LEBANESE UNIVERSITY
FACULTY OF ENGINEERING III
ELECTRICAL AND ELECTRONIC DEPARTMENT

Numerical Analysis and Modeling

Spring - SEM VI

Dr. Amjad Hajjar

Spring 2024-2025

Contents

Chapter 23

Recommended Exercices

Exercice 1 (chapter 2&3)

In the adjacent dataset, the attribute X_2 is categorical with domain $\{a, b\}$.

X_1	X_2
0.3	a
-0.3	b
0.4	a
-0.6	a
0.4	a
1.2	b
-0.1	a
-1.6	b
1.6	b
-1.3	a

1. Replace the column X_2 by two binary columns $(X_2 = a)$ and $(X_2 = b)$.
2. Calculate the mean vector.
3. Calculate the centered data matrix.
4. Calculate the covariance between X_1 and $(X_2 = a)$.

Solution

Part 1: Convert $X2$ to Binary Columns

$X1$	$(X2 = a)$	$(X2 = b)$
0.3	1	0
-0.3	0	1
0.4	1	0
-0.6	1	0
0.4	1	0
1.2	0	1
-0.1	1	0
-1.6	0	1
1.6	0	1
-1.3	1	0

Part 2: Mean Vector

$$\text{Mean Vector} = \begin{pmatrix} \text{mean}(X1) \\ \text{mean}(X2 = a) \\ \text{mean}(X2 = b) \end{pmatrix} = \begin{pmatrix} 0.0 \\ 0.6 \\ 0.4 \end{pmatrix}$$

Part 3: Centered Data Matrix

$X1 - 0.0$	$(X2 = a) - 0.6$	$(X2 = b) - 0.4$
0.3	0.4	−0.4
−0.3	−0.6	0.6
0.4	0.4	−0.4
−0.6	0.4	−0.4
0.4	0.4	−0.4
1.2	−0.6	0.6
−0.1	0.4	−0.4
−1.6	−0.6	0.6
1.6	−0.6	0.6
−1.3	0.4	−0.4

Part 4: Covariance between $X1$ and $(X2 = a)$

The covariance between $X1$ and $(X2 = a)$ is calculated as:

$$\text{Cov}(X1, X2 = a) = \frac{1}{n-1}\sum_{i=1}^{n}(X1_i - \text{mean}(X1)) \cdot ((X2 = a)_i - \text{mean}(X2 = a)) = -0.01$$

Exercice 2 (chapter 3)

In the adjacent table, we discrete X_1 into three bins:

- Negative: $X_1 \leq -0.5$
- Neutral: $-0.5 < X_1 < 0.5$
- Positive: $0.5 \leq X_1$

X_1	X_2
0.3	a
-0.3	b
0.4	a
-0.6	a
0.4	a
1.2	b
-0.1	a
-1.6	b
1.6	b
-1.3	a

1. Construct the new dataset with two categorical attributes (2 columns).
2. Construct the contingency table.
3. Construct the table frequencies e_{ij} expected under the independent hypothesis.
4. Calculate the number of degrees of freedom (q) and the value of the χ^2 statistic.

Solution

Part 1: Discretize $X1$ and Create New Categorical Dataset

$X1$ (Category)	$X2$
Neutral	a
Neutral	b
Neutral	a
Negative	a
Neutral	a
Positive	b
Neutral	a
Negative	b
Positive	b
Negative	a

Part 2: Contingency Table

	$X2 = a$	$X2 = b$
Negative	2	1
Neutral	4	1
Positive	0	2

Part 3: Expected Frequencies Table e_{ij}

	$X2 = a$	$X2 = b$
Negative	1.8	1.2
Neutral	3.0	2.0
Positive	1.2	0.8

Part 4: Degrees of Freedom and χ^2 Statistic

The degrees of freedom are:

$$q = (3-1)(2-1) = 2$$

The χ^2 statistic is calculated as:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 3.888$$

where n_{ij} are the observed frequencies and e_{ij} are the expected frequencies.

Exercice 3 (chapter 13)

Consider the following set of unidimensional points: $\{1, 1, 2, 3, 5, 8, 13, 21, 33, 54\}$.

1. Divide this set into two clusters using k-means with $k = 2$ and initial centroids 2 and 13.
2. Calculate the variance of each cluster.

   Suppose that the sample comes from a Gaussian mixture with two components, which are the above-calculated clusters.

3. Give the expression of the probability density function of the Gaussian mixture.

4. According to Bayes rule, calculate the probability of membership of the number 10 to each of the clusters.

Solution

Part 1: k-Means Clustering

Given points: $\{1, 1, 2, 3, 5, 8, 13, 21, 33, 54\}$.

1. Initial centroids: $\mu_1 = 2$ and $\mu_2 = 13$.
2. Assignments: Assign each point to the nearest centroid and recompute the centroids until convergence.

Finally we get two clusters:

- Cluster 1: $\{1\ 1\ 2\ 3\ 5\ 8\ 13\}$ with centroid $\mu_1 = 4.714$
- Cluster 2: $\{21\ 33\ 54\}$ with centroid $\mu_2 = 36$

Part 2: Variance of Each Cluster

The variance σ^2 for each cluster is given by:

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

where μ is the mean of the cluster and n is the number of points in the cluster.

We get:

- $\sigma_1^2 = 16.775$
- $\sigma_2^2 = 186$

Part 3: Gaussian Mixture Model Probability Density Function

The probability density function (PDF) of the Gaussian mixture model is:

$$f(x) = \pi_1 \cdot \mathcal{N}(x; \mu_1, \sigma_1^2) + \pi_2 \cdot \mathcal{N}(x; \mu_2, \sigma_2^2)$$

where:
- π_i is the prior probability of each cluster.
- $\mathcal{N}(x; \mu, \sigma^2)$ is the Gaussian PDF:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Part 4: Probability of Membership for 10 Using Bayes' Rule

Using Bayes' rule, the probability of membership for the number 10 is:

$$P(\text{Cluster } i | x = 10) = \frac{\pi_i \cdot \mathcal{N}(10; \mu_i, \sigma_i^2)}{\sum_{j=1}^{2} \pi_j \cdot \mathcal{N}(10; \mu_j, \sigma_j^2)}$$

After caculation we get:

- $P(\text{Cluster } 1 | x = 10) = 0.954$
- $P(\text{Cluster } 2 | x = 10) = 0.046$

Exercice 4 (chapter 2&13)

1. What is an outlier? What is a robust statistic?
2. If we run k-means on the same dataset with the same k but different initial centroids, will the algorithm necessarily converge to the same result? Why?

Solution

Part 1: Outlier and Robust Statistic

- **Outlier:** An outlier is a data point that significantly deviates from other observations in the dataset. It may result from variability in the data or experimental error. Outliers can distort statistical analyses, affecting metrics like the mean and variance.
- **Robust Statistic:** A robust statistic is a measure that is relatively unaffected by outliers or small deviations from assumptions. Robust statistics, such as the median or interquartile range (IQR), provide reliable insights even when the data includes outliers.

Part 2: Will k-Means Always Converge to the Same Result with Different Initial Centroids?

No, k-means may not always converge to the same result if we use different initial centroids. This is because:

- k-means is sensitive to initial centroids, and different starting points can lead the algorithm to converge to different local minima.
- The algorithm minimizes the sum of squared distances within clusters, which can lead to convergence at a local minimum rather than a global minimum, depending on the initial centroids chosen.

Exercice 5 (chapter 2&7&13)

True/False Questions

1. A biased statistic is weakly affected by outliers. T F
2. A robust statistic is weakly affected by outliers. T F
3. Mean is biased. T F
4. Mean is robust. T F
5. Median is robust. T F
6. Mode is robust. T F
7. Variance is robust. T F
8. IQR is robust. T F
9. Covariance depends on the units of measure. T F
10. Correlation depends on the units of measure. T F

11. Normalization reduces the effect of units of measure. T F
12. Covariance matrix is always diagonal. T F
13. Covariance matrix is always symmetric. T F
14. The first principal component is the attribute that that has the greatest variance. T F
15. The covariance between the principal components is null. T F
16. The axes of the multivariable normal distribution ellipsoid are parallel to the reference axes. T F
17. The axes of the multivariable normal distribution ellipsoid are parallels to the principal components axes. T F
18. The objective of K-means is to maximize the likelihood. T F
19. The objective of EM is to minimize the SSE. T F
20. Both K-means and EM give results that depend on the initial choice of centroids. T F

Solution

1. A biased statistic is weakly affected by outliers. **False**
2. A robust statistic is weakly affected by outliers. **True**
3. Mean is biased. **False**
4. Mean is robust. **False**
5. Median is robust. **True**
6. Mode is robust. **True**
7. Variance is robust. **False**
8. IQR is robust. **True**
9. Covariance depends on the units of measure. **True**
10. Correlation depends on the units of measure. **False**
11. Normalization reduces the effect of units of measure. **True**
12. Covariance matrix is always diagonal. **False**
13. Covariance matrix is always symmetric. **True**
14. The first principal component is the attribute that has the greatest variance. **True**
15. The covariance between the principal components is null. **True**
16. The axes of the multivariable normal distribution ellipsoid are parallel to the reference axes. **False**

17. The axes of the multivariable normal distribution ellipsoid are parallel to the principal component axes. **True**
18. The objective of k-means is to maximize the likelihood. **False**
19. The objective of EM is to minimize the SSE. **False**
20. Both k-means and EM give results that depend on the initial choice of centroids. **True**

Exercice 6 (chapter 2&7)

For the following dataset

1. Calculate the centered data matrix.
2. Calculate the covariance matrix.
3. Determine the eigenvalues and the first principal component.
4. Calculate the projection of the vector (1,1) on the direction of the principal component.

X_1	X_2
2	1
0	1
2	-1
0	-2
1	1

Solution

Part 1: Centered Data Matrix

Calculate the mean of $X1$ and $X2$:

$$\text{mean}(X1) = 1, \quad \text{mean}(X2) = 0$$

The centered data matrix is:

$$Z = \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 1 & -1 \\ -1 & -2 \\ 0 & 1 \end{pmatrix}$$

Part 2: Covariance Matrix

The covariance matrix Σ is calculated as:

$$\Sigma = \frac{1}{n} Z^T Z$$

After performing the matrix multiplication and dividing by 5, we obtain the covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{X1}^2 & \sigma_{X1,X2} \\ \sigma_{X1,X2} & \sigma_{X2}^2 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 1.6 \end{pmatrix}$$

Part 3: Eigenvalues and First Principal Component

The eigenvalues of the covariance matrix are calculated by solving the characteristic polynomial:

$$\det(\Sigma - \lambda I) = 0$$

The largest eigenvalue corresponds to the first principal component. The eigenvector associated with this eigenvalue is the direction of maximum variance in the data.

From the calculation we get: $\lambda_1 = 1.64$ and $\lambda_2 = 0.752$
and the first principal component is $v = (0.223, 0.947)$

Part 4: Projection of Vector $(1, 1)$ onto the First Principal Component

To project the vector $(1, 1)$ onto the direction of the first principal component we should calculate the following dot product:

$$\text{Projection} = (1, 1) \cdot \text{Principal Component Vector} = 1.1708$$

note that we didn't divide by the norm of the first principal component because it is already unitary.

Exercice 7 (chapter 3)

The following dataset has nine data points and two attributes: We discretize X_1 into three equal-frequency bins and call them (Low, Medium, High).

X_1	X_2
-5	A
-3	B
-1	A
0	A
1	A
1	A
2	B
2	B
3	B

1. Show the new dataset with two categorical attributes.
2. Construct the contingency table.
3. Construct the table of expected frequencies under the independence hypothesis.
4. Calculate the value of the χ^2 statistic.

Solution

Part 1: Discretize $X1$ and Create New Categorical Dataset

$X1$ (Category)	$X2$
Low	A
Low	B
Low	A
Medium	A
Medium	A
Medium	A
High	B
High	B
High	B

Part 2: Contingency Table

	$X2 = A$	$X2 = B$
Low	2	1
Medium	3	0
High	0	3

Part 3: Expected Frequencies Table e_{ij}

The expected frequencies under the independence hypothesis are:

	$X2 = A$	$X2 = B$
Low	1.67	1.33
Medium	1.67	1.33
High	1.67	1.33

Part 4: χ^2 Statistic

The χ^2 statistic is calculated as:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 6.3$$

where n_{ij} are the observed frequencies and e_{ij} are the expected frequencies.

Exercice 8 (chapter 13)

Consider the following set of numbers: {1, 2, 2, 2, 3, 4, 5, 5, 5, 6}

1. Apply k-means with $k = 2$ and initial centroids 1 and 4.

   Consider applying an iteration of EM with the following two clusters having the same prior probability: $C_1 : \{\mu_1 = 2, \sigma_1 = 1\}$ and $C_2 : \{\mu_2 = 5, \sigma_2 = 1\}$.

2. Calculate the membership probabilities.

   To simplify the calculation, we assume that in the normal distribution,

   $$|\mathbf{x} - \boldsymbol{\mu}| = 2.5\,\boldsymbol{\sigma} \implies \mathrm{N}(\mathbf{x}) = 0$$

3. Calculate the new clusters parameters.

Solution

Part 1: k-Means Clustering with Initial Centroids 1 and 4

Given points: $\{1, 2, 2, 2, 3, 4, 5, 5, 5, 6\}$.

1. Initial centroids: $\mu_1 = 1$ and $\mu_2 = 4$.
2. Assignments: Assign each point to the nearest centroid and recompute the centroids until convergence.

We get:

- Cluster 1: { 1 2 2 2 3 } with centroid $\mu_1 = 2$

- Cluster 2: { 4 5 5 5 6 } with centroid $\mu_2 = 5$

Part 2: E-step of the EM Algorithm

Given:

- $C_1 : \mu_1 = 2, \sigma_1 = 1$
- $C_2 : \mu_2 = 5, \sigma_2 = 1$

1. E-Part: Calculate the posterior probabilities $P(C_i|x_j)$ for each data point x_j:

$$P(C_i|x_j) = \frac{\pi_i \cdot N(x_j|\mu_i, \sigma_i^2)}{\sum_{k=1}^{2} \pi_k \cdot N(x_j|\mu_k, \sigma_k^2)}$$

where:
- π_i is the prior probability (assumed equal for both clusters),
- $N(x|\mu, \sigma^2)$ is the Gaussian probability density function:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

2. For data points where $|x - \mu| > 2.5\sigma$, we set $N(x) = 0$ to simplify calculations.

We get:

	C1	**C2**
$P(C_j\|1)$	0.999999993	0.000000007
$P(C_j\|2)$	0.999986993	0.000013007
$P(C_j\|3)$	0.977022630	0.022977370
$P(C_j\|4)$	0.022977370	0.977022630
$P(C_j\|5)$	0.000013007	0.999986993
$P(C_j\|6)$	0.000000007	0.999999993

Part 3: M-step of the EM Algorithm

Cluster 1

- **Mean:** $\mu_1 = \frac{\sum_{j=1}^{n} w_{1j} \cdot x_j}{\sum_{j=1}^{n} w_{1j}} = 2.0046$
- **Standard Deviation:** $\sqrt{\sigma_1^2} = \sqrt{\frac{\sum_{j=1}^{n} w_{1j} \left(x_j - \mu_1\right)^2}{\sum_{j=1}^{n} w_{1j}}} = \sqrt{0.6433} = 0.802$
- **Prior:** $\pi_1 = P\left(C_1\right) = \frac{\sum_{j=1}^{n} w_{1j}}{n} = 0.5$

Cluster 2

- **Mean:** $\mu_2 = \dfrac{\sum_{j=1}^{n} w_{2j} \cdot x_j}{\sum_{j=1}^{n} w_{2j}} = 4.9954$
- **Standard Deviation:** $\sqrt{\sigma_2^2} = \sqrt{\dfrac{\sum_{j=1}^{n} w_{2j} \left(x_j - \mu_2\right)^2}{\sum_{j=1}^{n} w_{2j}}} = \sqrt{0.6433} = 0.802$
- **Prior:** $\pi_2 = P\left(C_2\right) = \dfrac{\sum_{j=1}^{n} w_{2j}}{n} = 0.5$

Exercice 9 (chapter 14)

The following distance matrix shows the distances between four data points. Apply agglomerative clustering and show the dendrograms in each of the following cases:

1. Single Link.
2. Complete Link.
3. Mean distance (between centroids)

	B	C	D	E
A	2	8	8	7
B		7	6	6
C			3	5
D				4

Solution

Distance Matrix

Given distance matrix:

	B	C	D	E
A	2	8	8	7
B		7	6	6
C			3	5
D				4

Part 1: Agglomerative Clustering (Single Link)

Using the Single Link method:

- Identify the closest pair of clusters at each step.
- Merge the clusters with the minimum distance between any pair of points.
- Update the distance matrix until all points are in a single cluster.

Part 2: Agglomerative Clustering (Complete Link)

Using the Complete Link method:

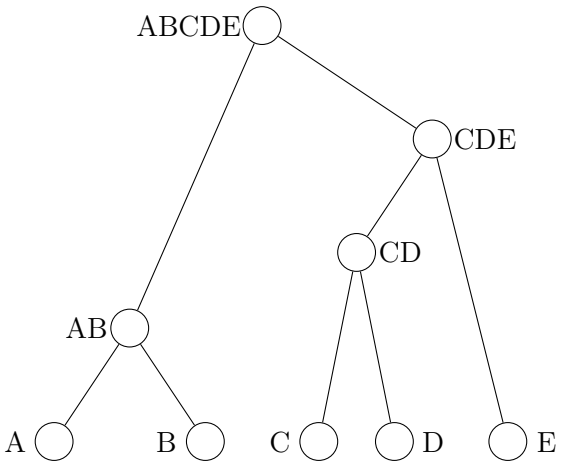
- Identify the farthest pair of clusters at each step.
- Merge clusters based on the maximum distance between any pair of points.
- Update the distance matrix iteratively.

Part 3: Agglomerative Clustering (Mean Distance / Centroid)

Using the Mean Distance method:

- Calculate the centroid of each cluster.
- Merge clusters based on the average distance between all points in each cluster.
- Update centroids and distances until a single cluster is formed.

In all three cases, we got the same dendrogram



Exercice 10 (chapter 15&17)

Consider the following eight points in $\mathbb{R}^2$:

$$A_1 = (2, 10) \quad A_2 = (2, 5) \quad A_3 = (8, 4) \quad A_4 = (5, 8)$$
$$A_5 = (7, 5) \quad A_6 = (6, 4) \quad A_7 = (1, 2) \quad A_8 = (4, 9)$$

1. Draw a table of the Euclidean distances between the points.
2. Apply DBSCAN on this dataset using $\varepsilon = 2$ and $MinPts = 2$. Indicate which points are core points, which are edge points and which outliers (if any) are. Show the resulting clusters.
3. Calculate the BetaCV measure of the resulted clustering.

   ps: $BetaCV = \dfrac{W_{in}/N_{in}}{W_{out}/N_{out}}$

Solution

Part 1: Euclidean Distance Table

The Euclidean distance d between points (x_1, y_1) and (x_2, y_2) is:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The distance table between the points $A1$ through $A8$ is as follows:

	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$	$A8$
$A1$	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}	d_{17}	d_{18}
$A2$		d_{23}	d_{24}	d_{25}	d_{26}	d_{27}	d_{28}
$A3$			d_{34}	d_{35}	d_{36}	d_{37}	d_{38}
$A4$				d_{45}	d_{46}	d_{47}	d_{48}
$A5$					d_{56}	d_{57}	d_{58}
$A6$						d_{67}	d_{68}
$A7$							d_{78}

We get the following table

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0.00	5.00	8.49	3.61	7.07	7.21	8.06	2.24
A2		0.00	6.08	4.24	5.00	4.12	3.16	4.47
A3			0.00	5.00	1.41	2.00	7.28	6.40
A4				0.00	3.61	4.12	7.21	1.41
A5					0.00	1.41	6.71	5.00
A6						0.00	5.39	5.39
A7							0.00	7.62
A8								0.00

Part 2: DBSCAN Clustering with $\varepsilon = 2$ and MinPts $= 2$

1. Calculate the ε-neighborhood for each point.
2. Identify core points (points with at least 2 neighbors), edge points, and outliers.
3. Cluster the points based on connectivity within ε-neighborhoods.

We get the following two clusters:

- Cluster 1: {A3, A5, A6}
- Cluster 2: {A4, A8}

The points are distributed as follows:

- Core Points: {A3, A4, A5, A6, A8}
- Border Points: { }
- Noise Points: {A1, A2, A7}

Part 3: Calculate BetaCV Measure

The BetaCV measure is calculated as:

$$\text{BetaCV} = \frac{\text{Win/Nin}}{\text{Wout/Nout}}$$

where:

- Win: Sum of intra-cluster distances.
- Nin: Number of intra-cluster point pairs.
- Wout: Sum of inter-cluster distances.
- Nout: Number of inter-cluster point pairs.

We get the following:

- $N_{in} = 4.0$
- $W_{in} = 6.2426$
- $N_{out} = 24.0$
- $W_{out} = 132.4715$

Finally we get

- $BetaCV = 0.2827$

Exercice 11 (chapter 19)

The following table is a training dataset for a binary classifier with two classes: Positive (+) and negative (-). Attributes a_1 and a_2 are categorical T/F and a_3 is numerical. We are envisaging a decision tree classifier. We want to find the best split for the first iteration.

Entity	a_1	a_2	a_3	Class
1	T	T	1	+
2	T	T	5	-
3	T	F	5	-
4	F	F	5	+
5	F	T	5	-
6	F	T	3	+
7	F	F	5	+
8	T	F	5	-
9	F	T	5	-

1. Calculate the entropy of the whole dataset.
2. Calculate the information gain if we split by a_1.
3. Calculate the information gain if we split by a_2.
4. Calculate the information gain for every possible split by a_3.
5. Conclude with the best split.

ps: $Entropy = \sum P(c_i) \log_2 P(c_i)$

Solution

Formulas

Entropy

The entropy H of a set of probabilities $\{p_1, p_2, \ldots, p_n\}$ is given by:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Information Gain

The information gain IG for an attribute A is defined as:

$$IG(A) = H(T) - \sum_{v \in Values(A)} \frac{|T_v|}{|T|} H(T_v)$$

where $H(T)$ is the entropy of the entire set T, and $H(T_v)$ is the entropy of subset T_v for which attribute A has value v.

Calculations

Entropy of the Whole Dataset

First, we calculate the entropy of the entire dataset:

$$H(T) = -\left(\frac{4}{9}\log_2\left(\frac{4}{9}\right) + \frac{5}{9}\log_2\left(\frac{5}{9}\right)\right) \approx 0.991$$

Information Gain for a_1

$$H(T_{a_1=T}) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) \approx 0.811$$

$$H(T_{a_1=F}) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) \approx 0.971$$

$$IG(a_1) = 0.991 - \left(\frac{4}{9}\cdot 0.811 + \frac{5}{9}\cdot 0.971\right) \approx 0.091$$

Information Gain for a_2

$$H(T_{a_2=T}) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) \approx 0.811$$

$$H(T_{a_2=F}) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) \approx 0.971$$

$$IG(a_2) = 0.991 - \left(\frac{4}{9}\cdot 0.811 + \frac{5}{9}\cdot 0.971\right) \approx 0.007$$

Information Gain for $a3$

For the threshold $a3 \leq 2.0$:

$$IG(a3 \leq 2.0) = H(\text{Total}) - \left(\frac{1}{9}H(\leq 2.0) + \frac{8}{9}H(> 2.0)\right)$$

$$IG(a3 \leq 2.0) \approx 0.1427$$

For the threshold $a3 \leq 4.0$:

$$IG(a3 \leq 4.0) = H(\text{Total}) - \left(\frac{2}{9}H(\leq 4.0) + \frac{7}{9}H(> 4.0)\right)$$

$$IG(a3 \leq 4.0) \approx 0.3198$$

Results

- Information gain for split by $a1$: 0.0911

- Information gain for split by $a2$: 0.0072
- Information gain for splits by $a3$:
  - $a3 \leq 2.0$: 0.1427
  - $a3 \leq 4.0$: 0.3198

The best split is $a3 \leq 4.0$ with an information gain of 0.3198.

Exercice 12 (chapter 22)

Suppose that we use the same dataset for testing and evaluating some classifier. When tested, the classifier predicts the entities 1, 2, 3 and 4 as positive, the others as negative.

1. Draw the confusion matrix.
2. Calculate the sensitivity (aka True Positive Rate).
3. Calculate the specificity (aka True Negative Rate).

Solution

- True Positive (TP): Predicted positive and actually positive
- False Positive (FP): Predicted positive but actually negative
- True Negative (TN): Predicted negative and actually negative
- False Negative (FN): Predicted negative but actually positive

Given the predicted and actual classes:

- TP = 2 (Entities 1 and 4)
- FN = 2 (Entities 6 and 7)
- FP = 2 (Entities 2 and 3)
- TN = 3 (Entities 5, 8, and 9)

Part 1: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	$TP = 2$	$FN = 2$
Actual Negative	$FP = 2$	$TN = 3$

Part 2: Sensitivity (True Positive Rate)

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{2}{2+2} = 0.5$$

Part 3: Specificity (True Negative Rate)

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{3}{3+2} = 0.6$$

Exercice 13 (chapter 18)

Suppose that we build a Naive Bayes classifier for the same dataset of the previous question. Calculate the posterior probabilities of the point (T, F, 3.5) and classify it.

Entity	a_1	a_2	a_3	Class
1	T	T	1	+
2	T	T	5	-
3	T	F	5	-
4	F	F	5	+
5	F	T	5	+
6	F	T	3	-
7	F	F	5	+
8	T	F	5	-
9	F	T	5	-

Solution

Step 1: Calculate Priors

$$P(+) = \frac{4}{9} \approx 0.444$$

$$P(-) = \frac{5}{9} \approx 0.556$$

Step 2: Calculate Probabilities for a_1 and a_2

For a_1:

- Positive (+) class:

$$P(T|+) = \frac{1}{4} = 0.25$$

$$P(F|+) = \frac{3}{4} = 0.75$$

- Negative (-) class:

$$P(T|-) = \frac{3}{5} = 0.6$$

$$P(F|-) = \frac{2}{5} = 0.4$$

For a_2:

- Positive (+) class:

$$P(T|+) = \frac{2}{4} = 0.5$$

$$P(F|+) = \frac{2}{4} = 0.5$$

- Negative (-) class:

$$P(T|-) = \frac{3}{5} = 0.6$$

$$P(F|-) = \frac{2}{5} = 0.4$$

Step 3: Calculate Likelihoods for a_3

Using Gaussian likelihood:

$$P(a3 = 3.5|+) = \frac{1}{\sqrt{2\pi \cdot 4}} e^{-\frac{(3.5-4)^2}{2 \cdot 4}} \approx 0.193$$

$$P(a3 = 3.5|-) = \frac{1}{\sqrt{2\pi \cdot 0.8}} e^{-\frac{(3.5-4.6)^2}{2 \cdot 0.8}} \approx 0.209$$

Step 4: Combine Priors and Likelihoods

$$P(+|X) \propto P(+) \cdot P(a_1 = T|+) \cdot P(a_2 = F|+) \cdot P(a3 = 3.5|+)$$
$$P(-|X) \propto P(-) \cdot P(a_1 = T|-) \cdot P(a_2 = F|-) \cdot P(a3 = 3.5|-)$$

$$P(+|X) \propto 0.444 \cdot 0.25 \cdot 0.5 \cdot 0.193 \approx 0.0107$$
$$P(-|X) \propto 0.556 \cdot 0.6 \cdot 0.4 \cdot 0.209 \approx 0.027$$

Step 5: Normalize to Get Posterior Probabilities

$$P(+|X) = \frac{0.0107}{0.0107 + 0.027} \approx 0.284$$
$$P(-|X) = \frac{0.027}{0.0107 + 0.027} \approx 0.716$$

Thus, the point (T, F, 3.5) is more likely to belong to the positive (-) class, with a probability of around 71.6%.

Exercice 14 (chapter 13)

Consider the following eight 2D points:

$A_1 = (2, 10)$ $A_2 = (2, 5)$ $A_3 = (8, 4)$
$A_4 = (5, 8)$ $A_5 = (7, 5)$ $A_6 = (6, 4)$
$A_7 = (1, 2)$ $A_8 = (4, 9)$

The adjacent table gives you the Euclidean distances between these points.

Starting with the centroids A_1, A_4 and A_7, apply two iterations of the K-means algorithm.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A_2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A_3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A_4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A_5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A_6						0	$\sqrt{29}$	$\sqrt{29}$
A_7							0	$\sqrt{58}$
A_8								0

Solution

Initial centroids are:

$$C_1 = A_1 = (2, 10)$$
$$C_2 = A_4 = (5, 8)$$
$$C_3 = A_7 = (1, 2)$$

Iteration 1

New centroids after first iteration:

$$C_1 = (2, 10) \quad C_2 = (6, 6) \quad C_3 = (1.5, 3.5)$$

Clusters:

- Cluster 1: $[(2, 10)]$
- Cluster 2: $[(8, 4), (5, 8), (7, 5), (6, 4), (4, 9)]$
- Cluster 3: $[(2, 5), (1, 2)]$

Iteration 2

New centroids after second iteration:

$$C_1 = (3, 9.5) \quad C_2 = (6.5, 5.25) \quad C_3 = (1.5, 3.5)$$

Clusters:

- Cluster 1: $[(2, 10), (4, 9)]$
- Cluster 2: $[(8, 4), (5, 8), (7, 5), (6, 4)]$
- Cluster 3: $[(2, 5), (1, 2)]$

Exercice 15 (chapter 14)

The adjacent table shows the distances between pairs of data points. Apply agglomerative clustering and show the dendograms in each of the following cases:

1. single link
2. complete link

	B	C	D	E	F
A	3	12	21	9	8
B		6	4	19	14
C			3	17	21
D				11	6
E					15

Solution

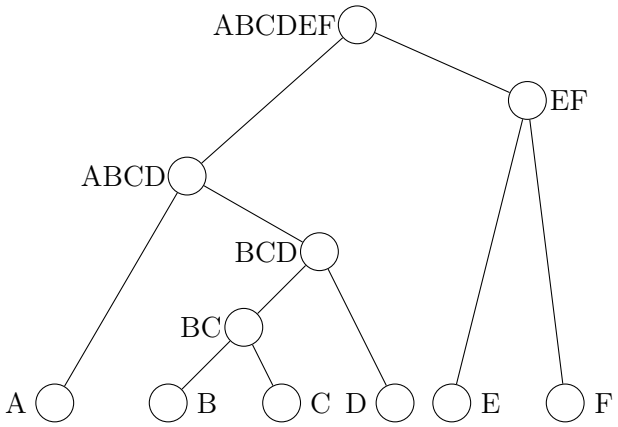
Given distance matrix:

	B	C	D	E	F
A	3	12	21	9	8
B		6	4	19	14
C			3	17	21
D				11	6
E					15

Part 1: Agglomerative Clustering (Single Link)

Using the Single Link method:

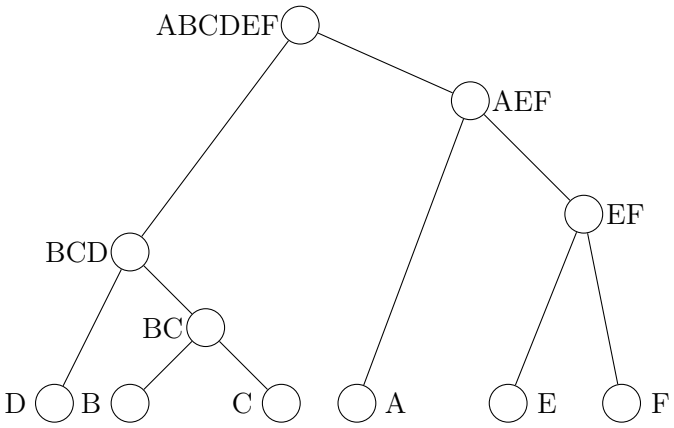
- Identify the closest pair of clusters at each step.
- Merge the clusters with the minimum distance between any pair of points.
- Update the distance matrix until all points are in a single cluster.



Part 2: Agglomerative Clustering (Complete Link)

Using the Complete Link method:

- Identify the farthest pair of clusters at each step.
- Merge clusters based on the maximum distance between any pair of points.
- Update the distance matrix iteratively.



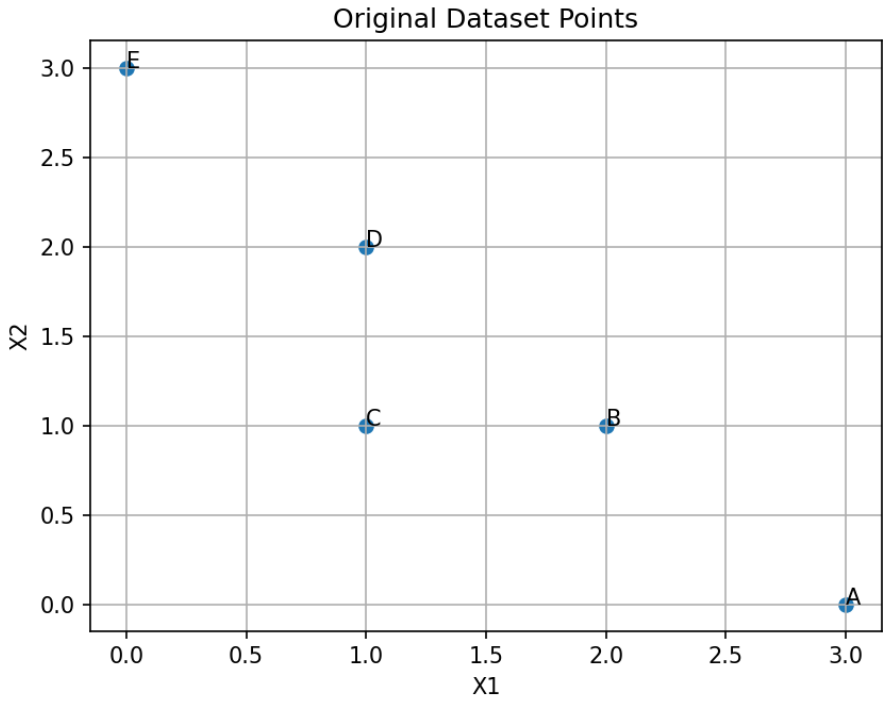
Exercice 16 (chapter 2&7)

Consider a $\mathbb{R}^2$ dataset with these 5 points: $(3,0), (2,1), (1,1), (1,2), (0,3)$.

Draw these points on a graph, then calculate:

1. the centered dataset, the covariance matrix, and the correlation between the two attributes
2. the eigenvalues, and the first principal component
3. the percentage of variance captured by the first principal component
4. the "projected dataset" on the first principal component

Solution



Part 1: Center the Dataset, Covariance Matrix, and Correlation

The centered dataset is calculated by subtracting the mean of each attribute from the dataset:

$$\text{mean}(X) = \left(\frac{3+2+1+1+0}{5}, \frac{0+1+1+2+3}{5}\right) = (1.4, 1.4)$$

$$\text{Centered Dataset} = Z = \begin{pmatrix} 3-1.4 & 0-1.4 \\ 2-1.4 & 1-1.4 \\ 1-1.4 & 1-1.4 \\ 1-1.4 & 2-1.4 \\ 0-1.4 & 3-1.4 \end{pmatrix} = \begin{pmatrix} 1.6 & -1.4 \\ 0.6 & -0.4 \\ -0.4 & -0.4 \\ -0.4 & 0.6 \\ -1.4 & 1.6 \end{pmatrix}$$

The covariance matrix is calculated as:

$$\Sigma = \frac{1}{n} Z^T Z = \begin{pmatrix} 1.04 & -0.96 \\ -0.96 & 1.04 \end{pmatrix}$$

The correlation coefficient is:

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{-0.96}{1.04 \times 1.04} = 0.887$$

Part 2: Eigenvalues and Principal Component

The eigenvalues λ and eigenvectors v of the covariance matrix are given by solving:

$$\det(\Sigma - \lambda I) = 0$$

Eigenvalues:

$$\lambda_1 = 2, \quad \lambda_2 = 0.08$$

First Principal Component:

$$v_1 = \begin{pmatrix} -0.7071 \\ 0.7071 \end{pmatrix}$$

Part 3: Variance Captured

The variance captured by the first principal component is:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{2}{2 + 0.08} = 0.962$$

Part 4: Projected Dataset

Project the dataset onto the first principal component:

$$\text{Projected Data} = Z \cdot v_1 = \begin{pmatrix} -2.121 \\ -0.707 \\ 0 \\ 0.707 \\ 2.121 \end{pmatrix}$$

Exercice 17 (chapter 13)

Consider the following set of numbers: $\{1, 2, 2, 3, 4, 5, 5, 5, 6, 6\}$

1. Apply k-means clustering with $k = 2$ and initial centroids 3 and 6, and calculate the SSE of the resulting clustering

2. Consider applying one iteration of EM with the following two clusters having the same prior probability: $C_1 : \{\mu_1 = 2, \sigma_1 = 1\}$, $C_2 : \{\mu_2 = 5, \sigma_2 = 1\}$.

   In order to simplify the calculation, we assume that $\mathrm{f}(x_j|C_i) = 0$ when $|x_j - \mu_i| > 2.5\sigma_i$

E-Part: calculate the posterior probabilities $P(C_i|x_j)$ in this iteration.

M-Part: calculate the new parameter set after this iteration.

Solution

Part 1: k-Means Clustering with Initial Centroids 3 and 6

Given points: $\{1, 2, 2, 3, 4, 5, 5, 5, 6, 6\}$.

1. Initial centroids: $\mu_1 = 3$ and $\mu_2 = 6$.

2. Assignments: Assign each point to the nearest centroid and recompute the centroids until convergence.

We get:

- **Iteration 1:**
  - Cluster 1: {1, 2, 2, 3, 4} with centroid $\mu_1 = 2.4$
  - Cluster 2: {5, 5, 5, 6, 6} with centroid $\mu_2 = 5.4$
- **Iteration 2:**
  - Cluster 1: {1, 2, 2, 3} with centroid $\mu_1 = 2.0$
  - Cluster 2: {4, 5, 5, 5, 6, 6} with centroid $\mu_2 = 5.167$
- **Iteration 3:**
  - Clusters remain the same:
  - Cluster 1: {1, 2, 2, 3} with centroid $\mu_1 = 2.0$
  - Cluster 2: {4, 5, 5, 5, 6, 6} with centroid $\mu_2 = 5.167$

$$\text{Final SSE} = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2 = 4.833$$

Part 2: E-step of the EM Algorithm

Given:

- $C_1 : \mu_1 = 2, \sigma_1 = 1$
- $C_2 : \mu_2 = 5, \sigma_2 = 1$

1. E-Step: Calculate the posterior probabilities $P(C_i|x_j)$ for each data point x_j:

$$P(C_i|x_j) = \frac{\pi_i \cdot N(x_j|\mu_i, \sigma_i^2)}{\sum_{k=1}^{2} \pi_k \cdot N(x_j|\mu_k, \sigma_k^2)}$$

where:

- π_i is the prior probability (assumed equal for both clusters),
- $N(x|\mu, \sigma^2)$ is the Gaussian probability density function:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We get:

	C1	**C2**
$P(C_j\|1)$	0.9994	0.0005
$P(C_j\|2)$	0.9890	0.0109
$P(C_j\|3)$	0.8175	0.1824
$P(C_j\|4)$	0.1824	0.8175
$P(C_j\|5)$	0.0109	0.0.9890
$P(C_j\|6)$	0.0005	0.9994

Part 3: M-step of the EM Algorithm

Cluster 1

- **Mean:** $\mu_1 = \frac{\sum_{j=1}^{n} w_{1j} \cdot x_j}{\sum_{j=1}^{n} w_{1j}} = 2.071$
- **Standard Deviation:** $\sqrt{\sigma_1^2} = \sqrt{\frac{\sum_{j=1}^{n} w_{1j} \left(x_j - \mu_1\right)^2}{\sum_{j=1}^{n} w_{1j}}} = 0.841$
- **Prior:** $\pi_1 = P(C_1) = \frac{\sum_{j=1}^{n} w_{1j}}{n} = 0.5$

Cluster 2

- **Mean:** $\mu_2 = \frac{\sum_{j=1}^{n} w_{2j} \cdot x_j}{\sum_{j=1}^{n} w_{2j}} = 5.125$
- **Standard Deviation:** $\sqrt{\sigma_2^2} = \sqrt{\frac{\sum_{j=1}^{n} w_{2j} \left(x_j - \mu_2\right)^2}{\sum_{j=1}^{n} w_{2j}}} = 0.782$
- **Prior:** $\pi_2 = P(C_2) = \frac{\sum_{j=1}^{n} w_{2j}}{n} = 0.5$

Exercice 18 (chapter 15)

Consider this set of five $\mathbb{R}^3$ points and the corresponding Euclidean distance table.

	X	Y	Z
A	3	1	2
B	0	2	1
C	3	0	5
D	1	1	1
E	4	2	2

	A	B	C	D	E
A	0	3.32	3.16	2.24	1.41
B	3.32	0	5.39	1.41	4.12
C	3.16	5.39	0	4.58	3.74
D	2.24	1.41	4.58	0	3.32
E	1.41	4.12	3.74	3.32	0

We want to apply DBSCAN with parameters $\epsilon = 2.5$, minPts $= 3$.

Classify each point as core ($|\epsilon$-neighborhood$| \geq 3$), edge or noise.

Note: when counting the ϵ-neighborhood of a point, include the point itself in the count.

Then provide the result of DBSCAN clustering.

Solution

DBSCAN Clustering with $\varepsilon = 2.5$ and MinPts $= 3$

1. **Calculate the ε-neighborhood for each point.**
   - Point A: ε- Neighborhood: [A, D, E]
   - Point B: ε- Neighborhood: [B, D]
   - Point C: ε- Neighborhood: [C]
   - Point D: ε- Neighborhood: [A, B, D]
   - Point E: ε- Neighborhood: [A, E]
2. **Identify core points, edge points, and noise points.**
   - Core Points: A, D
   - Edge Points: B, E
   - Noise Points: C
3. **Cluster the points based on connectivity within ε-neighborhoods.**
   - Cluster 0: {A, B, D, E}
   - Noise: {C}

Exercice 19 (chapter 13&17)

We want to create 3 clusters from the following dataset of numbers:

1, 2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 9

and then evaluate the quality of the result.

ps: some numbers are repeated twice, they represent different rows in the dataset.

1. We apply K-means with initial centroids 1, 4, 9.
   (The results are 2, 5, 8. Show the iterations)
2. Calculate for each cluster i (i denoting 2, 5 or 8):
   - the number of point pairs $N_{in,i}$
   - the sum of distance between points, $W_{in,i}$ ($= \frac{1}{2}\sum W(C_i, C_i)$ as denoted in the course)
3. Accordingly, calculate the total W_{in}, N_{in} and the overall intra-cluster mean distance.
4. You are given the following data for the sums of inter-cluster distances:

   $$W(2,5) = W(5,8) = 48 \qquad W(2,8) = 96$$

   Calculate accordingly the average inter-cluster distance W_{out}/N_{out} then conclude by calculating the BetaCV index.
5. You are given that the smallest 18 pairwise distances in the data sum up to 16, and the largest 18 such distances sum up to 106. Using this data and your previous calculations, calculate the C-index.

Exercice 20 (chapter 18)

Consider the labeled dataset containing:

- a numeric attribute "Num"
- a categorical attribute "Cat"
- a class label (positive "+" or negative "-")

Num	Cat	Label
-2.5	a	-
-1.5	a	-
-1	b	-
-0.5	a	-
0.5	a	-
-0.5	b	+
0.5	b	+
1	a	+
1.5	b	+
2.5	b	+

We use this DS to train a naive Bayes classifier.

1. Calculate the class parameters (profiles)
2. Compute the posterior probabilities of the point (0.25, a) and classify it accordingly.

Solution

Part 1: Class Parameters (Profiles)

Priors

$$P(-) = 0.5$$
$$P(+) = 0.5$$

Mean and Variance for Numeric Attribute (Num)

$$\begin{aligned}\text{Mean}(\text{Num} \mid -) &= -1.0\\ \text{Variance}(\text{Num} \mid -) &= 1.0\\ \text{Mean}(\text{Num} \mid +) &= 1.0\\ \text{Variance}(\text{Num} \mid +) &= 1.0\end{aligned}$$

Conditional Probabilities for Categorical Attribute (Cat)

$$\begin{aligned}P(a \mid -) &= 0.8\\ P(a \mid +) &= 0.2\\ P(b \mid -) &= 0.2\\ P(b \mid +) &= 0.8\end{aligned}$$

Part 2: Classifying the Point (0.25, a)

Posterior Probabilities

The point to classify is $(0.25, a)$. We use the Naive Bayes classifier to compute the posterior probabilities for each class.

Gaussian Function for Numeric Attribute

For a numeric attribute x given class c with mean μ_c and variance σ_c^2:

$$P(x \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right)$$

Calculations

$$\begin{aligned}P(0.25 \mid -) &= \frac{1}{\sqrt{2\pi \cdot 1.0}} \exp\left(-\frac{(0.25-(-1.0))^2}{2 \cdot 1.0}\right)\\ &= 0.1826\\ P(0.25 \mid +) &= \frac{1}{\sqrt{2\pi \cdot 1.0}} \exp\left(-\frac{(0.25-1.0)^2}{2 \cdot 1.0}\right)\\ &= 0.3011\end{aligned}$$

Posterior Probabilities for Each Class

$$\begin{aligned}P(- \mid 0.25, a) &= P(-) \cdot P(0.25 \mid -) \cdot P(a \mid -)\\ &= 0.5 \cdot 0.1826 \cdot 0.8\\ &= 0.07304\\ P(+ \mid 0.25, a) &= P(+) \cdot P(0.25 \mid +) \cdot P(a \mid +)\\ &= 0.5 \cdot 0.3011 \cdot 0.2\\ &= 0.03011\end{aligned}$$

Normalization

$$P(- \mid 0.25, a) = \frac{0.07304}{0.07304 + 0.03011} = 0.7081$$
$$P(+ \mid 0.25, a) = \frac{0.03011}{0.07304 + 0.03011} = 0.2919$$

Classification

The point $(0.25, a)$ is classified as:

$$\boxed{-}$$

Exercice 21 (chapter 19)

We use the same dataset to train a Decision Tree classifier.

Calculate:

1. The entropy of the whole dataset.
2. The info gain by splitting on the *Cat* attribute.
3. The info gain by splitting on the *Num* attribute at each of the following split values:
   - Num < 0?
   - Num < 0.75?
4. From these calculations, which split should be preferred as a first split of the data?

Solution

Part 1: Entropy of the Whole Dataset

The entropy $H(D)$ of the entire dataset D is calculated as follows:

$$H(D) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

where p_i is the probability of class i.
Given the dataset:

$$\text{Negative class}(-) : 5$$
$$\text{Positive class}(+) : 5$$

The probabilities are:

$$P(-) = \frac{5}{10} = 0.5$$
$$P(+) = \frac{5}{10} = 0.5$$

Thus, the entropy is:

$$\begin{aligned} H(D) &= -\left(0.5 \log_2 0.5 + 0.5 \log_2 0.5\right) \\ &= -\left(0.5 \cdot (-1) + 0.5 \cdot (-1)\right) \\ &= 1.0 \end{aligned}$$

Part 2: Information Gain by Splitting on `Cat`

To calculate the information gain when splitting on the categorical attribute `Cat`, we first calculate the entropy of each subset:

Subset `a`

$$\begin{aligned} \text{Negative class}(-) &: 4 \\ \text{Positive class}(+) &: 1 \\ P(- \mid \texttt{a}) &= \frac{4}{5} \\ P(+ \mid \texttt{a}) &= \frac{1}{5} \\ H(D_{\texttt{a}}) &= -\left(\frac{4}{5} \log_2 \frac{4}{5} + \frac{1}{5} \log_2 \frac{1}{5}\right) \\ &= 0.721928 \end{aligned}$$

Subset `b`

$$\begin{aligned} \text{Negative class}(-) &: 1 \\ \text{Positive class}(+) &: 4 \\ P(- \mid \texttt{b}) &= \frac{1}{5} \\ P(+ \mid \texttt{b}) &= \frac{4}{5} \\ H(D_{\texttt{b}}) &= -\left(\frac{1}{5} \log_2 \frac{1}{5} + \frac{4}{5} \log_2 \frac{4}{5}\right) \\ &= 0.721928 \end{aligned}$$

The weighted entropy for the split is:

$$H(D_{\texttt{Cat}}) = \frac{5}{10} \cdot 0.721928 + \frac{5}{10} \cdot 0.721928 = 0.721928$$

The information gain is:

$$IG(\texttt{Cat}) = H(D) - H(D_{\texttt{Cat}}) = 1.0 - 0.721928 = 0.278072$$

Part 3: Information Gain by Splitting on `Num`

Splitting at `Num` $<$ 0

$$\begin{aligned}
\text{Below } 0 &: [-2.5, -1.5, -1, -0.5, -0.5] \\
\text{Negative class}(-) &: 4 \\
\text{Positive class}(+) &: 1 \\
H(D_{\texttt{Num}<0}) &= 0.721928 \\
\text{Above or equal } 0 &: [0.5, 0.5, 1, 1.5, 2.5] \\
\text{Negative class}(-) &: 1 \\
\text{Positive class}(+) &: 4 \\
H(D_{\texttt{Num}\geq 0}) &= 0.721928 \\
H(D_{\texttt{Num}<0}) &= \frac{5}{10} \cdot 0.721928 + \frac{5}{10} \cdot 0.721928 = 0.721928 \\
IG(\texttt{Num} < 0) &= 1.0 - 0.721928 = 0.278072
\end{aligned}$$

Splitting at `Num` $<$ 0.75

$$\begin{aligned}
\text{Below } 0.75 &: [-2.5, -1.5, -1, -0.5, -0.5, 0.5, 0.5] \\
\text{Negative class}(-) &: 4 \\
\text{Positive class}(+) &: 3 \\
H(D_{\texttt{Num}<0.75}) &= 0.985229 \\
\text{Above or equal } 0.75 &: [1, 1.5, 2.5] \\
\text{Negative class}(-) &: 1 \\
\text{Positive class}(+) &: 2 \\
H(D_{\texttt{Num}\geq 0.75}) &= 0.918296 \\
H(D_{\texttt{Num}<0.75}) &= \frac{7}{10} \cdot 0.985229 + \frac{3}{10} \cdot 0.918296 = 0.604184 \\
IG(\texttt{Num} < 0.75) &= 1.0 - 0.604184 = 0.395816
\end{aligned}$$

Part 4: Preferred Split

Based on the information gains calculated:

- Information gain by splitting on `Cat`: 0.278072
- Information gain by splitting on `Num` $<$ 0: 0.278072
- Information gain by splitting on `Num` $<$ 0.75: 0.395816

The preferred split for the first division of the data is `Num` $<$ 0.75, as it provides the highest information gain.

Exercice 22 (chapter 22)

Suppose that we use this dataset as a test dataset for a classifier that is used to detect the positive class, and simply works in the following way:

$$Num > 0.75 \text{ AND } Cat = \text{'b'} \iff Class = +$$

a. Draw the confusion matrix, indicating the terms TP, TN, FP, FN.

b. Calculate the recall and the accuracy.

c. Calculate the F-measure.

Exercice 23 (chapter 2)

In the $\mathbb{R}^3$ space, consider the point $\mathbf{x} = (0, 2, 1)^T$ and the plane S defined by the two orthogonal unit vectors $\mathbf{v_1} = (\sqrt{2}, \sqrt{2}, 0)^T$ and $\mathbf{v_2} = (\sqrt{3}, -\sqrt{3}, \sqrt{3})^T$. Let $\mathbf{x}'$ be the projection of $\mathbf{x}$ on S. Calculate:

1. The coordinates of $\mathbf{x}'$ with respect to the frame of reference $(\mathbf{v_1}, \mathbf{v_2})$.
2. The projection matrix $\mathbf{P}$.
3. The coordinates of $\mathbf{x}'$ with respect to the initial frame of reference.

Solution

Part 1: The coordinates of $\mathbf{x}'$ with respect to the frame of reference $(\mathbf{v_1}, \mathbf{v_2})$

First, we need to normalize $\mathbf{v}_1$ and $\mathbf{v}_2$ to convert them into unit vectors:

$$\hat{\mathbf{v}}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \frac{(\sqrt{2}, \sqrt{2}, 0)^T}{2} = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0\right)^T$$

$$\hat{\mathbf{v}}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \frac{(\sqrt{3}, -\sqrt{3}, \sqrt{3})^T}{3} = \left(\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right)^T$$

We construct the matrix A using these unit vectors:

$$\mathbf{A} = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{3}}{3} \\ 0 & \frac{\sqrt{3}}{3} \end{pmatrix}$$

To find the coordinates of $\mathbf{x}'$ with respect to $(\mathbf{v}_1, \mathbf{v}_2)$, we compute:

$$\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

Applying $\mathbf{A}^\dagger$ to $\mathbf{x}$, we get:

$$\mathbf{x}'_{coords} = \mathbf{A}^\dagger\mathbf{x} = \begin{pmatrix} 1.41421356 \\ -0.57735027 \end{pmatrix}$$

Part 2: The projection matrix P

The projection matrix is given by:

$$\mathbf{P} = \mathbf{A}\mathbf{A}^\dagger$$

After calculations, we get:

$$\mathbf{P} = \begin{pmatrix} 0.83333333 & 0.16666667 & 0.33333333 \\ 0.16666667 & 0.83333333 & -0.33333333 \\ 0.33333333 & -0.33333333 & 0.33333333 \end{pmatrix}$$

Part 3: The coordinates of $\mathbf{x}'$ with respect to the initial frame of reference

Applying the projection matrix P to x:

$$\mathbf{x}' = \mathbf{P}\mathbf{x} = \begin{pmatrix} 0.66666667 \\ 1.33333333 \\ -0.33333333 \end{pmatrix}$$

Exercice 24 (chapter 2&7)

Consider a $\mathbb{R}^2$ dataset with these points: $(3, -1), (2, 0), (1, 2), (0, 4), (-1, 5)$.

1. Calculate the centered dataset, the covariance matrix, and the correlation between the two attributes.

2. Calculate the eigenvalues and the first principal component.

3. What is the percentage of variance captured by the first principal component?

4. What is the SSE (sum of squared errors) if we project the dataset on the first principal component?

Solution

1. Centered Dataset

Given the dataset:

$$\begin{pmatrix} -3 & -1 \\ 2 & 0 \\ 1 & 2 \\ 0 & 4 \\ -1 & 5 \end{pmatrix}$$

The mean of the dataset is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \left(\frac{-3+2+1+0-1}{5}, \frac{-1+0+2+4+5}{5}\right) = (0, 2)$$

Centering the dataset:

$$\mathbf{Z} = \begin{pmatrix} 2 & -3 \\ 1 & -2 \\ 0 & 0 \\ -1 & 2 \\ -2 & 3 \end{pmatrix}$$

2. Covariance Matrix

The covariance matrix is calculated by dividing by n:

$$\Sigma = \frac{1}{n}\mathbf{Z}^{\mathbf{T}}\mathbf{Z} = \begin{pmatrix} 2 & -3.2 \\ -3.2 & 5.2 \end{pmatrix}$$

3. Correlation Between Attributes

The correlation between the two attributes is:

$$\text{Correlation}(X, Y) = \frac{\sigma_X}{\sigma_X \sigma_Y} = -0.9922778767136677$$

4. Eigenvalues and Principal Component

The eigenvalues and eigenvectors of the covariance matrix are:

$$\lambda_1 = 7.17770876, \quad \lambda_2 = 0.02229124$$

The first principal component is the eigenvector corresponding to the largest eigenvalue:

$$\mathbf{v}_1 = \begin{pmatrix} 0.52573111 \\ -0.85065081 \end{pmatrix}$$

5. Percentage of Variance

The percentage of variance is given by:

$$\text{Variance percentage} = \frac{\lambda_1}{\lambda_1 + \lambda_2} \times 100\% = 99.69039949999532\%$$

6. SSE (Sum of Squared Errors)

Projecting the data onto the first principal component and reconstructing:

$$\text{Projection} = \mathbf{P} \cdot \text{Centered Data} = \begin{pmatrix} 3.60341465 \\ 2.22703273 \\ 0 \\ -2.22703273 \\ -3.60341465 \end{pmatrix}$$

The SSE is:

$$\text{SSE} = \sum_{i=1}^{n} \left\| X_i - \hat{X}_i \right\|^2 = 0.11145618000168248$$

Exercice 25 (chapter 3&13)

The adjacent dataset has nine data points and two attributes: $\mathbf{X_1}$ numerical and $\mathbf{X_2}$ categorical.

X_1	X_2
-5	A
-3	B
-1	A
0	A
1	A
1	A
2	B
3	B
5	B

1. We consider the numerical attribute $\mathbf{X_1}$ alone. Apply K-Means with $K = 3$ clusters, and initial centroids $C_1 : 0, C_2 : 1, C_3 : 2$.

   Now we discretize the attribute X_1: we replace it by a categorical attribute $\mathbf{Y_1}$ according to which cluster the value of $\mathbf{X_1}$ belongs. Hence, $\mathbf{Y_1}$ has 3 possible values: C_1, C_2 or C_3.

2. Show the new dataset composed of attributes $\mathbf{Y_1}$ and $\mathbf{X_2}$.

3. Construct the contingency table between Y_1 and X_2, then calculate the χ^2 statistic.

Solution

Part 1: Apply K-Means with $K = 3$ clusters

1. Initial centroids: $\mu_1 = 0$ and $\mu_2 = 1$ and $\mu_3 = 2$.
2. Assignments: Assign each point to the nearest centroid and recompute the centroids until convergence.

- **Iteration 1:**

$$\text{Centroids: } [-2.25], [1], [3.33]$$

$$\text{Assignments: } [C_1, C_1, C_1, C_2, C_2, C_2, C_2, C_3, C_3]$$

- **Iteration 2:**

$$\text{Centroids: } [-3], [1], [4]$$

$$\text{Assignments: } [C_1, C_1, C_1, C_2, C_2, C_2, C_2, C_3, C_3]$$

- **Iteration 3:**

$$\text{Centroids: } [-3], [1], [4]$$

$$\text{Assignments: } [C_1, C_1, C_1, C_2, C_2, C_2, C_2, C_3, C_3]$$

Part 2: Combine with $\mathbf{X_2}$

We create a new dataset combining $\mathbf{Y_1}$ (cluster assignments) and $\mathbf{X_2}$:

$\mathbf{Y_1}$	$\mathbf{X_2}$
C_1	A
C_1	B
C_1	A
C_2	A
C_2	A
C_2	A
C_2	B
C_3	B
C_3	B

Part 3: Construct Contingency Table and Calculate χ^2 Statistic

The contingency table is a summary of the counts between the categorical variables $\mathbf{Y_1}$ and $\mathbf{X_2}$:

	$\mathbf{X_2}$		
$\mathbf{Y_1}$	A	B	Total
C_1	2	1	3
C_2	3	1	4
C_3	0	2	2
Total	5	4	9

	$\mathbf{X_2}$		
$\mathbf{Y_1}$	A	B	Total
C_1	1.67	1.33	3
C_2	2.22	1.78	4
C_3	1.11	0.89	2
Total	5	4	9

The χ^2 statistic is calculated as:

$$\chi^2 = \sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Where n_{ij} are the observed frequencies and e_{ij} are the expected frequencies under the null hypothesis of independence.

For our data:

$$\chi^2 = 3.2625$$