Machine Learning

Linear and Quadratic Discriminant Analysis

Topics of previous lectures

- ✓ Ingredients of Machine Learning
- ✓ Classification Basics
- ✓ Basic Linear Classifier
- ✓ K-Nearest Neighbours Classifier
- ✓ Naive Bayes Classifier

Topics of today's lecture

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Maximum Likelihood Estimation (MLE)

Motivation for LDA and QDA

- We can obtain an optimal classifier on training data, if we follow the **maximum a posteriori (MAP)** decision rule

$$\hat{y} = f(\mathbf{x}) = \underset{y \in \mathbb{Y}}{\operatorname{argmax}} P(Y = y | \mathbf{X} = \mathbf{x}) =$$

$$= \underset{y \in \mathbb{Y}}{\operatorname{argmax}} \frac{P(\mathbf{X} = \mathbf{x} | Y = y) \cdot P(Y = y)}{P(\mathbf{X} = \mathbf{x})}$$

  That is, predict the class that has the highest probability conditional to the given feature values.

- For 2 classes $\oplus$ and $\ominus$, for input $\mathbf{x}$ we would predict $\oplus$ if

$$P(Y = \oplus | \mathbf{X} = \mathbf{x}) > P(Y = \ominus | \mathbf{X} = \mathbf{x})$$

$$\frac{P(Y = \oplus | \mathbf{X} = \mathbf{x})}{P(Y = \ominus | \mathbf{X} = \mathbf{x})} = \frac{P(\mathbf{X} = \mathbf{x} | Y = \oplus) \cdot P(Y = \oplus)}{P(\mathbf{X} = \mathbf{x} | Y = \ominus) \cdot P(Y = \ominus)} > 1$$

Motivation for LDA and QDA

- In case of Naive Bayes, we assumed conditional independence of the features given the label and obtained

$$\hat{y} = \underset{y \in \mathbb{Y}}{\operatorname{argmax}} P(Y = y) \prod_{i=1}^{m} P(\mathbf{X}_i = \mathbf{x}_i | Y = y)$$

Motivation for LDA and QDA

- In case of Naive Bayes, we assumed conditional independence of the features given the label and obtained

$$\hat{y} = \underset{y \in \mathbb{Y}}{\operatorname{argmax}} P(Y = y) \prod_{i=1}^{m} P(\mathbf{X}_i = \mathbf{x}_i | Y = y)$$

- In case of LDA and QDA, we assume that $P(\mathbf{X} = \mathbf{x}|Y = k)$ for each class k is modeled as a multivariate Gaussian distribution with PDF

$$P(\mathbf{X} = \mathbf{x}|Y = k) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\Big),$$

where d is the number of features, that is $\mathbf{x} \in \mathbb{R}^d$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the mean vector and $\mathbf{\Sigma}_k \in \mathbb{R}^{d \times d}$ is the covariance matrix for class k. From here on, we will denote

$$f_k(\mathbf{x}) := P(\mathbf{X} = \mathbf{x}|Y = k) \quad p_k = P(Y = k).$$

Linear Discriminant Analysis (LDA)

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \Big)$$

In case of LDA, we further assume that the classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma} \quad \forall k$. When comparing two classes $\oplus$ and $\ominus$, we can look at the log-ratio to obtain the decision boundary between the classes

$$\log\left(\frac{P(Y = \oplus | \mathbf{X} = \mathbf{x})}{P(Y = \ominus | \mathbf{X} = \mathbf{x})}\right) =$$

Linear Discriminant Analysis (LDA)

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \Big)$$

In case of LDA, we further assume that the classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma} \quad \forall k$. When comparing two classes $\oplus$ and $\ominus$, we can look at the log-ratio to obtain the decision boundary between the classes

$$\log\left(\frac{P(Y = \oplus | \mathbf{X} = \mathbf{x})}{P(Y = \ominus | \mathbf{X} = \mathbf{x})}\right) = \log\left(\frac{f_\oplus(\mathbf{x}) \cdot p_\oplus}{f_\ominus(\mathbf{x}) \cdot p_\ominus}\right) =$$

Linear Discriminant Analysis (LDA)

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \Big)$$

In case of LDA, we further assume that the classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma} \quad \forall k$. When comparing two classes $\oplus$ and $\ominus$, we can look at the log-ratio to obtain the decision boundary between the classes

$$\log\left(\frac{P(Y = \oplus | \mathbf{X} = \mathbf{x})}{P(Y = \ominus | \mathbf{X} = \mathbf{x})}\right) = \log\left(\frac{f_\oplus(\mathbf{x}) \cdot p_\oplus}{f_\ominus(\mathbf{x}) \cdot p_\ominus}\right) =$$

$$= \log f_\oplus(\mathbf{x}) + \log p_\oplus - (\log f_\ominus(\mathbf{x}) + \log p_\ominus) = \log\left(\frac{f_\oplus(\mathbf{x})}{f_\ominus(\mathbf{x})}\right) + \log\left(\frac{p_\oplus}{p_\ominus}\right) =$$

Linear Discriminant Analysis (LDA)

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\Big)$$

In case of LDA, we further assume that the classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma} \quad \forall k$. When comparing two classes $\oplus$ and $\ominus$, we can look at the log-ratio to obtain the decision boundary between the classes

$$\log\left(\frac{P(Y=\oplus|\mathbf{X}=\mathbf{x})}{P(Y=\ominus|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{f_\oplus(\mathbf{x})\cdot p_\oplus}{f_\ominus(\mathbf{x})\cdot p_\ominus}\right) =$$

$$= \log f_\oplus(\mathbf{x}) + \log p_\oplus - (\log f_\ominus(\mathbf{x}) + \log p_\ominus) = \log\left(\frac{f_\oplus(\mathbf{x})}{f_\ominus(\mathbf{x})}\right) + \log\left(\frac{p_\oplus}{p_\ominus}\right) =$$

$$= \log\left(\frac{p_\oplus}{p_\ominus}\right) - \frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_\oplus)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\oplus) - (\mathbf{x}-\boldsymbol{\mu}_\ominus)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ominus)\Big) =$$

Linear Discriminant Analysis (LDA)

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\Big)$$

In case of LDA, we further assume that the classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma} \quad \forall k$. When comparing two classes $\oplus$ and $\ominus$, we can look at the log-ratio to obtain the decision boundary between the classes

$$\log\left(\frac{P(Y=\oplus|\mathbf{X}=\mathbf{x})}{P(Y=\ominus|\mathbf{X}=\mathbf{x})}\right) = \log\left(\frac{f_\oplus(\mathbf{x})\cdot p_\oplus}{f_\ominus(\mathbf{x})\cdot p_\ominus}\right) =$$

$$= \log f_\oplus(\mathbf{x}) + \log p_\oplus - (\log f_\ominus(\mathbf{x}) + \log p_\ominus) = \log\left(\frac{f_\oplus(\mathbf{x})}{f_\ominus(\mathbf{x})}\right) + \log\left(\frac{p_\oplus}{p_\ominus}\right) =$$

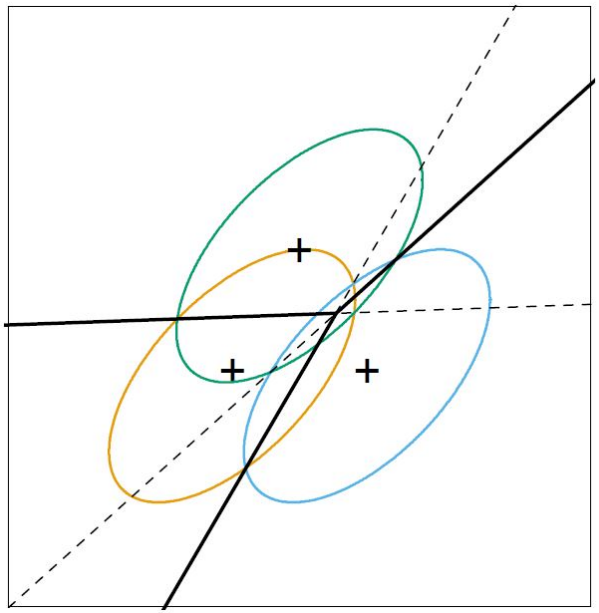
$$= \log\left(\frac{p_\oplus}{p_\ominus}\right) - \frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu}_\oplus)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\oplus) - (\mathbf{x}-\boldsymbol{\mu}_\ominus)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\ominus)\Big) =$$

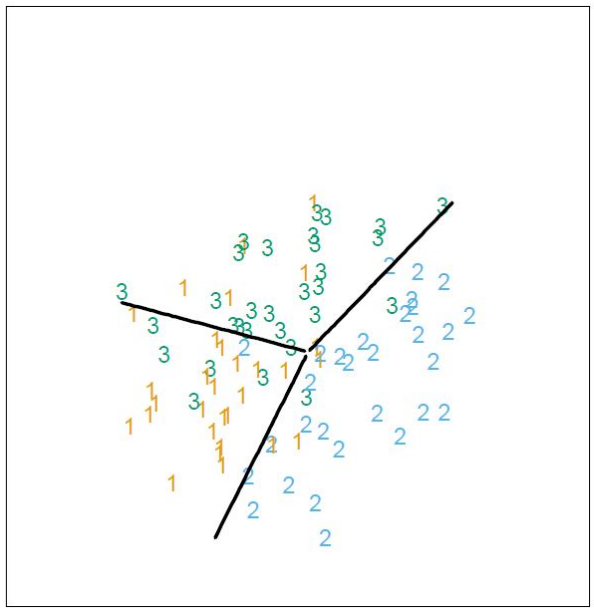
$$= \log\left(\frac{p_\oplus}{p_\ominus}\right) - \frac{1}{2}(\boldsymbol{\mu}_\oplus+\boldsymbol{\mu}_\ominus)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_\oplus-\boldsymbol{\mu}_\ominus) + \mathbf{x}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_\oplus-\boldsymbol{\mu}_\ominus) = 0$$

Linear Discriminant Analysis (LDA)

$$\log\left(\frac{p_{\oplus}}{p_{\ominus}}\right) - \frac{1}{2}(\boldsymbol{\mu}_{\oplus} + \boldsymbol{\mu}_{\ominus})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\oplus} - \boldsymbol{\mu}_{\ominus}) + \mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\oplus} - \boldsymbol{\mu}_{\ominus}) = 0$$

- The above linear function is the decision boundary between classes $\oplus$ and $\ominus$
- For more than two classes, we can obtain the pairwise decision boundaries similarly





Linear Discriminant Analysis (LDA)

- The MAP decision rule can be equivalently represented in terms of the **linear discriminant functions**

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log p_k,$$

so that

$$\hat{y} = f(\mathbf{x}) = \underset{k \in Y}{\operatorname{argmax}}\, \delta_k(\mathbf{x})$$

Linear Discriminant Analysis (LDA)

- The MAP decision rule can be equivalently represented in terms of the **linear discriminant functions**

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log p_k,$$

so that

$$\hat{y} = f(\mathbf{x}) = \underset{k \in Y}{\operatorname{argmax}}\, \delta_k(\mathbf{x})$$

- In practice we don't know the parameters of the Normal distribution and we need to **estimate** them using the training data:
  - $\hat{p}_k = \frac{N_k}{N}$, where N_k is the number of class-k observations and N is the total number of observations
  - $\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{i:Y=k} \mathbf{x}_i$
  - $\hat{\boldsymbol{\Sigma}} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i:Y=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$

Quadratic Discriminant Analysis (QDA)

- The assumption that the inputs of every class have the same covariance $\mathbf{\Sigma}$ is quite restrictive

Quadratic Discriminant Analysis (QDA)

- The assumption that the inputs of every class have the same covariance $\mathbf{\Sigma}$ is quite restrictive
- In case of QDA, we also estimate $\mathbf{\Sigma}_k$ for each class, and get **quadratic discriminant functions**

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log p_k =$$

Quadratic Discriminant Analysis (QDA)

- The assumption that the inputs of every class have the same covariance $\boldsymbol{\Sigma}$ is quite restrictive
- In case of QDA, we also estimate $\boldsymbol{\Sigma}_k$ for each class, and get **quadratic discriminant functions**

$$\begin{aligned}\delta_k(\mathbf{x}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log p_k = \\ &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \log p_k\end{aligned}$$

Quadratic Discriminant Analysis (QDA)

- The assumption that the inputs of every class have the same covariance $\boldsymbol{\Sigma}$ is quite restrictive
- In case of QDA, we also estimate $\boldsymbol{\Sigma}_k$ for each class, and get **quadratic discriminant functions**

$$\begin{aligned}\delta_k(\mathbf{x}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log p_k = \\ &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \log p_k\end{aligned}$$

- The decision boundary between each pair of classes $\oplus$ and $\ominus$ is described by a quadratic equation $\{\mathbf{x} : \delta_{\oplus}(\mathbf{x}) = \delta_{\ominus}(\mathbf{x})\}$

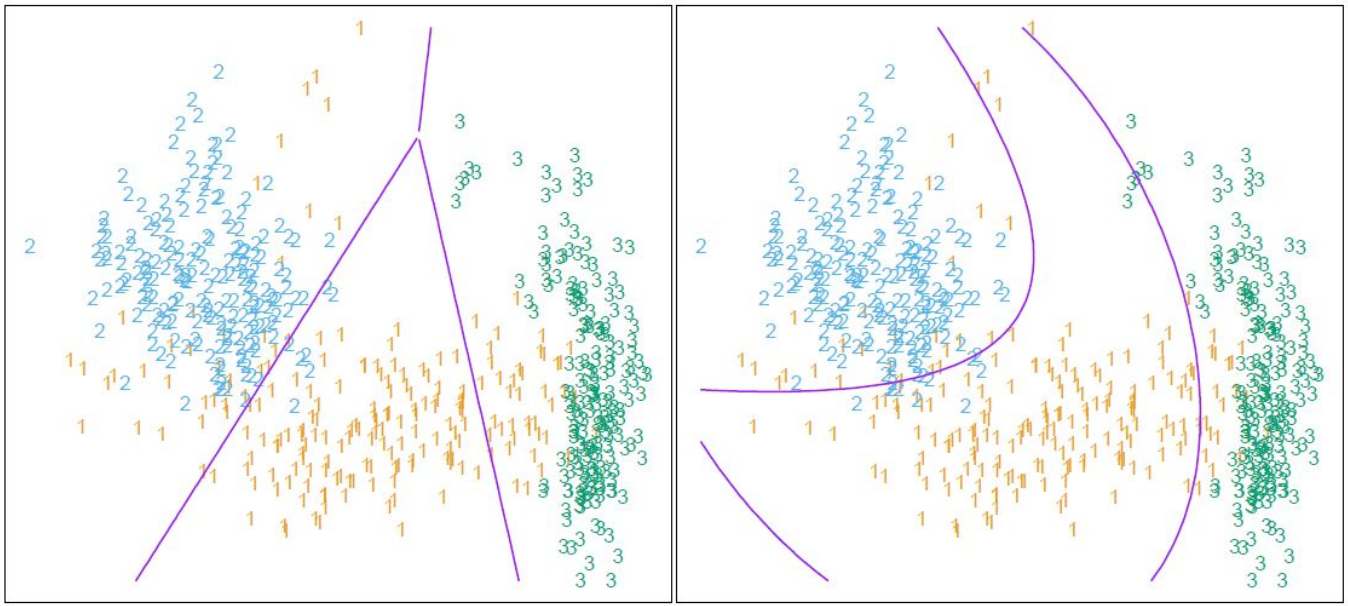
Quadratic Discriminant Analysis (QDA)

- The assumption that the inputs of every class have the same covariance $\boldsymbol{\Sigma}$ is quite restrictive
- In case of QDA, we also estimate $\boldsymbol{\Sigma}_k$ for each class, and get **quadratic discriminant functions**

$$\begin{aligned}\delta_k(\mathbf{x}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log p_k = \\ &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \log p_k\end{aligned}$$

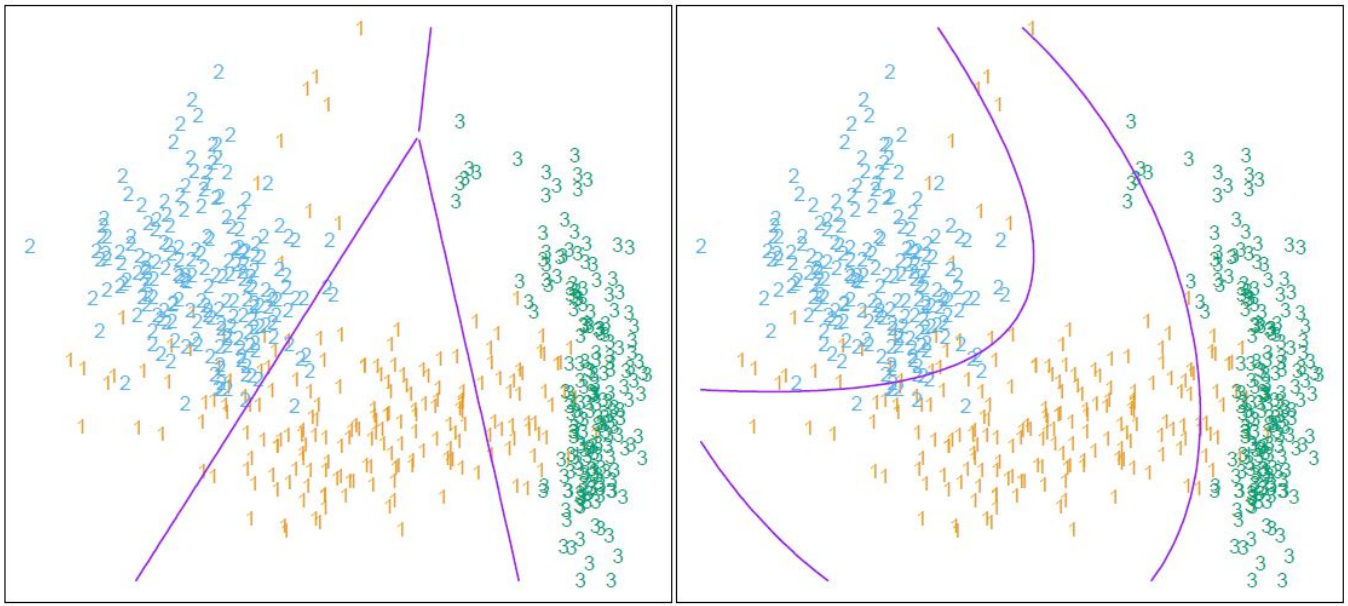
- The decision boundary between each pair of classes $\oplus$ and $\ominus$ is described by a quadratic equation $\{\mathbf{x} : \delta_{\oplus}(\mathbf{x}) = \delta_{\ominus}(\mathbf{x})\}$
- The estimates for QDA are similar to those for LDA, except that separate covariance matrices must be estimated for each class.

LDA and QDA



Pro: Provides fast classification and is easy to implement

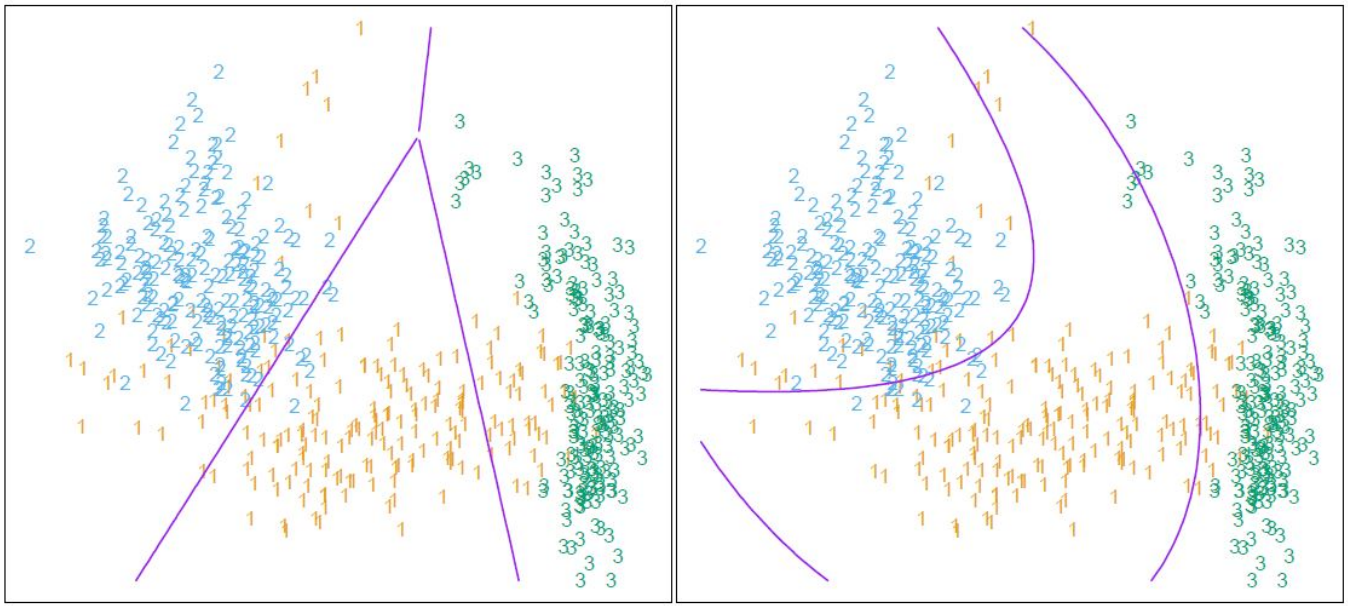
LDA and QDA



Pro: Provides fast classification and is easy to implement

Pro: LDA & QDA are often preferred when there are more than 2 labels to predict

LDA and QDA

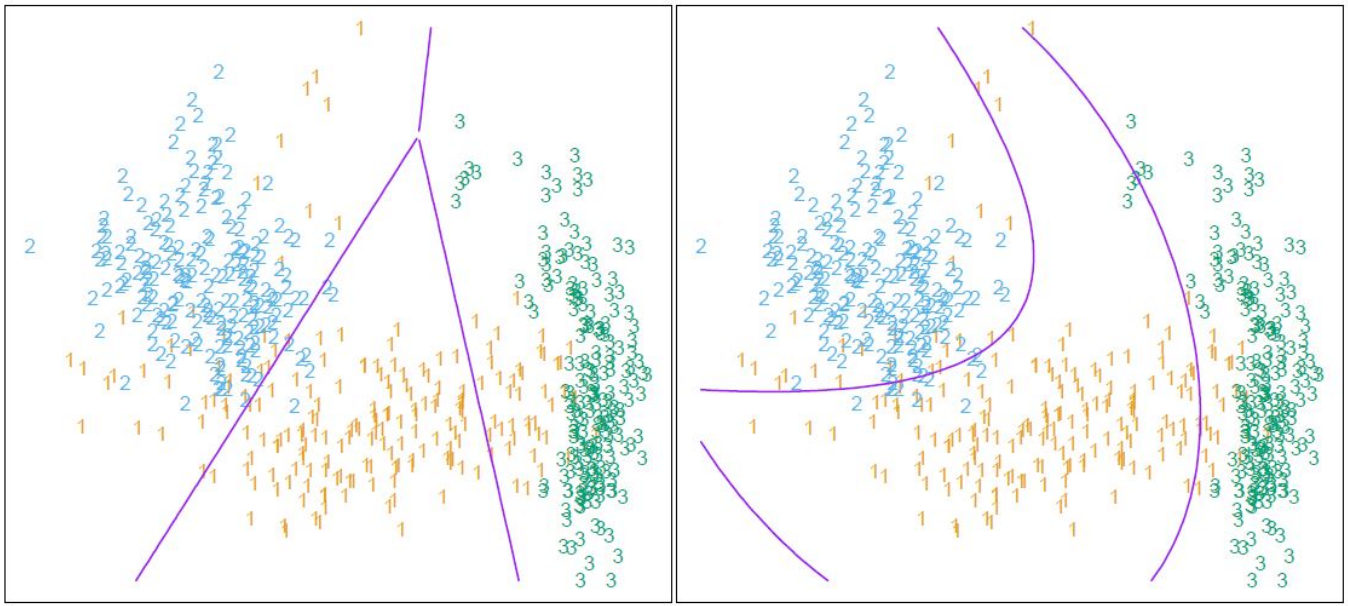


Pro: Provides fast classification and is easy to implement

Pro: LDA & QDA are often preferred when there are more than 2 labels to predict

Con: The normality assumption may not hold in our data

LDA and QDA



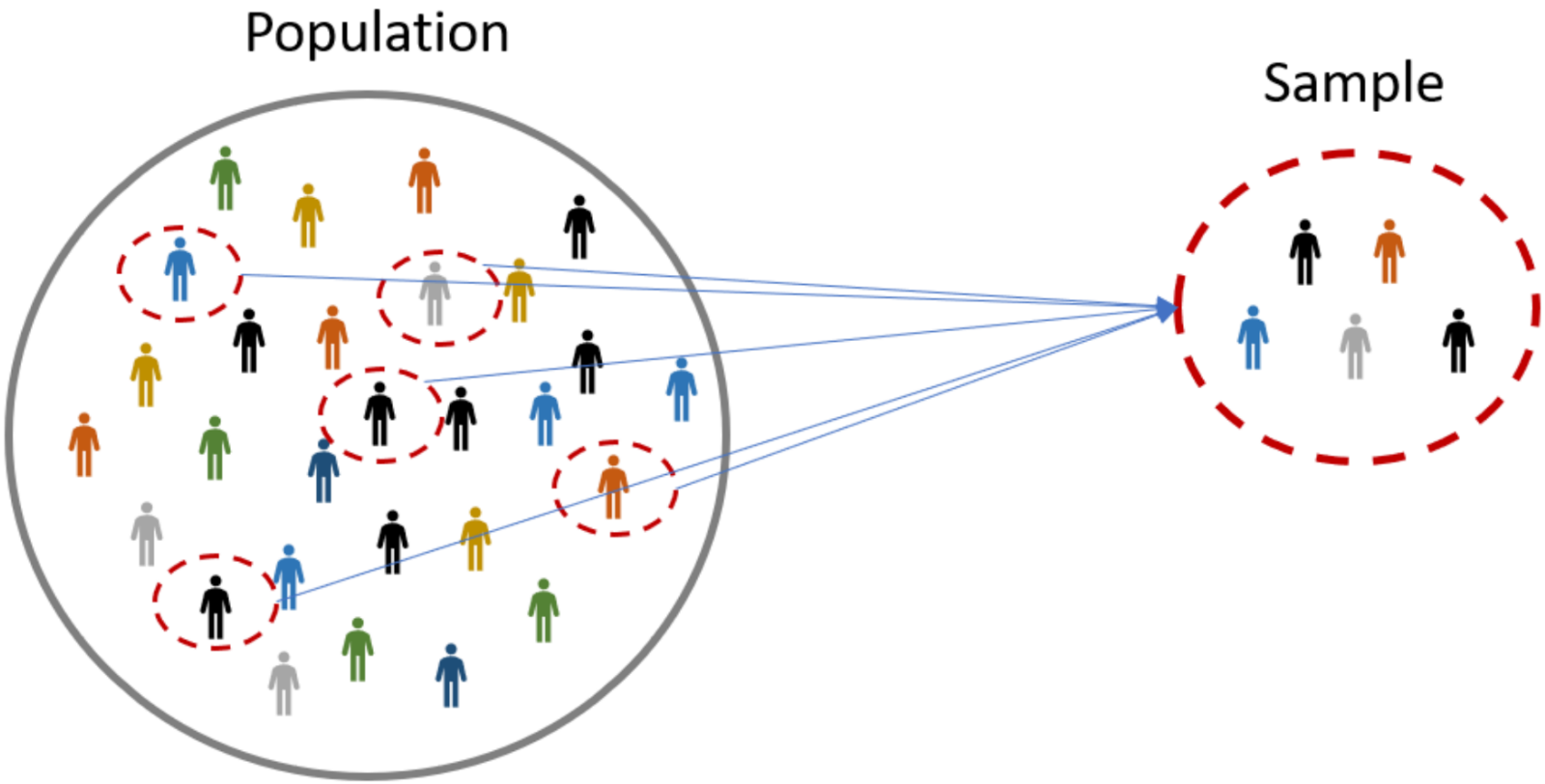
Pro: Provides fast classification and is easy to implement

Pro: LDA & QDA are often preferred when there are more than 2 labels to predict

Con: The normality assumption may not hold in our data

Con: Sensitive to class imbalance.

Population vs Sample



Maximum Likelihood Estimation (MLE)

Definition (Likelihood function)

Let $f(x_1, \ldots, x_n; \theta)$, $\boldsymbol{\theta} \in \mathbb{R}^k$ be the joint PMF or PDF of random variables $X_1, \ldots, X_n$ with sample values $x_1, \ldots, x_n$. The **likelihood function** of the sample is given by

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = L(\boldsymbol{\theta}) = f(x_1, \ldots, x_n; \boldsymbol{\theta})$$

If $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.)

Maximum Likelihood Estimation (MLE)

Definition (Likelihood function)

Let $f(x_1, \ldots, x_n; \theta)$, $\boldsymbol{\theta} \in \mathbb{R}^k$ be the joint PMF or PDF of random variables $X_1, \ldots, X_n$ with sample values $x_1, \ldots, x_n$. The **likelihood function** of the sample is given by

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = L(\boldsymbol{\theta}) = f(x_1, \ldots, x_n; \boldsymbol{\theta})$$

If $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.)

- discrete random variable with PMF $p(x, \boldsymbol{\theta})$, then

$$L(\boldsymbol{\theta}) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p(x_i, \boldsymbol{\theta})$$

Maximum Likelihood Estimation (MLE)

Definition (Likelihood function)

Let $f(x_1, \ldots, x_n; \theta)$, $\boldsymbol{\theta} \in \mathbb{R}^k$ be the joint PMF or PDF of random variables $X_1, \ldots, X_n$ with sample values $x_1, \ldots, x_n$. The **likelihood function** of the sample is given by

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = L(\boldsymbol{\theta}) = f(x_1, \ldots, x_n; \boldsymbol{\theta})$$

If $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.)

- discrete random variable with PMF $p(x, \boldsymbol{\theta})$, then

$$L(\boldsymbol{\theta}) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p(x_i, \boldsymbol{\theta})$$

- continuous r.v. with density $f(x, \boldsymbol{\theta})$, then similarly

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$$

Maximum Likelihood Estimation (MLE)

Definition (Maximum Likelihood Estimators)

The maximum likelihood estimators (MLEs) are those values of the parameters that maximize the likelihood function with respect to the parameter θ. That is,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}; x_1, \ldots, x_n)$$

Maximum likelihood estimates give the parameter values for which the observed sample is **most likely** to have been generated.

Example

- Suppose the data $x_1, x_2, \ldots, x_n$ is drawn independently from a normal distribution $N(\mu, \sigma^2)$ with unknown μ and σ

Example

- Suppose the data $x_1, x_2, \ldots, x_n$ is drawn independently from a normal distribution $N(\mu, \sigma^2)$ with unknown μ and σ
- We want to estimate these unknown parameters from the data

Example

- Suppose the data $x_1, x_2, \ldots, x_n$ is drawn independently from a normal distribution $N(\mu, \sigma^2)$ with unknown μ and σ
- We want to estimate these unknown parameters from the data
- For which values of μ and σ it is **most likely** that our data comes from the corresponding normal distribution?

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} L(\mu, \sigma) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Example

- Suppose the data $x_1, x_2, \ldots, x_n$ is drawn independently from a normal distribution $N(\mu, \sigma^2)$ with unknown μ and σ
- We want to estimate these unknown parameters from the data
- For which values of μ and σ it is **most likely** that our data comes from the corresponding normal distribution?

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} L(\mu, \sigma) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

We can obtain a more convenient, but equivalent optimization problem if we consider the **log likelihood** function

Example

- Suppose the data $x_1, x_2, \ldots, x_n$ is drawn independently from a normal distribution $N(\mu, \sigma^2)$ with unknown μ and σ
- We want to estimate these unknown parameters from the data
- For which values of μ and σ it is **most likely** that our data comes from the corresponding normal distribution?

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} L(\mu, \sigma) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

We can obtain a more convenient, but equivalent optimization problem if we consider the **log likelihood** function

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} =$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \log L(\mu, \sigma)}{\partial \sigma} =$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \log L(\mu, \sigma)}{\partial \sigma} = \frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} = 0 \Rightarrow$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \log L(\mu, \sigma)}{\partial \sigma} = \frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Example

$$\hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma) = -n \log(\sqrt{2\pi}) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

How can we find the optimal values $\hat{\mu}, \hat{\sigma}$?

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \log L(\mu, \sigma)}{\partial \sigma} = \frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

In statistical terms, we call $\hat{\mu}$ and $\hat{\sigma}$ **point estimators** or **statistics**.

Properties of Estimators

Definition (Point Estimator)

Let $x_1, x_2, \ldots, x_n$ be a set of i.i.d. data points. A **point estimator** or **statistic** is any function of the data

$$\hat{\theta} = \hat{\theta_n} = g(x_1, x_2, \ldots, x_n)$$

Properties of Estimators

Definition (Point Estimator)

Let $x_1, x_2, \ldots, x_n$ be a set of i.i.d. data points. A **point estimator** or **statistic** is any function of the data

$$\hat{\theta} = \hat{\theta_n} = g(x_1, x_2, \ldots, x_n)$$

Since the data is drawn from a random process, any function of the data is random. Therefore $\hat{\theta}$ is a random variable.

Properties of Estimators

Definition (Point Estimator)

Let $x_1, x_2, \ldots, x_n$ be a set of i.i.d. data points. A **point estimator** or **statistic** is any function of the data

$$\hat{\theta} = \hat{\theta_n} = g(x_1, x_2, \ldots, x_n)$$

Since the data is drawn from a random process, any function of the data is random. Therefore $\hat{\theta}$ is a random variable.

Definition (Bias)

The bias of an estimator is defined as

$$bias(\hat{\theta}_n) = \mathbb{E}(\hat{\theta_n}) - \theta,$$

where the expectation is taken over the data and θ is the true underlying value used to define the data-generating distribution.

Properties of Estimators

Definition (Unbiased Estimator)

An estimator $\hat{\theta_n}$ is said to be **unbiased** if $bias(\hat{\theta_n}) = 0$, which implies that $\mathbb{E}(\hat{\theta_n}) = \theta$.

Properties of Estimators

Definition (Unbiased Estimator)

An estimator $\hat{\theta_n}$ is said to be **unbiased** if $bias(\hat{\theta_n}) = 0$, which implies that $\mathbb{E}(\hat{\theta_n}) = \theta$.

Definition (Asymptotically Unbiased)

An estimator $\hat{\theta_n}$ is said to be **asymptotically unbiased** if $\lim_{n\to\infty} bias(\hat{\theta}_n) = 0$, which implies that $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$.

Properties of Estimators

Definition (Unbiased Estimator)

An estimator $\hat{\theta_n}$ is said to be **unbiased** if $bias(\hat{\theta_n}) = 0$, which implies that $\mathbb{E}(\hat{\theta_n}) = \theta$.

Definition (Asymptotically Unbiased)

An estimator $\hat{\theta_n}$ is said to be **asymptotically unbiased** if $\lim_{n\to\infty} bias(\hat{\theta}_n) = 0$, which implies that $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$.

Definition (Consistency)

An estimator is weak consistent, if $\hat{\theta}_n \xrightarrow{p} \theta \quad n \to \infty$ and strong consistent, if $\hat{\theta}_n \xrightarrow{a.s.} \theta \quad n \to \infty$.

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu =$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu =$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) - \mu = \mu - \mu = 0$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) - \mu = \mu - \mu = 0$$

The sample variance of a Gaussian distribution is biased, because

$$bias(\hat{\sigma}_n^2) = \mathbb{E}(\hat{\sigma}_n^2) - \sigma^2$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) - \mu = \mu - \mu = 0$$

The sample variance of a Gaussian distribution is biased, because

$$bias(\hat{\sigma}_n^2) = \mathbb{E}(\hat{\sigma}_n^2) - \sigma^2$$

$$\mathbb{E}(\hat{\sigma}_n^2) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2\Big) =$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) - \mu = \mu - \mu = 0$$

The sample variance of a Gaussian distribution is biased, because

$$bias(\hat{\sigma}_n^2) = \mathbb{E}(\hat{\sigma}_n^2) - \sigma^2$$

$$\mathbb{E}(\hat{\sigma}_n^2) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2\Big) = \frac{n-1}{n}\sigma^2 \text{(show it!)} \Rightarrow$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) - \mu = \mu - \mu = 0$$

The sample variance of a Gaussian distribution is biased, because

$$bias(\hat{\sigma}_n^2) = \mathbb{E}(\hat{\sigma}_n^2) - \sigma^2$$

$$\mathbb{E}(\hat{\sigma}_n^2) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2\Big) = \frac{n-1}{n}\sigma^2 \text{(show it!)} \Rightarrow$$

$$\Rightarrow bias(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Example

The MLE estimator of the Gaussian mean parameter is unbiased. Suppose $x_i \overset{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n$, then

$$bias(\hat{\mu}_n) = \mathbb{E}(\hat{\mu}_n) - \mu = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) - \mu = \mu - \mu = 0$$

The sample variance of a Gaussian distribution is biased, because

$$bias(\hat{\sigma}_n^2) = \mathbb{E}(\hat{\sigma}_n^2) - \sigma^2$$

$$\mathbb{E}(\hat{\sigma}_n^2) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2\Big) = \frac{n-1}{n}\sigma^2 \text{(show it!)} \Rightarrow$$

$$\Rightarrow bias(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

The unbiased sample variance estimator is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2.$$

What have we learned today?

- ✓ Linear Discriminant Analysis (LDA)
- ✓ Quadratic Discriminant Analysis (QDA)
- ✓ Maximum Likelihood Estimation (MLE)