Machine Learning

Unsupervised Learning

Academic year 2023/2024

Queen Mary University of London

EXERCISES

EXERCISE ♯1. Consider the following simple dataset, where ID denotes the sample identifier and x, y and z are three attributes:

ID	x	y	z
1	1	2	A
2	2	1	A
3	2	3	A
4	3	2	A
5	4	5	B
6	5	4	B
7	5	6	B
8	6	5	B
9	5	5	B

ID	x	y	z
10	7	5	B
11	3	5	B
12	5	3	B
13	5	7	B
14	4	4	B
15	6	6	B
16	6	4	B
17	4	6	B
18	5	5	B

Start by plotting the dataset in the (x, y) space.

a. Using a bin width of $\Delta = 1$ and bin centres at 0, 1, 2... obtain:
   - The histogram for the probability densities $p(x)$ and $p(y)$.
   - The histogram for the probability densities $p(x|z = A)$ and $p(x|z = B)$.

b. Assume that the probability density $p(x, y)$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:
   - Estimate $\boldsymbol{\mu}$ and plot it.
   - Do you expect $\boldsymbol{\Sigma}$ be a diagonal matrix?

c. Assume that the probability densities $p(x)$ and $p(y)$ are Gaussian with means μ_x and μ_y, respectively:
   - What is the relationship between the means μ_x and μ_y and the mean $\boldsymbol{\mu}$?
   - Can $p(x, y)$ be expressed as the product of $p(x)$ and $p(y)$?

d. Let $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ be the means of $p(x, y|z = A)$ and $p(x, y|z = B)$, respectively.
   - Estimate $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ and plot them.
   - Can you obtain $\boldsymbol{\mu}$ from $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$?

e. Let x' and y' be two new features obtained after applying principal components analysis to the raw features x and y.
   - Plot the transformed dataset in the (x', y') space.
   - Will the covariance matrix of the probability density $p(x', y')$ be diagonal?
   - Can $p(x', y')$ be expressed as the product of $p(x')$ and $p(y')$?

   Note: Except for the calculations of the means, all the questions in this exercise can be answered by resorting to your intuition and using some background theory. You are encouraged to obtain the exact answers in your data environment of choice to assess your intuition!

EXERCISE ♯2. Consider the following simple dataset:

ID	x	y
1	1	2
2	2	1
3	2	3
4	3	2
5	4	5
6	5	4
7	5	6
8	6	5

In this exercise, we will use k-means to define two clusters ($k = 2$). Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be the prototypes (centres) of each cluster. In k-means, the quality of a clustering solution is defined as the average of the square distances of each sample to its prototype and the best solution is usually obtained by applying iteratively an expectation-minimisation (EM) approach: first each sample is assigned to the closest prototype (*E* stage) and then the prototypes are updated as the average of the samples belonging to the newly defined cluster (*M* stage).

- Assuming that the initial values of the prototypes are $\boldsymbol{\mu}_1 = [0, 0]^T$ and $\boldsymbol{\mu}_2 = [7, 7]^T$, what clusters will the k-means algorithm identify? What is the quality of the clustering solution?
- Assuming that the initial values of the prototypes are $\boldsymbol{\mu}_1 = [1, 6]^T$ and $\boldsymbol{\mu}_2 = [6, 1]^T$, what clusters will the k-means algorithm identify? What is the quality of the clustering solution?
- Compare both results and explain the discrepancies observed by using the notion of local minima.