Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Discriminant analysis

- Model the distribution of X in each of the classes separately, and then use *Bayes theorem* to flip things around and obtain $Pr(Y|X)$
- Using normal (Gaussian) distributions for each class leads to linear or quadratic discriminant analysis

Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are unstable
- If the distribution of the predictors X is approximately normal in each of the classes and n is small, the linear discriminant model may be more accurate than the logistic regression model
- Linear discriminant analysis is popular when we have more than two response classes

Bayes theorem for classification

Bayes theorem:

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k) \cdot Pr(Y = k)}{Pr(X = x)}$$

For discriminant analysis:

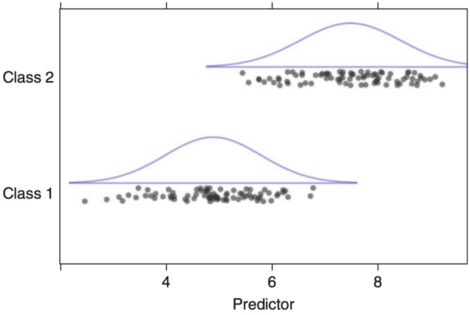
$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where

- $f_k(x) = Pr(X = x|Y = k)$ is the density for X in class k. Here we will use normal densities, separately in each class
- $\pi_k = Pr(Y = k)$ is the marginal or prior probability for class k

Classify to the highest density

- We classify a new point according to which density is highest
- When the priors are different, we take them into account as well, and compare $\pi_k f_k(x)$



Linear discriminant analysis when $p = 1$

- The Gaussian density has the form

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

- We assume that all the $\sigma_k = \sigma$ are the same
- Plugging this into Bayes formula, we get the following expression for $p_k(x) = Pr(Y = k|X = x)$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Discriminant functions

- Which of the $p_k(x)$ is largest?
- Equivalent to assigning x to the class with the largest discriminant score:

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

- $\delta_k(x)$ is a *linear* function of x
- If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$?

Example

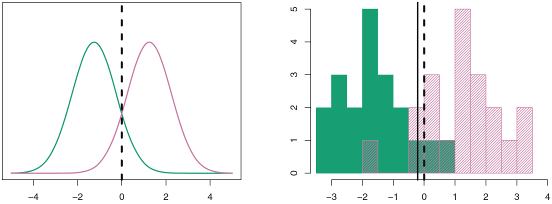


Figure: ISL 4.4

Example with $\mu_1 = -1.5, \mu_2 = 1.5, \pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$

- Typically we don't know these parameters
- Estimate the parameters and plug them into the rule

Estimating the parameters

- $\hat{\pi}_k = \frac{n_k}{n}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$
- $\hat{\sigma}_k^2 = \frac{1}{n_k-1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the k-th class

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

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \hat{\sigma}_k^2$$

Linear discriminant analysis when $p > 1$

- Multivariate normal density

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^\intercal \Sigma_k^{-1}(x-\mu_k)}$$

- Linear discriminant analysis (LDA): assume

$$\Sigma_k = \Sigma, \ \forall k$$

- Comparing two classes

$$\begin{aligned}
\log \frac{Pr(Y = k \mid X = x)}{Pr(Y = l \mid X = x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^\intercal \Sigma^{-1}(\mu_k - \mu_l) + x^\intercal \Sigma^{-1}(\mu_k - \mu_l) \\
&= \alpha_0 + \alpha^\intercal x
\end{aligned}$$

- Decision boundary?

LDA vs. logistic regression

- Logistic regression
  - Maximize conditional likelihood based on $Pr(Y \mid X)$
  - Normal assumption?
- LDA
  - Estimating π_k, μ_k, Σ amounts to estimating α, α_0
  - Full likelihood based on $Pr(Y, X)$

LDA when $p > 1$

- Discriminant function

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

- Bayes decision rule: $\arg\max_k \delta_k(x)$
- $\delta_k(x)$ is a linear in x
- π_k, μ_k, Σ can be estimated from the training data

$$\hat{\pi}_k = n_k / n, \quad \hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$

$$\hat{\Sigma} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\intercal$$

- $C(x) = \arg\max_k \hat{\delta}_k(x)$

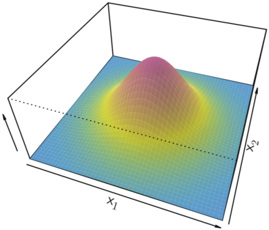
LDA when $p > 1$

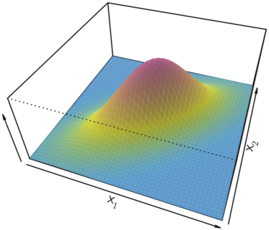
- Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

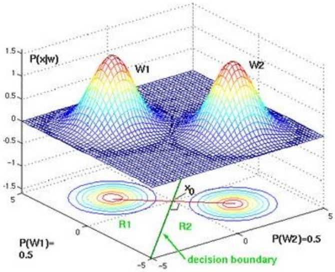
$$\widehat{Pr}(Y = k \mid X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

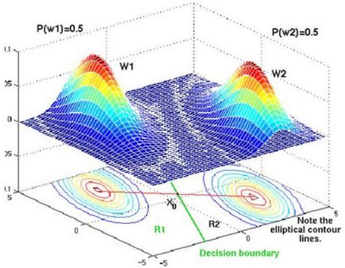
- Classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{Pr}(Y = k \mid X = x)$ is largest

Illustration: $p = 2$, $K = 2$ and $\pi_1 = \pi_2$









- Left: Covariance matrix $\sigma^2 I$ Right: Covariance matrix Σ

Illustration: $p = 2$ and $K = 3$

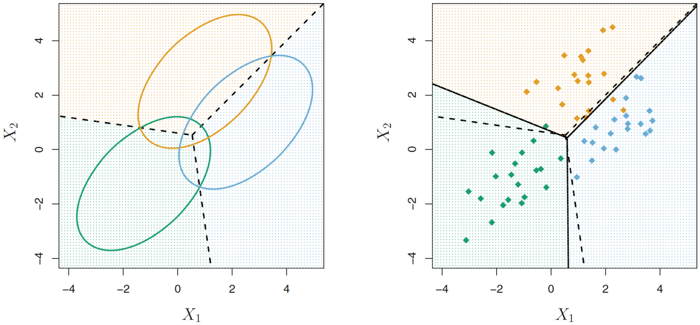


Figure: ESL 4.5

$\pi_1 = \pi_2 = \pi_3 = 1/3$
The dashed lines are Bayes decision boundaries

LDA computations

- Decision boundaries are useful for graphical purposes
- Simply compute $\hat{\delta}_k(x)$ for $k = 1, \ldots, K$ for classification
- Equivalently, minimize over k,

$$
\begin{aligned}
&\frac{1}{2}(x - \hat{\mu}_k)^\intercal \hat{\Sigma}^{-1}(x - \hat{\mu}_k) - \log \hat{\pi}_k \\
=&\frac{1}{2}\|x^* - \hat{\mu}_k^*\|_2^2 - \log \hat{\pi}_k
\end{aligned}
$$

- $\hat{\Sigma} = UDU^\intercal$ is the eigendecomposition
- Sphering the data: $x^* = D^{-1/2}U^\intercal x$, $\hat{\mu}_k^* = D^{-1/2}U^\intercal \hat{\mu}_k$

Classification can be achieved by sphering the data, and classifying to the closest centroid (adjusting for $\log \pi_k$) in the sphered space

Linear subspace spanned by sphered centroids

- LDA compares $\frac{1}{2}\|x^* - \hat{\mu}_k^*\|_2^2 - \log \hat{\pi}_k$ across k
- Consider the case where p is much larger than K
- The transformed centroids $(\hat{\mu}_1^*, \ldots, \hat{\mu}_K^*)$ span a subspace H (dimension $\leq K - 1$)
- Project x^* onto H, $x^* = P_H x^* + P_{H^\perp} x^*$

$$\|x^* - \hat{\mu}_k^*\|_2^2 = \|P_H(x^* - \hat{\mu}_k^*)\|_2^2 + \|P_{H^\perp} x^*\|_2^2$$

- Equivalently, LDA compares $\frac{1}{2}\|P_H(x^* - \hat{\mu}_k^*)\|_2^2 - \log \hat{\pi}_k$ across k

Since only the relative distance to the centroids count, one can confine the data to H, the subspace spanned by the centroids in the sphered space

LDA procedure summarized

LDA projects a feature space onto a smaller subspace while maintaining the class-discriminatory information

- Estimate $\pi_k, \mu_k, \Sigma, k = 1, \ldots, K$
- Transformation
  - Sphere the data points using $\hat{\Sigma}$
  - Project onto the subspace spanned by the sphered centroids
  - These can be summarized using $\tilde{x} = Ax \in \mathbb{R}^{K-1}$, where A is a $(K-1) \times p$ matrix
- Given x, transform to $\tilde{x} = Ax \in \mathbb{R}^{K-1}$, and classify according to the class for which $\frac{1}{2}\|\tilde{x} - \tilde{\mu}_k\|_2^2 - \log \hat{\pi}_k$ is minimized, where $\tilde{\mu}_k = A\hat{\mu}_k$

When there are K classes, LDA can be viewed in a $K-1$ dimensional plot

An alternative derivation: Fisher's optimization criteria

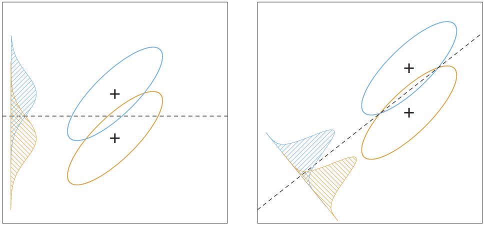
Find the linear combination $Z = a^\intercal X$ such that the between-class variance is maximized relative to the within-class variance

- W - within-class covariance matrix of X, i.e., $\hat{\Sigma}$ in LDA
- $B = \sum_{k=1}^{K} \hat{\pi}_k (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^\intercal$ - between-class covariance matrix
- Find a_1 that maximizes

$$\frac{a^\intercal B a}{a^\intercal W a}$$

- Find a_2 orthogonal in W to a_1 such that $\frac{a_2^\intercal B a_2}{a_2^\intercal W a_2}$ is maximized, and so on
- $a_1, a_2, \ldots$ are referred to as discriminant coordinates, $Z_l = a_l^\intercal X$ is the lth discriminant variable

Discriminant direction



- Left: direction of greatest centroid spread
- Right: discriminant direction minimizes this overlap for Gaussian data

Normality assumption

- LDA assumes normally distributed features
- For classification tasks, LDA can be quite robust to the distribution of the data

Extensions

$$Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- When $f_k(x)$ are Gaussian densities with the same covariance matrix Σ in each class, we get LDA
- With Gaussians but different Σ_k in each class, we get *quadratic discriminant analysis* (QDA)
- With $f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)$ (conditional independence model) in each class, we get *naive Bayes*
  - For Gaussian: Σ_k are diagonal

Quadratic Discriminant Analysis

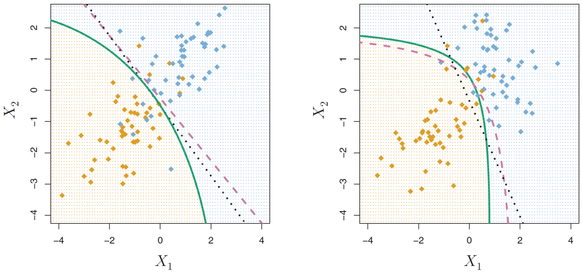
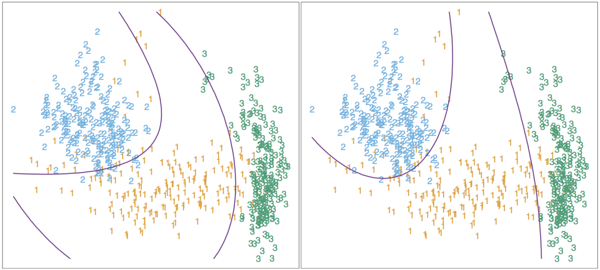


Figure: ISL 4.9

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^\intercal \Sigma_k^{-1}(x-\mu_k) + \log \pi_k - \frac{1}{2}\log|\Sigma_k|$$

Because the Σ_k are different, the quadratic terms matter

Quadratic boundaries



- Left: LDA using $X_1, X_2, X_1 X_2, X_1^2, X_2^2$
- Right: QDA
- Similar results

Naive Bayes

- Assumes features are independent in each class
- Useful when p is large, and so multivariate methods like LDA break down
- Gaussian naive Bayes assumes Σ_k is diagonal:

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\left[\frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} + \log \sigma_{kj}^2\right] + \log \pi$$

- Can use for mixed feature vectors (qualitative and quantitative)
- If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function over discrete categories
- Despite strong assumptions, naive Beyes often produces good classification results

k-Nearest-Neighbor classifiers

Predict class label given x_0

- Find the k training points $x_{(r)}$, $r = 1, \ldots, k$, closest in distance to x_0
- Classify using majority vote among the k neighbors

$$C(x_0) = j \text{ such that } \sum_{r=1}^{k} I(y_{(r)} = j) \text{ is largest}$$

- Tuning parameter: k
- Disadvantages
  - Limited insight into relationship between the predictors and the response
  - Computation: need the entire dataset when classifying a new point x_0 - prediction is slow

Summary

- Logistic regression is popular for classification especially when $K = 2$
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable, also when $K > 2$
- Naive Bayes is useful when p is very large