DISCRIMINANT ANALYSIS

Concha Bielza, Pedro Larrañaga

Computational Intelligence Group
Departmento de Inteligencia Artificial
Universidad Politécnica de Madrid

Machine Learning

Outline

1. Introduction
2. LDA. Equal spherical covariance matrices
3. LDA. Equal covariance matrices
4. QDA. Arbitrary covariance matrices
5. Conclusions

Outline

1. Introduction
2. LDA. Equal spherical covariance matrices
3. LDA. Equal covariance matrices
4. QDA. Arbitrary covariance matrices
5. Conclusions

Assumptions

- Assume the class-conditional density function $f(\mathbf{x}|c_r)$ follows a multivariate Gaussian $\mathbf{X}|c_r \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, i.e.

$$f(\mathbf{x}|c_r, \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_r|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T\boldsymbol{\Sigma}_r^{-1}(\mathbf{x}-\boldsymbol{\mu}_r)\right\},$$

  where $\boldsymbol{\mu}_r$ is the n-dimensional mean vector, $\boldsymbol{\Sigma}_r$ is the $n \times n$ covariance matrix and $|\boldsymbol{\Sigma}_r|$ its determinant, $r = 1, ..., R$

- Search for $c^* = \arg\text{máx}_r\, p(C = c_r|\mathbf{x})$, or equivalently maximize the discriminant function:

$$\begin{aligned} g_r(\mathbf{x}) &= \ln f(\mathbf{x}|c_r) + \ln p(C = c_r) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T\boldsymbol{\Sigma}_r^{-1}(\mathbf{x}-\boldsymbol{\mu}_r) - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_r| + \ln p(C = c_r) \end{aligned}$$

- Applying $g_r(\mathbf{x})$, the feature space is divided into R **decision regions**, $\mathcal{R}_1, ..., \mathcal{R}_R$: $\mathbf{x}$ is in $\mathcal{R}_r$ if $g_r(\mathbf{x}) = c_r$

Estimation of parameters

- Parameters estimated from data with their maximum likelihood estimates:

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_r &= \frac{1}{N_r} \sum_{i:c^i=c_r} \mathbf{x}^i \quad \textit{(sample mean)} \\
\hat{\boldsymbol{\Sigma}}_r &= \frac{1}{N_r} \sum_{i:c^i=c_r} (\mathbf{x}^i - \hat{\boldsymbol{\mu}}_r)(\mathbf{x}^i - \hat{\boldsymbol{\mu}}_r)^T \quad \textit{(sample covariance)} \\
\hat{p}(C = c_r) &= \frac{N_r}{N} \quad \textit{(relative frequency of class-}c_r\textit{ observations)}
\end{aligned}
$$

- Applying a multivariate Gaussian goodness-of-fit test will be necessary

Outline

1 Introduction

2 LDA. Equal spherical covariance matrices

3 LDA. Equal covariance matrices

4 QDA. Arbitrary covariance matrices

5 Conclusions

LDA. Equal spherical covariance matrices

Linear discriminant analysis

- $\mathbf{X}|c_r$ has zero covariances and the same variance σ^2 in $\mathbf{\Sigma}_r$ (for all c_r), i.e. X_i are conditionally independent
  - $\mathbf{\Sigma}_r = \sigma^2\mathbf{I} \Rightarrow |\mathbf{\Sigma}_r| = \sigma^{2n}$ and $\mathbf{\Sigma}_r^{-1} = (1/\sigma^2)\mathbf{I}$
- 2nd and 3rd addends in $g_r(\mathbf{x})$ can be ignored (do not depend on r) and

$$g_r(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu}_r)^T(\mathbf{x} - \boldsymbol{\mu}_r) + \ln p(C = c_r)$$

or equivalently ($\mathbf{x}^T\mathbf{x}$ does not depend on r) we obtain the linear function

$$g_r(\mathbf{x}) = \mathbf{w}_r^T\mathbf{x} + w_{r0}$$

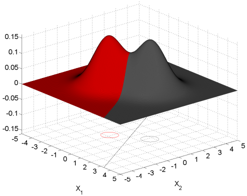
where

$$\begin{aligned} \mathbf{w}_r &= \frac{1}{\sigma^2}\boldsymbol{\mu}_r \\ w_{r0} &= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_r^T\boldsymbol{\mu}_r + \ln p(C = c_r) \end{aligned}$$

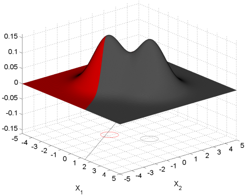
LDA. Equal spherical covariance matrices

- Decision boundary, defined by $g_r(\mathbf{x}) = g_k(\mathbf{x})$, is $\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\mu}_r - \boldsymbol{\mu}_k \\ \mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_r + \boldsymbol{\mu}_k) - \frac{\sigma^2}{(\boldsymbol{\mu}_r - \boldsymbol{\mu}_k)^T(\boldsymbol{\mu}_r - \boldsymbol{\mu}_k)} \ln \frac{p(c_r)}{p(c_k)}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_k) \end{aligned}$$



$p(c_1) = p(c_2) = .5$



$p(c_1) = .99, p(c_2) = .01$

The decision boundary is a hyperplane orthogonal to $\mathbf{w}$ (line linking the means) and passes through point $\mathbf{x}_0$.

- In (a), the hyperplane passes through the halfway point between the means: if $p(c_1) = p(c_2) \rightarrow \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$
- In (b), the decision is biased in favor of c_1

Outline

1 Introduction

2 LDA. Equal spherical covariance matrices

3 **LDA. Equal covariance matrices**

4 QDA. Arbitrary covariance matrices

5 Conclusions

LDA. Equal covariance matrices

Linear discriminant analysis

- Assume **homoscedasticity** $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}$, i.e., all equal although arbitrary
- The shared $\boldsymbol{\Sigma}$ is estimated using the whole data set as the pooled sample covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-R} \sum_{r=1}^{R} \sum_{i:c^i=c_r} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)^T$$

- The discriminant function is

$$g_r(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_r)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_r) + \ln p(C = c_r)$$

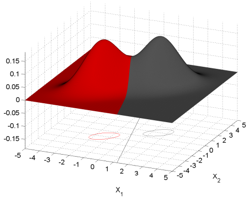
and since $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ does not depend on r either, $g_r(\mathbf{x}) = \mathbf{w}_r^T \mathbf{x} + w_{r0}$, where

$$\begin{aligned} \mathbf{w}_r &= \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_r \\ w_{r0} &= -\frac{1}{2} \boldsymbol{\mu}_r^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_r + \ln p(C = c_r) \end{aligned}$$

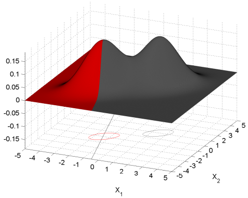
LDA. Equal covariance matrices

- Decision boundary is again a hyperplane $\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$, where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_k) \\ \mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_r + \boldsymbol{\mu}_k) - \frac{1}{(\boldsymbol{\mu}_r - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_k)} \ln \frac{p(c_r)}{p(c_k)} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_k) \end{aligned}$$



$p(c_1) = p(c_2) = .5$



$p(c_1) = .99, p(c_2) = .01$

The decision boundary is not necessarily orthogonal to the line linking the means

Outline

1. Introduction
2. LDA. Equal spherical covariance matrices
3. LDA. Equal covariance matrices
4. **QDA. Arbitrary covariance matrices**
5. Conclusions

QDA. Arbitrary covariance matrices

Quadratic discriminant analysis

- Assume different covariance matrices for each class label, $\mathbf{\Sigma}_r$
- Only 2nd addend in $g_r(\mathbf{x})$ can be ignored and g_r is now quadratic

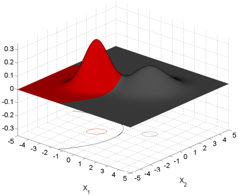
$$g_r(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_r \mathbf{x} + \mathbf{w}_r^T \mathbf{x} + w_{r0}$$

where

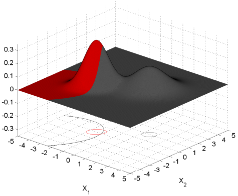
$$\begin{aligned}
\mathbf{W}_r &= -\frac{1}{2}\mathbf{\Sigma}_r^{-1} \\
\mathbf{w}_r &= \mathbf{\Sigma}_r^{-1}\boldsymbol{\mu}_r \\
w_{r0} &= -\frac{1}{2}\boldsymbol{\mu}_r^T \mathbf{\Sigma}_r^{-1} \boldsymbol{\mu}_r - \frac{1}{2}\ln|\mathbf{\Sigma}_r| + \ln p(C = c_r)
\end{aligned}$$

QDA. Arbitrary covariance matrices

- For C binary, the decision boundaries are hyperquadrics with any general form: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, etc.
- For more than two classes, the extension is straightforward and may result in many different and complicated regions



$p(c_1) = p(c_2) = .5$



$p(c_1) = .99, p(c_2) = .01$

Regions separated by the hyperbola

Outline

1. Introduction
2. LDA. Equal spherical covariance matrices
3. LDA. Equal covariance matrices
4. QDA. Arbitrary covariance matrices
5. Conclusions

Conclusions

Summary

- Gaussian assumption for class-conditional density
- Linear and quadratic cases
- More assumptions than logistic regression
- Since $g_r(\mathbf{x}) = \ln p(C = c_r, \mathbf{X} = \mathbf{x})$, then
  $g_r(\mathbf{x}) - g_R(\mathbf{x}) = \ln \frac{f(C=r,\mathbf{X}=\mathbf{x})}{f(C=R,\mathbf{X}=\mathbf{x})} = \ln \frac{p(C=r|\mathbf{x})}{p(C=R|\mathbf{x})}$, that in LDA is a linear combination
  $\beta'_{r0} + \beta'_{r1}x_1 + \cdots + \beta'_{rn}x_n$ *As in LOGREG*
  - $\Rightarrow$ logistic regression and LDA have the same form: the log-posterior odds for a pair of classes is a linear function of $\mathbf{x}$
  - $\Rightarrow$ However, parameters are estimated differently:
    - $f(\mathbf{x}, c) = f(\mathbf{x})p(c|\mathbf{x})$, with $p(c|\mathbf{x})$ in a logit-linear form *in both*
    - Logistic fits the parameters of $p(c|\mathbf{x})$ by maximizing the *conditional* log-likelihood. A **discriminative** classifier (and ignors $f(\mathbf{x})$)
    - LDA, by maximizing the *full* log-likelihood. A **generative** classifier based on the joint density $f(\mathbf{x}, c) = f(\mathbf{x}|c)p(c)$, where $f(\cdot|\cdot)$ is a Gaussian density (and $f(\mathbf{x})$ is a Gaussian mixture density, not ignored)

In Weka

LDA, QDA within *Functions* [install them from the Package Manager]

Bibliography

Texts

- Bielza, C., Larrañaga, P. (2021) *Data-Driven Computational Neuroscience. Machine Learning and Statistical Models*, Cambridge University Press [Chap. 8]
- R. Duda, P. Hart, D.G. Stork (2001) *Pattern Classification*, John Wiley & Sons, 2nd Ed. [Chap 2]

DISCRIMINANT ANALYSIS

Concha Bielza, Pedro Larrañaga

Computational Intelligence Group
Departmento de Inteligencia Artificial
Universidad Politécnica de Madrid

Machine Learning