1. Data preparation

1. Data types
2. Missing values
3. Outliers
4. Standardisation & normalization

2. Bayes Theorem

Let $\mathbf{x}$ denotes the feature vector, and $\{\omega_1, \omega_2, \cdots, \omega_c\}$ denotes the c classes. Then the posterior probability can be computed as follows:

$$p(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)p(\omega_j)}{p(\mathbf{x})}$$

prior probability

where

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)p(\omega_j)$$

class-conditional probability density function

posterior probability

Gaussian class-conditional probability density function:

$$p(\mathbf{x}|\omega_j) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]$$

covariance matrix of class *j*

mean vector of class *j*

Bayes decision rule:

The classifier assigns $\mathbf{x}$ to class ω_i if

$$p(\omega_i|\mathbf{x}) > p(\omega_j|\mathbf{x}) \text{ for all } j \neq i$$

Or use the variant:

$$p(\mathbf{x}|\omega_i)p(\omega_i) > p(\omega_j|\mathbf{x})p(\omega_j) \text{ for all } j \neq i$$

Naïve Bayes

In Naïve Bayes, we assume independence between features:

$$p(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)p(\omega_j)}{p(\mathbf{x})} = \frac{\prod_{i=1}^{n} p(x_i|\omega_j)p(\omega_j)}{p(\mathbf{x})}$$

If x_i follows normal distribution, then the class-conditional probability density function is as follow:

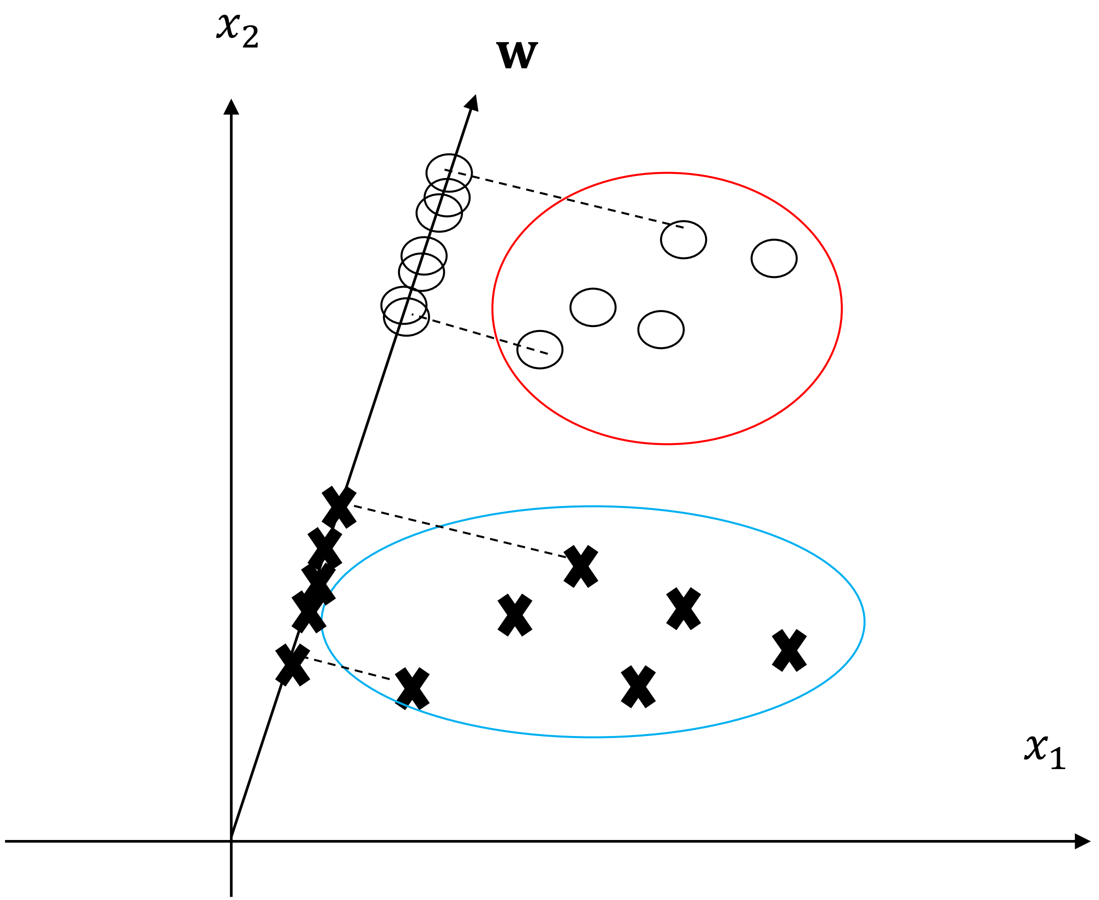
$$p(x_i|\omega_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu_{ij}}{\sigma_{ij}}\right)^2\right]$$

mean value of feature x_i for class ω_j (refer to slide 13)

standard deviation of feature x_i for class ω_j (refer to slide 12)

3. Fisher linear discriminant

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$



The Fisher linear discriminant finds such $\mathbf{w}$ that the following criterion is maximized:

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

i.e.

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

where

$$\boldsymbol{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

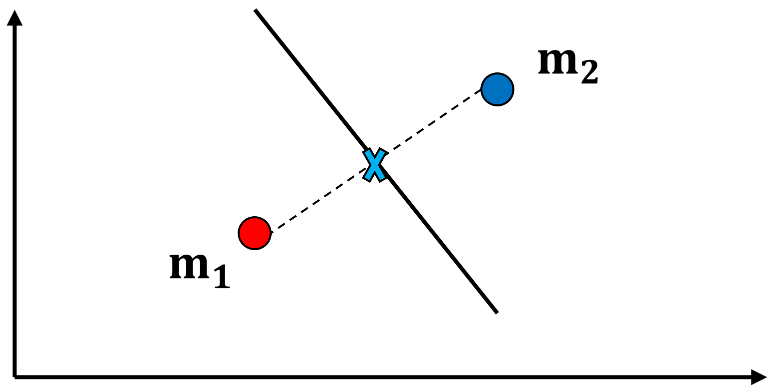
$$\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

$$\boldsymbol{S}_W = \boldsymbol{S}_\mathbf{1} + \boldsymbol{S}_\mathbf{2}$$

$\mathbf{w}$ can be found by

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

The bias or threshold w_0 is often so defined that the middle of two class means is on the hyperplane:



$$\mathbf{w}^T\mathbf{x} + w_0 = \boldsymbol{w^T}\frac{\mathbf{m}_1 + \mathbf{m}_2}{\mathbf{2}} + w_0 = 0$$

$$w_0 = -\frac{\mathbf{w}^T(\mathbf{m}_1 + \mathbf{m}_2)}{2}$$

Relationship between covariance matrix and scatter matrix

Scatter matrix

$$\mathbf{S}_{\mathbf{1}} = \sum_{\mathbf{x}_k \in D_1} (\mathbf{x}_k - \boldsymbol{\mu}_1)(\mathbf{x}_k - \boldsymbol{\mu}_1)^T$$

$$\mathbf{S}_{\mathbf{2}} = \sum_{\mathbf{x}_k \in D_2} (\mathbf{x}_k - \boldsymbol{\mu}_2)(\mathbf{x}_k - \boldsymbol{\mu}_2)^T$$

ML estimation of covariance matrix:

$$\boldsymbol{\Sigma}_{\mathbf{1}} = \frac{1}{n_1} \sum_{\mathbf{x}_k \in D_1} (\mathbf{x}_k - \boldsymbol{\mu}_1)(\mathbf{x}_k - \boldsymbol{\mu}_1)^T = \frac{1}{n_1} \mathbf{S}_1$$

$$\boldsymbol{\Sigma}_{\mathbf{2}} = \frac{1}{n_2} \sum_{\mathbf{x}_k \in D_2} (\mathbf{x}_k - \boldsymbol{\mu}_2)(\mathbf{x}_k - \boldsymbol{\mu}_2)^T = \frac{1}{n_2} \mathbf{S}_2$$

Example

σ_{11}

$$\boldsymbol{\Sigma}_1 = \frac{1}{n_1} \sum_{\mathbf{x}_k \in D_1} (\mathbf{x}_k - \boldsymbol{\mu}_1)(\mathbf{x}_k - \boldsymbol{\mu}_1)^T = \begin{bmatrix} 1.0253 & -0.0036 \\ -0.0036 & 0.8880 \end{bmatrix}$$

σ_{21}

σ_{12}

$$\boldsymbol{\Sigma}_2 = \frac{1}{n_2} \sum_{\mathbf{x}_k \in D_2} (\mathbf{x}_k - \boldsymbol{\mu}_2)(\mathbf{x}_k - \boldsymbol{\mu}_2)^T = \begin{bmatrix} 1.1884 & -0.013 \\ -0.013 & 1.0198 \end{bmatrix}$$

σ_{22}

$$\boldsymbol{\mu}_1 = \frac{1}{n_1}\sum\nolimits_{\mathbf{x}_k \in D_1} \mathbf{x}_k = \begin{bmatrix} -0.1055 \\ -0.0974 \end{bmatrix} \begin{array}{l} \mu_{11} \\ \mu_{21} \end{array}$$

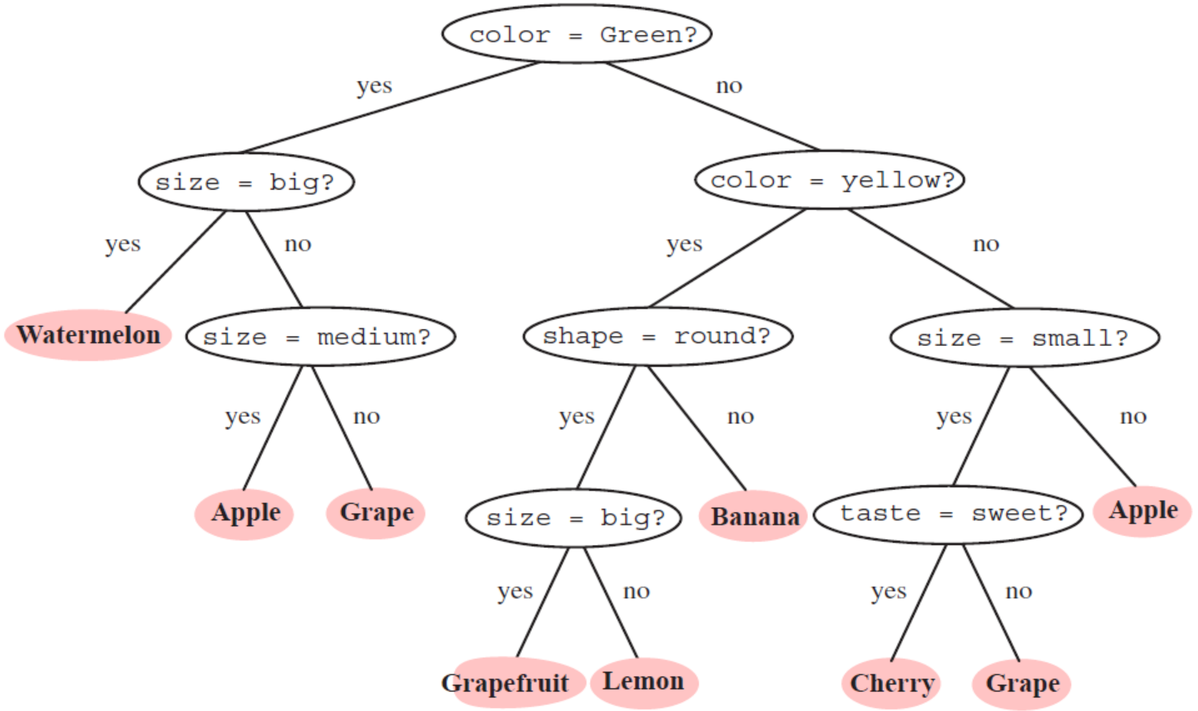
$$\boldsymbol{\mu}_2 = \frac{1}{n_2}\sum\nolimits_{\mathbf{x}_k \in D_2} \mathbf{x}_k = \begin{bmatrix} 2.0638 \\ 3.0451 \end{bmatrix} \begin{array}{l} \mu_{12} \\ \mu_{22} \end{array}$$

4. Support vector machines (SVM)

1. Margin of separation
2. Primal problem
3. Dual problem
4. Separable vs non-separable

5. Classification tree

1. Selection of attribute at a node
2. Question to ask at a node
3. Random forest



6. Regression

1. Multiple regression and OLS estimation of parameters
2. Ridge regression
3. Lasso

7. Classifier performance evaluation methods and metrics

Methods

1. Hold-out, repeated hold-out
2. K-fold cross validation
3. Repeated k-fold cross validation,
4. Leave-one-out

Metrics

1. Accuracy, error rate
2. Selectivity and specificity
3. Precision, recall, F-score
4. ROC curve and AUC

8. Feature subset selection

Two components in a feature subset selection algorithm

1. Search algorithm
2. Evaluation criterion

Methods (based on evaluation criterion)

1. Filter method
2. Wrapper method

Methods (based on search algorithm)

1. Optimal vs sub-optimal
2. Forward selection vs back elimination

9. Clustering analysis

1. Centroids-based clustering (partitioning methods)
   - ❑ K-means clustering

2. Hierarchical clustering
   - ❑ Agglomerative hierarchical clustering

3. Density-based clustering
   - ❑ DBSCAN

4. Distribution-based clustering
   - ❑ Gaussian mixture model

10. Clustering evaluation metrics

1. Silhouette coefficient
2. Dunn index
3. Davies-Bouldin index
4. Calinski-Harabasz index
5. Rand index