Lecture 9: A probabilistic approach to classification

Attendance code: RAK7HQXM

Iain Styles

8 November 2018

Learning Outcomes

By the end of this lecture you should be able to

- Develop a simple probabilistic model of a linearly separable binary class dataset
- Use the model to derive classification rules
- Generalise the model to non-linear boundaries and multiple classes
- Generate new samples using the model

Linear Discriminant Analysis

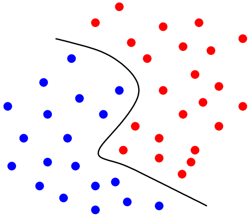
- Bayesian statistics will allow us to classify data in a probabilistic way

Linear Discriminant Analysis

- Bayesian statistics will allow us to classify data in a probabilistic way
- Build statistical models of classes

Linear Discriminant Analysis

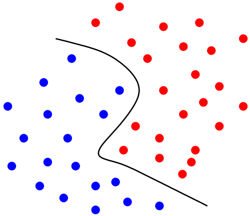
- Bayesian statistics will allow us to classify data in a probabilistic way
- Build statistical models of classes
- Construct explicit decision boundary



- Simplest base: linear binary classifier
- Linear combination of variables–a straight line–that best discriminates between two pre-defined groups of points
- Uses the probability distributions of the classes to find the line of equal probability dividing two classes

Linear Discriminant Analysis

- Bayesian statistics will allow us to classify data in a probabilistic way
- Build statistical models of classes
- Construct explicit decision boundary



- Simplest base: linear binary classifier
- Linear combination of variables–a straight line–that best discriminates between two pre-defined groups of points
- Uses the probability distributions of the classes to find the line of equal probability dividing two classes
- *Generative approach* to classification

General Approach

- Use Bayes rule: $P(A|B) = P(B|A)P(A)/P(B)$

General Approach

- Use Bayes rule: $P(A|B) = P(B|A)P(A)/P(B)$
- Define priors: expected proportions of data points in each class – $P(\mathrm{class})$

General Approach

- Use Bayes rule: $P(A|B) = P(B|A)P(A)/P(B)$
- Define priors: expected proportions of data points in each class – $P(\mathrm{class})$
- Define likelihoods: given class PDFs, where are the data points – $P(\mathrm{point}|\mathrm{class})$?

General Approach

- Use Bayes rule: $P(A|B) = P(B|A)P(A)/P(B)$
- Define priors: expected proportions of data points in each class – $P(\text{class})$
- Define likelihoods: given class PDFs, where are the data points – $P(\text{point}|\text{class})$?
- Apply Bayes: given data points, what are their classes – $P(\text{class}|\text{point})$

General Approach

- Use Bayes rule: $P(A|B) = P(B|A)P(A)/P(B)$
- Define priors: expected proportions of data points in each class – $P(\text{class})$
- Define likelihoods: given class PDFs, where are the data points – $P(\text{point}|\text{class})$?
- Apply Bayes: given data points, what are their classes – $P(\text{class}|\text{point})$
- Assign points to most probable classes

Formulating LDA

- Consider a data point $\mathbf{x}$ which we want to assign to one of two predefined classes Π_i $(i = 1, 2)$
- We write that $P(\mathbf{x} \in \Pi_i) = \pi_i$

Formulating LDA

- Consider a data point $\mathbf{x}$ which we want to assign to one of two predefined classes Π_i $(i = 1, 2)$
- We write that $P(\mathbf{x} \in \Pi_i) = \pi_i$: *Prior probability*

Formulating LDA

- Consider a data point $\mathbf{x}$ which we want to assign to one of two predefined classes Π_i $(i = 1, 2)$
- We write that $P(\mathbf{x} \in \Pi_i) = \pi_i$: *Prior probability*
- Class-conditional probability $P(\mathbf{x}|\Pi_i) = f_i(\mathbf{x})$

Formulating LDA

- Consider a data point $\mathbf{x}$ which we want to assign to one of two predefined classes Π_i $(i = 1, 2)$
- We write that $P(\mathbf{x} \in \Pi_i) = \pi_i$: *Prior probability*
- Class-conditional probability $P(\mathbf{x}|\Pi_i) = f_i(\mathbf{x})$: distribution of points in class Π_i

Formulating LDA

- Consider a data point $\mathbf{x}$ which we want to assign to one of two predefined classes Π_i $(i = 1, 2)$
- We write that $P(\mathbf{x} \in \Pi_i) = \pi_i$: *Prior probability*
- Class-conditional probability $P(\mathbf{x}|\Pi_i) = f_i(\mathbf{x})$: distribution of points in class Π_i
- These can be learned from the training data

Formulating LDA

- Consider a data point $\mathbf{x}$ which we want to assign to one of two predefined classes Π_i $(i = 1, 2)$
- We write that $P(\mathbf{x} \in \Pi_i) = \pi_i$: *Prior probability*
- Class-conditional probability $P(\mathbf{x}|\Pi_i) = f_i(\mathbf{x})$: distribution of points in class Π_i
- These can be learned from the training data
- Bayes rule: probability that $\mathbf{x}$ belongs to class Π_i is:

$$\begin{aligned} P(\Pi_i|\mathbf{x}) &= \frac{P(\mathbf{x}|\Pi_i)P(\Pi_i)}{P(\mathbf{x})} \\ &= \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2} \end{aligned}$$

Formulating LDA

- Given class-conditional likelihoods $f_i(\mathbf{x})$, and priors $\pi_x(\mathbf{x})$

Formulating LDA

- Given class-conditional likelihoods $f_i(\mathbf{x})$, and priors $\pi_x(\mathbf{x})$
- New points are assigned to the group with the highest probability:

$$\frac{P(\Pi_1|\mathbf{x})}{P(\Pi_2|\mathbf{x})} > 1 \mapsto \mathbf{x} \in \Pi_1 \quad \text{else}\, \mathbf{x} \in \Pi_2$$

Formulating LDA

- Given class-conditional likelihoods $f_i(\mathbf{x})$, and priors $\pi_x(\mathbf{x})$
- New points are assigned to the group with the highest probability:

$$\frac{P(\Pi_1|\mathbf{x})}{P(\Pi_2|\mathbf{x})} > 1 \mapsto \mathbf{x} \in \Pi_1 \quad \text{else}\, \mathbf{x} \in \Pi_2$$

- In terms of f and π, $\mathbf{x}$ belongs to Π_1 if

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

- otherwise to Π_2.

Formulating LDA

- Given class-conditional likelihoods $f_i(\mathbf{x})$, and priors $\pi_x(\mathbf{x})$
- New points are assigned to the group with the highest probability:

$$\frac{P(\Pi_1|\mathbf{x})}{P(\Pi_2|\mathbf{x})} > 1 \mapsto \mathbf{x} \in \Pi_1 \quad \text{else}\, \mathbf{x} \in \Pi_2$$

- In terms of f and π, $\mathbf{x}$ belongs to Π_1 if

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}$$

- otherwise to Π_2.
- Equal ratios, randomly assign to either class.

Defining the Classes

- We need to define the class distributions $f_i(\mathbf{x})$.
- For a simple two-class binary classifier we assume:
  - Each data point belongs to exactly one of exactly two distinct and identifiable groups, Π_1 and Π_2

Defining the Classes

- We need to define the class distributions $f_i(\mathbf{x})$.
- For a simple two-class binary classifier we assume:
  - Each data point belongs to exactly one of exactly two distinct and identifiable groups, Π_1 and Π_2
  - The two groups are normally distributed with different means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ but identical covariances $\boldsymbol{\Sigma}$.

Defining the Classes

- We need to define the class distributions $f_i(\mathbf{x})$.
- For a simple two-class binary classifier we assume:
  - Each data point belongs to exactly one of exactly two distinct and identifiable groups, Π_1 and Π_2
  - The two groups are normally distributed with different means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ but identical covariances $\mathbf{\Sigma}$.
- Covariance?

Covariance

- Variance: measure of variation in a variable

Covariance

- Variance: measure of variation in a variable
- Covariance: measure of how two variables vary with each

Covariance

- Variance: measure of variation in a variable
- Covariance: measure of how two variables vary with each
- $\Sigma_{X,Y} = \mathbb{E}\left[(X - \bar{X})(Y - \bar{Y})\right]$

Covariance

- Variance: measure of variation in a variable
- Covariance: measure of how two variables vary with each
- $\Sigma_{X,Y} = \mathbb{E}\left[(X - \bar{X})(Y - \bar{Y})\right]$
- Correlation $\rho_{x,y} = \Sigma_{x,y}/\sigma_x\sigma_y$.

Covariance

- Variance: measure of variation in a variable
- Covariance: measure of how two variables vary with each
- $\Sigma_{X,Y} = \mathbb{E}\left[(X - \bar{X})(Y - \bar{Y})\right]$
- Correlation $\rho_{x,y} = \Sigma_{x,y}/\sigma_x\sigma_y$.
- Multivariate problem: *covariance matrix* $\mathbf{\Sigma}$ with components

$$\Sigma_{ij} = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_i^{(n)} - \bar{x}_i\right)\left(x_j^{(n)} - \bar{x}_j\right) \tag{1}$$

Class-conditional Likelihood

- Model the classes as normally distributed with different means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ but identical covariances $\boldsymbol{\Sigma}$.
- Thus, for $n = \{1, 2\}$ the groups distributions are $P(\mathbf{x}|\Pi_i) = f_i(\mathbf{x})$:

$$f_n(\mathbf{x}) = \frac{1}{(2\pi)^{r/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\,[-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_n)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_n)]$$

The Separation Rule

- We know the ratio $\frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = 1$ separates the groups.

The Separation Rule

- We know the ratio $\frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = 1$ separates the groups.
- Taking logs: $\ln \frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} - \ln \frac{\pi_1}{\pi_2}$ with:

$$\begin{aligned} \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= (\mathbf{x} - \bar{\mathbf{x}}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_1) - (\mathbf{x} - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_2) \\ &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \end{aligned}$$

The Separation Rule

- We know the ratio $\frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = 1$ separates the groups.
- Taking logs: $\ln \frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} - \ln \frac{\pi_1}{\pi_2}$ with:

$$\begin{aligned}
\ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= (\mathbf{x} - \bar{\mathbf{x}}_1)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_1) - (\mathbf{x} - \bar{\mathbf{x}}_2)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_2) \\
&= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)
\end{aligned}$$

- The first term on the RHS is linear in $\mathbf{x}$
- The second term on the RHS is "constant" (no $\mathbf{x}$)

The Separation Rule

- We know the ratio $\frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = 1$ separates the groups.
- Taking logs: $\ln \frac{f_1(\mathbf{x})\pi_2}{f_2(\mathbf{x})\pi_1} = \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} - \ln \frac{\pi_1}{\pi_2}$ with:

$$\begin{aligned}
\ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= (\mathbf{x} - \bar{\mathbf{x}}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_1) - (\mathbf{x} - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_2) \\
&= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)
\end{aligned}$$

- The first term on the RHS is linear in $\mathbf{x}$
- The second term on the RHS is “constant” (no $\mathbf{x}$)
- This is a straight line / plane / hyperplane

The Separation Rule

- Writing $\mathbf{M} = \boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$
- And $c = -\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right) + \log_e \frac{\pi_1}{\pi_2}$

The Separation Rule

- Writing $\mathbf{M} = \boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$
- And $c = -\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right) + \log_e \frac{\pi_1}{\pi_2}$
- We have
$$L(\mathbf{x}) = \ln \frac{f_1(\mathbf{x})\pi_1}{f_2(\mathbf{x})\pi_2} = \mathbf{M}^{\mathrm{T}}\mathbf{x} + c,$$

The Separation Rule

- Writing $\mathbf{M} = \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
- And $c = -(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \log_e \frac{\pi_1}{\pi_2}$
- We have

$$L(\mathbf{x}) = \ln \frac{f_1(\mathbf{x})\pi_1}{f_2(\mathbf{x})\pi_2} = \mathbf{M}^{\mathrm{T}}\mathbf{x} + c,$$

- Given the separation rule $\frac{f_1(\mathbf{x})\pi_1}{f_2(\mathbf{x})\pi_2} = 1$ we have:

$$\text{if } L(\mathbf{x}) > 0 \text{ assign } \mathbf{x} \text{ to } \Pi_1 \text{ else } \Pi_2$$

- This is *Gaussian LDA*
- $\mathbf{M}^{\mathrm{T}}\mathbf{x}$ is *Fisher's linear discriminant function*.

Quadratic Discriminant Analysis

- Different class covariances require more complex analysis

Quadratic Discriminant Analysis

- Different class covariances require more complex analysis
- If the covariances of the two classes are $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ the discriminant becomes:

$$Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$$

Quadratic Discriminant Analysis

- Different class covariances require more complex analysis
- If the covariances of the two classes are $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ the discriminant becomes:

$$Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$$

a *Quadratic discriminant* with

$$\begin{aligned}
\mathbf{A} &= -\frac{1}{2}\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right) \\
\mathbf{b} &= \boldsymbol{\Sigma}_1^{-1}\bar{\mathbf{x}}_1 - \boldsymbol{\Sigma}_2^{-1}\bar{\mathbf{x}}_2 \\
c &= -\frac{1}{2}\left(\log_e \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \bar{\mathbf{x}}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\bar{\mathbf{x}}_2\right) - \log_e \frac{\pi_1}{\pi_2}
\end{aligned}$$

Quadratic Discriminant Analysis

- Different class covariances require more complex analysis
- If the covariances of the two classes are $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ the discriminant becomes:

$$Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$$

a *Quadratic discriminant* with

$$\begin{aligned}
\mathbf{A} &= -\frac{1}{2}\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right) \\
\mathbf{b} &= \boldsymbol{\Sigma}_1^{-1}\bar{\mathbf{x}}_1 - \boldsymbol{\Sigma}_2^{-1}\bar{\mathbf{x}}_2 \\
c &= -\frac{1}{2}\left(\log_e \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \bar{\mathbf{x}}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\bar{\mathbf{x}}_2\right) - \log_e \frac{\pi_1}{\pi_2}
\end{aligned}$$

The classification rule is:

$$\text{if } Q(\mathbf{x}) > 0 \text{ assign } \mathbf{x} \text{ to } \Pi_1$$

else assign $\mathbf{x}$ to Π_2.

Multiclass LDA

- Generalisation to multiclass data is relatively simple. . . .

Multiclass LDA

- Generalisation to multiclass data is relatively simple. . . .
- Compute the pairwise relative probabilities as before and form the discriminant (points i and j)

$$L_{ij}(\mathbf{x}) = \log_e \left(\frac{P(\Pi_i|\mathbf{x})}{P(\Pi_j|\mathbf{x}} \right) = \log_e \left(\frac{f_i(\mathbf{x})\pi_i}{f_j(\mathbf{x})\pi_j} \right)$$

Multiclass LDA

- Generalisation to multiclass data is relatively simple. . . .
- Compute the pairwise relative probabilities as before and form the discriminant (points i and j)

$$L_{ij}(\mathbf{x}) = \log_e \left(\frac{P(\Pi_i|\mathbf{x})}{P(\Pi_j|\mathbf{x}} \right) = \log_e \left(\frac{f_i(\mathbf{x})\pi_i}{f_j(\mathbf{x})\pi_j} \right)$$

- $\mathbf{x}$ is assigned to Π_i if $L_{ij} > 0$ for all $j \neq i$.
- The discriminant function between classes i and j is then

$$L_{ij}(\mathbf{x}) = \mathbf{m}_{ij}^{\mathrm{T}}\mathbf{x} + c_{ij}$$

with

$$\begin{aligned} \mathbf{m}_{ij} &= (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \text{ and} \\ c_{ij} &= -\frac{1}{2} \left(\bar{\mathbf{x}}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_j \right) + \log_e \frac{\pi_i}{\pi_j}. \end{aligned}$$

Summary

- LDA is a *probabilistic* approach to classification

Summary

- LDA is a *probabilistic* approach to classification
- Makes strong assumptions about data

Summary

- LDA is a *probabilistic* approach to classification
- Makes strong assumptions about data
- May still work if assumptions invalid

Summary

- LDA is a *probabilistic* approach to classification
- Makes strong assumptions about data
- May still work if assumptions invalid
- *Generative*: can sample for class-conditional likelihoods

Summary

- LDA is a *probabilistic* approach to classification
- Makes strong assumptions about data
- May still work if assumptions invalid
- *Generative*: can sample for class-conditional likelihoods