Intelligent Data Analysis - Assignment 1

Problems

Felipe Orihuela-Espina

Contents

1 Preliminaries

This assignment is scored on a scale of 0 to 10 and its weight represents 20% of the final mark of the module.

- **Deadline**: **14 November 2024 at 16:00 UK time**.
- **Late submissions policy**: There will be a penalty of 10% per hour or part thereof for late submissions with a final cut-off time five hours after the deadline.
- **What to submit?**: Just answer the quiz in canvas.
- **Learning outcomes covered**:
  1. Demonstrate knowledge and understanding of core ideas of pattern analysis, data analytics and data mining
  2. Demonstrate understanding of broader issues of generalisation in intelligent data analysis
  3. Demonstrate the ability to apply the main approaches to unseen examples.
- **Where has the skill to complete this assignment been learned in the module**:
  - Block 1: Nature of data, Weeks 1-2 covered core ideas
  - Block 2: Data as vectors / Point clouds, Weeks 3-5 covered the understanding of data analysis for cloud points.

2 About the assignment

The assignment has 2 parts:

1. This collection of **exercises**, with no time limit other than the deadline indicated above. You can do this off-line, and you do *not* need to submit your attempt of the solution in canvas, but you'll need the solutions to answer the quiz.

2. A **quiz** in canvas. You complete the quiz in canvas. This quiz is *time constrained*; you have 2 hours to complete the quiz, and each question *can only be attempted once*.

Make sure you complete the exercises before you attempt the quiz. **All the assessment of this assignment is done via the quiz in canvas**. Therefore, you do *not* need to submit your attempt of the solutions in canvas. However, for some of the questions in the quiz, you'll need to have solved the exercises in advance.

The questions in the quiz related to the exercises may be both, about the final outcome as well as about intermediate calculations. Further, the quiz may also include a few questions unrelated to the exercises though, e.g. more theoretical.

The quiz contains a number of 10 questions randomized from a pool. **The quiz is time constrained (2h) and can only be attempted once**. For not having late penalty, you ought to have *finished* the quiz *before* the deadline. In other words i.e. the latest you should start the quiz is 2h before the deadline.

The quiz can only be attempted once. Do not initiate the quiz until you are ready to answer the questions as questions only show once. You will not be allowed to go back.

3 Block 1: Nature of data (Wks 1-2)

Exercise 3.1. Random variables. Let be the sample in Table 1.

4.7	3.6	3.8	2.2	4.7	4.1	3.6	4.0	4.4	5.1
4.2	4.1	4.4	5.0	3.7	3.6	2.9	3.7	4.7	3.4
3.9	4.8	3.3	3.3	3.6	4.6	3.4	4.5	3.3	4.0
3.4	4.0	3.8	4.1	3.8	4.4	4.9	4.9	4.3	6.0

Tab. 1: Blood glucose levels for first-year medical students [mmol/liter].

a) **Making intervals of size 0.5 (closed by the left, open by the right e.g. $[a, b)$), draw the histogram of the distribution.**
b) **Observing the histogram determine whether the distribution is symmetrical, skewed to the right or skewed to the left.**
c) **Calculate the standard deviation using the Bessel correction (denominator $n - 1$).**
d) **Set the interval corresponding to two standard deviations about the mean**

Exercise 3.2. Distributions. Let $f(X)$ be a probability density function. Are the following statement true or false?:

1. $Pr(X \leq x) = \int_{-\infty}^{x} f(X)dx$
2. $Pr(X \geq x) = \int_{x}^{\infty} f(X)dx$
3. $Pr(X = x) = 0$

4 Block 2: Data as vectors / Point clouds (Wks 3-5)

Exercise 4.1. PCA. Let be the point cloud:

$$
\begin{aligned}
X &= [X_1 X_2 X_3 X_4] \\
&= \begin{bmatrix}
4.5 & -4.9 & 13.5 & -0.9 \\
2.2 & 0.4 & 6.6 & 4.4 \\
-1.0 & -4.0 & -3.0 & 0 \\
3.3 & -3.5 & 9.9 & 0.5 \\
-3.7 & 1.3 & -11.1 & 5.3 \\
-4.4 & 3.6 & -13.2 & 7.6 \\
-4.2 & 4.7 & -12.6 & 8.7 \\
-3.4 & 0.7 & -10.2 & 4.7 \\
-1.8 & 5.0 & -5.4 & 9.0 \\
-2.0 & 0.5 & -6.0 & 4.5 \\
-5.0 & -5.0 & -15.0 & -1.0 \\
5.0 & 5.0 & 15.0 & 9.0
\end{bmatrix}
\end{aligned}
$$

Using PCA establish the intrinsic dimensionality of X.

Exercise 4.2. Statistical independence. In a scientific study with rats, the population S of rats has the following demographics:

ID	Age	Sex	Hair	ID	Age	Sex	Hair
1	0	M	White	11	5	M	Brown
2	5	M	Gray	12	4	M	Brown
3	4	F	White	13	7	F	White
4	3	M	White	14	5	F	Black
5	6	M	Gray	15	6	M	Brown
6	4	M	White	16	3	M	White
7	4	F	Gray	17	3	F	Brown
8	6	F	Gray	18	6	M	Brown
9	6	F	Gray	19	1	F	Brown
10	4	M	White	20	1	F	Brown

donde:

- **Age: In years**
- **Sex: M: Male, F: Female**

During the scientific experiment, half of the rat population had a tumor induced. If the tumor affects males and females equally, what is the probability that a rat is female or has a tumor?