Linear Discriminant Analysis

Predictive Modeling & Statistical Learning

Gaston Sanchez

CC BY-SA 4.0

Linear Discriminant Analysis

Note

In these slides, I am assuming that all variables (predictors and response) are centered (mean $= 0$)!

Probabilistic Discriminant Analysis

A couple of slides ago we described the Bayesian approach for classification purposes:

$$P(Y = k | X = x) = \frac{\pi_k \, f_k(x)}{\sum_{k=1}^{K} \pi_k \, f_k(x)}$$

- $P(Y = k) = \pi_k$, the **prior** probability for class k.
- $P(X = x | Y = k) = f_k(x)$, the **density** for X in class k.

Keep in mind

However, the Bayes formula

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

does NOT tell us:

- how to calculate priors π_k
- what form should we use for densities $f_k(x)$

There is plenty of room to play with π_k and $f_k(x)$

Welcome to LDA

Linear Discriminant Analysis involves considering Normal distributions for the densities $f_k(x)$

Welch (1939), based on Fisher's works (1936, 1938) was the first one to assume normal densities.

LDA with one predictor $p = 1$

The Normal density has the form:

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

where:

- μ_k is the mean (in class k)
- σ_k^2 is the variance (in class k)

LDA with one predictor $p = 1$

Plugging the Normal density $f_k(x)$ into the Bayes formula we get a rather complex expression for $p_k(x) = P(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}}{\sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}}$$

LDA with one predictor $p = 1$

If we assume constant variances: $\sigma_k^2 = \sigma^2$ we get:

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

LDA with one predictor $p = 1$

Typically, we don't know the class priors π_k, the class means μ_k, and the variance σ^2, so we estimate them with the training data:

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

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in G_k} x_i = g_k$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i \in G_k} (x_i - g_k)^2$$

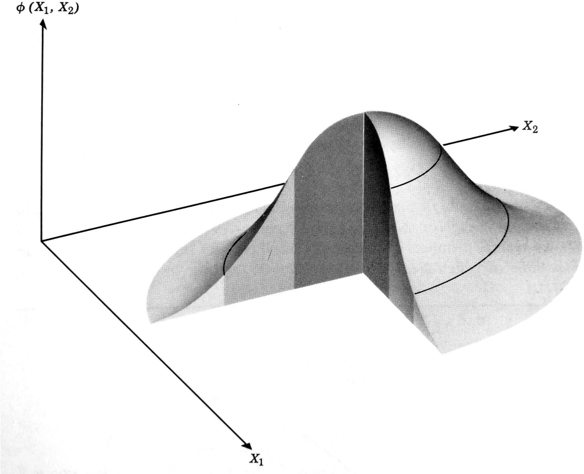
Multivariate Normal Distribution

Bivariate Normal Distribution

For the case of two variable X_1 and X_2, the bivariate normal density function is:

$$f(x)$$

Bivariate Normal Density Surface



Bivariate normal density surface (Tatsuoka, 1988, p.67)

Bivariate Normal Distribution

Multivariate Nomal (MVN) distribution

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\mathsf{T}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu_k})}$$

Bayes Classifier

Bayes classifier involves choosing class k for which $\pi_k f_k(x)$ is maximum

$$\hat{\pi}_k \hat{f}_k(\mathbf{x}) = \underset{k}{argmax} \left\{ \pi_k f_k(\mathbf{x}) \right\}$$

Score or discriminant function

The discriminant score $\delta_k(x)$ is given by:

$$\delta_k(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu_k})^\mathsf{T} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu_k}) + log(|\boldsymbol{\Sigma}_k|) - 2log(\pi_k)$$

Score or discriminant function

When all classes have the same covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}, \forall k$, the discriminant score $\delta_k(x)$ is given by:

$$\delta_k(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu_k})^\mathsf{T} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu_k}) - 2log(\pi_k)$$

Score or discriminant function

If in addition to having the same covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}, \forall k$, we also assume same prior probabilities $\pi_k = \pi$, the discriminant score $\delta_k(x)$ becomes:

$$\delta_k(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu_k})^\mathsf{T} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu_k})$$

which is the Bayesian rule that assigns $\mathbf{x}$ to the closest centroid $\boldsymbol{\mu_k}$ according to the Mahalanobis distance.

Summary

Assumptions	**Bayesian Rule**
Multinormal Distribution	Quadratic (QDA)
Multinormal Distribution + Same variances	Linear (LDA)
Multinormal Distribution + Same variances + Same priors	Linear, equivalent to geometric rule (CDA)

In practice ...

Typically, we don't know the class priors π_k, the class means $\boldsymbol{\mu}_k$, and the variance matrices $\boldsymbol{\Sigma}_k$, so we estimate them with:

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

$$\hat{\boldsymbol{\mu}}_k = \mathbf{g_k}$$

$$\hat{\boldsymbol{\Sigma}}_k = \mathbf{W_k} = \frac{1}{n_k} \mathbf{X_k^\mathsf{T}} \mathbf{X_k}$$

LDA for multiple predictors

LDA with multiple predictors $p \geq 2$

A multivariate normal density is:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^{\mathsf{T}}\mathbf{\Sigma}(x-\mu)}$$

LDA with multiple predictors $p \geq 2$

The discriminant function $\delta_k(x)$ is:

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

From $\delta_k(x)$ to probabilities

Once we have estimates $\hat{\delta}_k(x)$, we can use to estimate class probabilities:

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

So classifying to the largest $\delta_k(x)$ amounts to classifying to the class for which $\hat{Pr}(Y = k|X = x)$ is largest.

Bibliography

- **The use of multiple measurements in taxonomic problems** by R.A. Fisher (1936). *Annals of Eugenics, 7, 179-188*.
- **Principles of Multivariate Analysis: A User's Perspective** by W.J. Krzanowski (1988). *Chapter 11: Incorporating group structure: descriptive methods*. Wiley.
- **On the generalized distance in statistics** by P.C. Mahalanobis (1936). *Proceedings of the National Institute of Science, India, 12, 49-55*.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*.
- **Multivariate Analysis** by Maurice Tatsuoka (1988). *Chapter 7: Discriminant Analysis and Canonical Correlation*.
- **Discriminant Analysis** by Tatsuoka and Tiedeman (1954). *Review of Educational Research, 25, 402-420*.

French Literature

- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3: Analyse factorielle discriminante*. Dunod, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique explicative appliquee** by Nakache and Confais (2003). *Chapter 1: Analyse discriminante sur variables quantitatives*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 10: L'analyse discriminante*. Dunod, Paris.