathbf{x})}.$$

Thus, to compare two classes $k, k' \in C$ with $k \neq k'$

$$\begin{aligned} & \mathbb{P}(Y = k \mid X = \mathbf{x}) \geq \mathbb{P}(Y = k' \mid X = \mathbf{x}) \\ \iff & \mathbb{P}(X = \mathbf{x} \mid Y = k)\mathbb{P}(Y = k) \geq \mathbb{P}(X = \mathbf{x} \mid Y = k')\mathbb{P}(Y = k') \end{aligned}$$

Notation for discriminant analysis

Suppose we have K classes, $C = \{0, 1, 2, \ldots, K-1\}$. For any $k \in C$,

- We write
$$\pi_k := \mathbb{P}(Y = k)$$
as the **prior** probability that a randomly chosen observation comes from the kth class.

- Write
$$f_k(\mathbf{x}) := \mathbb{P}(X = \mathbf{x} \mid Y = k)$$
as the **conditional density function** of $X = \mathbf{x}$ from class k.

- In discriminant analysis, **parametric assumption is assumed on $f_k(\mathbf{x})$**.

The Bayes rule

- By the Bayes' theorem,

$$p_k(\mathbf{x}) := \mathbb{P}(Y = k \mid X = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell \in C} \pi_\ell f_\ell(\mathbf{x})}$$

is called the **posterior** probability, i.e. the probability that an observation belongs to the kth class given its feature.

- According to the Bayes classifier, we should classify a new point $\mathbf{x}$ according to

$$\arg\max_{k \in C} \; p_k(\mathbf{x}) = \arg\max_{k \in C} \; \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell \in C} \pi_\ell f_\ell(\mathbf{x})} = \arg\max_{k \in C} \; \pi_k f_k(\mathbf{x}).$$

Discriminant Analysis for $p = 1$

- Assume that

$$X \mid Y = k \sim N(\mu_k, \sigma_k^2), \qquad \forall k \in C,$$

namely,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}.$$

- **Linear Discriminant Analysis (LDA)** further assumes

$$\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_{K-1}^2 = \sigma^2.$$

Linear Discriminant Analysis for $p = 1$

- As a result,

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{\ell \in C} \pi_\ell f_\ell(x)} = \frac{\pi_k e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}}{\sum_{\ell \in C} \pi_\ell e^{-\frac{1}{2\sigma^2}(x-\mu_\ell)^2}}.$$

- The Bayes rule classifies $X = x$ to

$$\begin{aligned} \arg\max_{k \in C} \; p_k(x) &= \arg\max_{k \in C} \; \log\left(p_k(x)\right) \\ &= \arg\max_{k \in C} \; \underbrace{\frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k}_{\delta_k(x)} \qquad \text{(verify!)} \end{aligned}$$

The name LDA is due to the fact that the **discriminant function** $\delta_k(x)$ is a linear function in x.

Linear Discriminant Analysis for $p = 1$

For binary case, i.e. $K = 2$,

$$\arg\max_{k\in\{0,1\}} p_k(x) = \arg\max_{k\in\{0,1\}} \left[\frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \right]$$

- If the priors are equal $\pi_0 = \pi_1$ and suppose $\mu_1 \geq \mu_0$, then the Bayes classifier assigns $X = x$ to

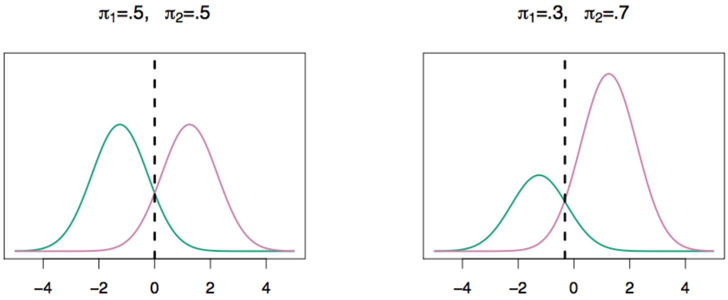
$$\begin{cases} 0 & \text{if } x < \frac{\mu_0+\mu_1}{2} \\ 1 & \text{if } x \geq \frac{\mu_0+\mu_1}{2} \end{cases}$$

The line $x = (\mu_0 + \mu_1)/2$ is called **the Bayes decision boundary**.

Example of LDA in binary classification

Consider $\mu_0 = -1.5$, $\mu_1 = 1.5$, and $\sigma = 1$. The curves are $p_0(x)$ (green) and $p_1(x)$ (red). The dashed vertical lines are the Bayes decision boundary.

$$f^*(x) = \begin{cases} 0 & \text{if } x < \frac{\mu_0+\mu_1}{2} = 0 \\ 1 & \text{if } x \geq \frac{\mu_0+\mu_1}{2} = 0 \end{cases}$$



Compute the Bayes classifier

- If we know $\mu_0, \ldots, \mu_{K-1}$, σ^2 and $\pi_0, \ldots, \pi_{K-1}$, then we can construct the Bayes rule

$$\arg\max_{k \in C} \delta_k(x) = \arg\max_{k \in C} \left\{ \frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \right\}.$$

- However, we typically don't know these parameters. We need to use the training data to estimate them!

Estimation under LDA

Given training data $(x_1, y_1), \ldots, (x_n, y_n)$, for all $k \in C$,

- We have

$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}.$$

- We estimate π_k by

$$\hat{\pi}_k = \frac{n_k}{n}.$$

- We estimate μ_k and σ^2 by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{1 \le i \le n : y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{K} \sum_{1 \le i \le n : y_i = k} (x_i - \hat{\mu}_k)^2.$$

These are actually the MLEs.

The LDA classifier

- We estimate $\delta_k(x)$ by the plug-in estimator

$$\hat{\delta}_k(x) = \frac{\hat{\mu}_k}{\hat{\sigma}^2}x - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k.$$

- The LDA classifier assigns x to

$$\arg\max_{k \in C} \ \hat{\delta}_k(x).$$

- How about the case when $p > 1$?

Linear Discriminant Analysis for $p > 1$

- Recall that the posterior probability has the form

$$P(Y = k \mid X = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell \in C} \pi_\ell f_\ell(\mathbf{x})},$$

- Now, we assume

$$X \mid Y = k \sim N_p(\mu_k, \Sigma), \quad \forall k \in C,$$

that is,

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^\top \Sigma^{-1}(\mathbf{x}-\mu_k)}.$$

- The discriminant function becomes

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1} \mu_k + \log \pi_k$$

  c.f. the univariate case

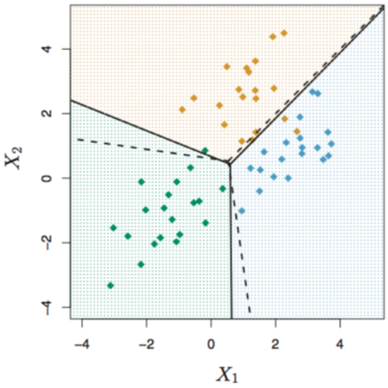
$$\delta_k(\mathbf{x}) = \frac{\mu_k}{\sigma^2}\mathbf{x} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k.$$

- The Bayes decision boundaries are the set of $\mathbf{x}$ for which

$$\delta_k(\mathbf{x}) = \delta_\ell(\mathbf{x}), \quad \forall k \neq \ell,$$

  which are again **linear hyperplanes** in $\mathbb{R}^p$.

Example



There are three classes (orange, green and blue) with two features X_1 and X_2. Dashed lines are the Bayes decision boundaries. Solid lines are their estimates based on the LDA.

Estimation under LDA for $p > 1$

Given the training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, for any $k \in C$,

- We have
$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}.$$
- We estimate π_k by
$$\hat{\pi}_k = \frac{n_k}{n}.$$

The slight difference is to estimate μ_k and Σ by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{1 \le i \le n: y_i = k} \mathbf{x}_i$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{1 \le i \le n: y_i = k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^\top.$$

A plugin rule for estimating discriminant functions

- We use the plugin estimator

$$\hat{\delta}_k(\mathbf{x}) = \mathbf{x}^\top \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^\top \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k, \quad \forall k \in C.$$

- The resulting LDA classifier is

$$\arg\max_{k \in C} \ \hat{\delta}_k(\mathbf{x}).$$

Logistic Regression v.s. LDA: similarity

For binary classification of LDA , one can show that

$$\log\left(\frac{p_1(\mathbf{x})}{1-p_1(\mathbf{x})}\right) = \log\left(\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\right)$$
$$= c_0 + c_1 x_1 + \cdots + c_p x_p,$$

where the $c_0, c_1, \ldots, c_p$ depends on $\pi_0, \pi_1, \mu_0, \mu_1$ and Σ.

The log-odds under LDA is also a linear form in both the parameters and the features (c.f. the logistic regression).

Logistic Regression v.s. LDA: differences

1. LDA makes more assumption by specifying $X \mid Y$.

2. The parameters are estimated differently.
   - Logistic regression uses the conditional likelihood based on $\mathbb{P}(Y|X)$ (known as discriminative learning).
   - LDA uses the full likelihood based on $\mathbb{P}(X, Y)$ (known as generative learning).

3. If classes are well-separated, then logistic regression is not advocated.

Other forms of Discriminant Analysis

LDA specifies

$$X \mid Y = k \sim N(\mu_k, \Sigma), \qquad \forall k \in C.$$

Other discriminant analyses change the specifications for $X \mid Y = k$.

- **Quadratic discriminant analysis** (QDA) assumes

  $$X \mid Y = k \sim N(\mu_k, \Sigma_k), \qquad \forall k \in C,$$

  by allowing different Σ_k across all classes.

Quadratic Discriminant Analysis: $p = 1$

- Assume that

$$X \mid Y = k \sim N(\mu_k, \sigma_k^2), \qquad \forall k \in C,$$

namely,

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}.$$

- As a result,

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{\ell \in C} \pi_\ell f_\ell(x)} = \frac{\frac{\pi_k}{\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}}{\sum_{\ell \in C} \frac{\pi_\ell}{\sigma_\ell} e^{-\frac{1}{2\sigma_\ell^2}(x-\mu_\ell)^2}}.$$

Decision boundary of QDA

The Bayes rule classifies $X = x$ to

$$\begin{aligned}\arg\max_{k\in C}\ p_k(x) &= \arg\max_{k\in C}\ \log\left(p_k(x)\right)\\ &= \arg\max_{k\in C}\ \log\left[\frac{\pi_k}{\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}\right]\\ &= \arg\max_{k\in C}\ \underbrace{-\frac{\mathbf{x^2}}{\mathbf{2\sigma_k^2}} + \frac{\mu_k}{\sigma_k^2}x - \frac{\mu_k^2}{2\sigma_k^2} + \log\pi_k - \log(\sigma_\mathbf{k})}_{\delta_k(x)}\end{aligned}$$

The name QDA is due to the fact that $\delta_k(x)$ is **quadratic** in x.

Quadratic Discriminant Analysis: $p \geq 1$

$$X \mid Y = k \sim N_p(\mu_k, \Sigma_k)$$

The discriminant function becomes

$$\delta_k(\mathbf{x}) = \log\left[\frac{\pi_k}{|\boldsymbol{\Sigma}_\mathbf{k}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^\top \boldsymbol{\Sigma}_\mathbf{k}{}^{-1}(\mathbf{x}-\mu_k)}\right]$$

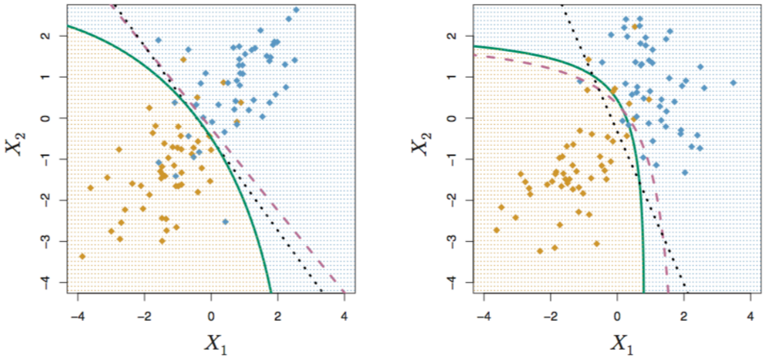
$$= \mathbf{x}^\top \Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma_k^{-1}\mu_k + \log \pi_k - \mathbf{\frac{1}{2}}\mathbf{x}^\top \boldsymbol{\Sigma}_\mathbf{k}^{-1}\mathbf{x} - \mathbf{\frac{1}{2}}\log|\boldsymbol{\Sigma}_\mathbf{k}|.$$

The **decision boundary** between any class k and class ℓ

$$\{\mathbf{x} \in \mathbb{R}^p : \delta_k(\mathbf{x}) = \delta_\ell(\mathbf{x})\}$$

is **quadratic** in x

Decision boundaries of LDA and QDA



Decision boundaries of the Bayes classifier (purple dashed), LDA (black dotted), and QDA (green solid) in two scenarios.

Estimation of QDA

Given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, for any $k \in C$,

- We have
$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}.$$
- We estimate π_k by
$$\hat{\pi}_k = \frac{n_k}{n}.$$
- We estimate μ_k and Σ_k by
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{1 \le i \le n: y_i = k} \mathbf{x}_i$$
$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{1 \le i \le n: y_i = k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^\top.$$
- Plugin estimator for $\delta(\mathbf{x})$.

Potential problems for LDA and QDA in high dimension

- LDA: we have
$$(K-1) + pK + \frac{p(p+1)}{2}$$
number of parameters to estimate.

- QDA: we have
$$(K-1) + pK + \frac{p(p+1)}{2}K$$
number of parameters to estimate.

- The estimation error is large when p is large comparing to n.

Naive Bayes

Other discriminant analyses: different density of $X \mid Y = k$, including non-parametric approaches.

- **Naive Bayes** assumes

$$X_1, \ldots, X_p \text{ are independent given } Y = k$$

so that

$$f_k(\mathbf{x}) = \prod_{j=1}^{p} f_{k,j}(x_j)$$

- It is easy to deal with both quantitative and categorical features.

- Despite the strong independence assumption within class, naive Bayes often produces good classification results.

Naive Bayes

- For Gaussian density

$$X_j \mid Y = k \sim N(\mu_{k,j}, \sigma^2_{k,j}),$$

this means that $\Sigma_k = \mathrm{diag}(\sigma^2_{k,1}, \sigma^2_{k,2}, \dots, \sigma^2_{k,p})$ and

$$f_k(\mathbf{x}) = \prod_{j=1}^{p} \frac{1}{\sigma_{k,j}\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2_{k,j}}(x_j - \mu_{k,j})^2}$$

- The discriminant function is

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(x_j - \mu_{kj})^2}{\sigma^2_{kj}} + \log \pi_k - \frac{1}{2}\sum_{j=1}^{p} \log \sigma^2_{kj}.$$