Project description for report 1

Objective: The objective of this report is to apply the methods you have learned in the first section of the course,*"Data: Feature extraction, and visualization"* on your own data set to get a basic understanding of your data prior to the further analysis (project report 2).

Material: You can use the 02450Toolbox on Inside to see how the various methods learned in the course are used in Matlab, R or Python. In particular, you should review exercise 1 to 5 in order to see how the various tasks can be carried out.

Mandatory section

In order to have your report evaluated, it must contain the following two items:

- According to the DTU regulations, each students contribution to the report must be clearly specified. Therefore, for each section, specify which student was responsible for it (use a list or table). **A report must contain this documentation to be accepted**. The responsibility assignment must be individualized[1]

- Solutions, or attempted solutions, for at least four of the exam problems found at the end of this document[2] . The solutions do not have to be long (a couple of lines, perhaps a calculation) but must show the gist of your reasoning so as to verify you have worked independently on the problem. We suggest they are given in an itemized format:

    1. Option $A/B/C/D$: To see this ...

[1]For reports made by 3 students: Each section must have a student who is 40% or more responsible. For reports made by 2 students: Each section must have a student who is 60% or more responsible.

[2]We ask you to do this because it has been our experience some students are unfamiliar with the written exam format until days before the exam, and we think this is the best way to ensure the requirements of the written exam are made clear early on. We don't evaluate your answers for correctness because that aspect of the course will be tested at the exam and would be redundant here.

2. Option $A/B/C/D$: We solve this by using..

Don't know is obviously not allowed, but you can take inspiration from the homework problems (and solutions given at the end of the notes). The purpose is to demonstrate that you have worked on the exam problems but not to test for correctness, and you can therefore hand in solutions which describes your best attempt at solving the problem (but you know are wrong). Keep in mind the solutions (fraction correct etc.) will not affect your evaluation, but rather whether the report is evaluated at all.

Your report cannot be evaluated unless it contains these items.

Handin checklist

- Make sure the mandatory section is included
- Make sure the report clearly display the **names *and* study numbers** of all group members. Make sure study numbers are correct.
- Your handin should consist of exactly two files: A `.pdf` file containing the report, and a `.zip` file containing the code you have used (extensions: `.py`, `.R` or `.m`; do **not** upload your data). The reports are not evaluated based on the quality of the code (comments, etc.), however we ask the code is included to avoid any potential issues of illegal collaboration between groups. Please do not compress or convert these files.
- Reports are evaluated based on how well they address the questions below. Therefore, to get the best evaluation, address all questions
- Use the group handin feature. **Do not upload separate reports for each team member as this will lead to duplicate work and unhappy instructors**
- **Deadline for handin is no later than 8 March at 13:00**. Late handins will not be accepted under normal circumstances

Description

Understanding the data you are trying to model well is very important. You can apply very sophisticated machine learning methods but if you are not aware of potential issues with the data the further modeling will be difficult. Thus, the aim of

this first project is to get a thorough understanding of your data and describe how you expect the data can be used in the later reports.

Report 1 should cover what you have learned in the lectures and exercises of week 1 to 4 covering the section *"Data: Feature extraction, and visualization"* . You should consider yourself as a new employee in a company who has just been given a data set. Your job is to make a useful description of the data set for your co-workers and make some basic plots. In particular, the report ***must*** include the following items and the report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality. For readability and brevity consider not using one subsection for each item.

1. **A description of your data set.**

   - Explain what your data is about. I.e. what is the overall problem of interest?
   - Provide a reference to where you obtained the data.
   - Summarize previous analysis of the data. (i.e. go through one or two of the original source papers and read what they did to the data and summarize their results).
   - You will be asked to apply (1) classification and (2) regression on your data in the next report. For now, we want you to consider how this should be done. Therefore:
     Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?.
     Explain which attribute you wish to predict in the regression based on which other attributes? Which class label will you predict based on which other attributes in the classification task?
     If you need to transform the data in order to carry out these tasks, explain roughly how you plan to do this.

   One of these tasks (1)–(5) is likely more relevant than the rest and will be denoted the **main machine learning aim** in the following. The purpose of the following questions, which asks you to describe/visualize the data, is to allow you to reflect on the feasibility of this task.

2. **A detailed explanation of the attributes of the data.**

- Describe if the attributes are discrete/continuous, Nominal/Ordinal/Interval/Ratio,
- Give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so.
- Include basic summary statistics of the attributes.

If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense).

3. **Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).**
Touch upon the following subjects, use visualizations when it appears sensible. *Keep in mind the ACCENT principles and Tufte's guidelines when you visualize the data.*

   - Are there issues with outliers in the data,
   - do the attributes appear to be normal distributed,
   - are variables correlated,
   - does the primary machine learning modeling aim appear to be feasible based on your visualizations.

   There are three aspects that needs to be described when you carry out the PCA analysis for the report:

   - The amount of variation explained as a function of the number of PCA components included,
   - the principal directions of the considered PCA components (either find a way to plot them or interpret them in terms of the features),
   - the data projected onto the considered principal components.

   If your attributes have different scales you should include the step where the data is standardizes by the standard deviation prior to the PCA analysis.

4. **A discussion explaining what you have learned about the data.**
Summarize here the most important things you have learned about the data and give also your thoughts on whether your primary machine learning aim appears to be feasible based on your visualization.

Collaboration

The usual DTU rules for collaboration applies for the reports. The main rule is that if you hand in a report, you must have authored or co-authored the content of the report for this assignment, and if your report contains text you did not write, then it must be with attribution. Notice in particular:

- If you are taking the course again, you are allowed to re-use content from a report that you previously authored or co-authored.
- If you are authoring a report together with a person who has previously taken the course, you cannot re-use that report since you did not originally author it. We recommend that you simply choose another dataset and re-write the text such that the new report can be considered original joint work by both authors.
- You are of course allowed to use the scripts, etc. supplied in this course for the reports.

The report should be 5-10 pages long (and no longer!) including figures and tables and give a precise and coherent introduction to and overview of the dataset you have chosen.

Transferring/reusing reports from previous semesters

If you are retaking the course, you are allowed to reuse your previous report. You can either have the report transferred in it's entirety, or re-work sections of the report and have it evaluated anew.

To have a report transferred, *do absolutely nothing*. Reports from previous semesters are automatically transferred. Therefore, please do not upload old reports to Inside as this will lead to duplicate work. As a safeguard, we will contact all students who are missing reports shortly after the exam.

If you wish to redo parts of a report you have already handed in as part of a group in a previous semester, then to avoid any issues about plagiarism please keep attribution to the original group members for those sections you choose not to redo.

1 Exam problems for the project

Problems

Question 1. Spring 2019 question 1: The main dataset used in this exam is the Urban Traffic dataset[3] described in table 1. We will consider the type of an attribute as the highest level it obtains in the type-hierarchy (nominal, ordinal, interval and ratio). Which of the following statements are true about the types of the attributes in the Urban Traffic dataset?

No.	Attribute description	Abbrev.
x_1	30-minute interval (coded)	Time of day
x_2	Number of broken trucks	Broken Truck
x_3	Number of accident victims	Accident victim
x_4	Number of immobile busses	Immobilized bus
x_5	Number of trolleybus network defects	Defects
x_6	Number of broken traffic lights	Traffic lights
x_7	Number of run over accidents	Running over
y	Level of congestion/slowdown (low to high)	Congestion level

Table 1: Description of the features of the Urban Traffic dataset used in this exam. The dataset describes urban traffic behaviour of the city of Sao Paulo in Brazil. Each observation corresponds to a 30-minute interval between 7:00 and 20:30, indicated by the integer x_1, such that $x_1 = 1$ corresponds to 7:00-7:30 and so on up to $x_1 = 27$ that corresponds to 20:00-20:30. The other attributes $x_2, \ldots, x_7$ corresponds to a number of occurences of the given type in that 30-minute interval. We will consider the primary goal to be classification, namely to predict y which is the level of congestion of the bus network in the given interval. The dataset used here consists of $N = 135$ observations and the attribute y is discrete taking values $y = 1$ (corresponding to no congestion), $y = 2$ (corresponding to a light congestion), $y = 3$ (corresponding to an intermediate congestion), and $y = 4$ (corresponding to a heavy congestion).

A x_1 (*Time of day*) is nominal, x_2 (*Broken Truck*) is ratio, x_7 (*Running over*) is ratio, and y (*Congestion level*) is ordinal

B x_1 (*Time of day*) is ratio, x_4 (*Immobilized bus*) is nominal, x_6 (*Traffic lights*) is ratio, and y (*Congestion level*) is ordinal

C x_1 (*Time of day*) is ordinal, x_6 (*Traffic lights*) is ratio, x_7 (*Running over*) is ratio, and y (*Congestion level*) is ordinal

D x_1 (*Time of day*) is interval, x_6 (*Traffic lights*) is ratio, x_7 (*Running over*) is ratio, and y (*Congestion level*) is ordinal

E Don't know.

Question 2. Spring 2019 question 2: Consider again the Urban Traffic dataset from table 1 and in particular the 14 and 18'th observation

$$\boldsymbol{x}_{14} = \begin{bmatrix} 26 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_{18} = \begin{bmatrix} 19 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Which of the following statements about the p-norm distance $d_p(\cdot, \cdot)$ is correct?

A $d_{p=\infty}(\boldsymbol{x}_{14}, \boldsymbol{x}_{18}) = 7.0$

B $d_{p=3}(\boldsymbol{x}_{14}, \boldsymbol{x}_{18}) = 3.688$

C $d_{p=1}(\boldsymbol{x}_{14}, \boldsymbol{x}_{18}) = 1.286$

D $d_{p=4}(\boldsymbol{x}_{14}, \boldsymbol{x}_{18}) = 4.311$

E Don't know.

Question 3. Spring 2019 question 3: A Principal Component Analysis (PCA) is carried out on the Urban Traffic dataset in table 1 based on the attributes $x_1,\ x_2,\ x_3,\ x_4,\ x_5$.

The data is standardized by (i) substracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is

[3]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Behavior+of+the+urban+traffic+of+the+city+of+Sao+Paulo+in+Brazil`

then carried out on the standardized data matrix to obtain the decomposition $\boldsymbol{USV}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.49 & -0.5 & 0.08 & -0.49 & 0.52 \\ 0.58 & 0.23 & -0.01 & 0.71 & 0.33 \\ 0.56 & 0.23 & 0.43 & -0.25 & -0.62 \\ 0.31 & 0.09 & -0.9 & -0.19 & -0.24 \\ -0.06 & 0.8 & 0.03 & -0.41 & 0.43 \end{bmatrix} \quad (1)$$

$$\boldsymbol{S} = \begin{bmatrix} 13.9 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 12.47 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 11.48 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 10.03 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 9.45 \end{bmatrix}.$$

Which one of the following statements is true?

A The variance explained by the first four principal components is greater than 0.8

B The variance explained by the last three principal components is greater than 0.51

C The variance explained by the first two principal components is less than 0.5

D The variance explained by the first three principal components is less than 0.7

E Don't know.

Question 4. Spring 2019 question 4: Consider again the PCA analysis for the Urban Traffic dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in eq. (2). Which one of the following statements is true?

$$\boldsymbol{V} = \begin{bmatrix} 0.49 & -0.5 & 0.08 & -0.49 & 0.52 \\ 0.58 & 0.23 & -0.01 & 0.71 & 0.33 \\ 0.56 & 0.23 & 0.43 & -0.25 & -0.62 \\ 0.31 & 0.09 & -0.9 & -0.19 & -0.24 \\ -0.06 & 0.8 & 0.03 & -0.41 & 0.43 \end{bmatrix} \quad (2)$$

A An observation with a low value of **Time of day**, a low value of **Broken Truck**, a high value of **Accident victim**, a high value of **Immobilized bus**, and a low value of **Defects** will typically have a positive value of the projection onto principal component number 5.

B An observation with a low value of **Accident victim**, and a high value of **Immobilized bus** will typically have a positive value of the projection onto principal component number 3.

C An observation with a low value of **Time of day**, a high value of **Broken Truck**, a low value of **Accident victim**, and a low value of **Defects** will typically have a negative value of the projection onto principal component number 4.

D An observation with a low value of **Time of day**, a high value of **Broken Truck**, a high value of **Accident victim**, and a high value of **Defects** will typically have a positive value of the projection onto principal component number 2.

E Don't know.

Question 5. Spring 2019 question 14: Suppose s_1 and s_2 are two text documents containing the text:

$$s_1 = \left\{ \begin{matrix} \text{the bag of words representation} \\ \text{becomes less parsimoneous} \end{matrix} \right\}$$
$$s_2 = \{\text{if we do not stem the words} \quad \}$$

The documents are encoded using a bag-of-words encoding assuming a total vocabulary size of $M = 20000$. No stopwords lists or stemming is applied to the dataset. What is the Jaccard similarity between documents s_1 and s_2?

A Jaccard similarity of s_1 and s_2 is 0.153846

B Jaccard similarity of s_1 and s_2 is 0.000650

C Jaccard similarity of s_1 and s_2 is 0.000100

D Jaccard similarity of s_1 and s_2 is 0.136977

E Don't know.

Question 6. Spring 2019 question 27: Consider the Urban Traffic dataset from table 1.

Suppose the attributes have been binarized such that $\hat{x}_2 = 0$ corresponds to $x_2 \leq 1$ (and otherwise $\hat{x}_2 = 1$) and $\hat{x}_7 = 0$ corresponds to $x_7 \leq 0$ (and otherwise $\hat{x}_7 = 1$). Suppose the probability for each of the configurations of $\hat{x}_2$ and $\hat{x}_7$ conditional on the congestion level y are as given in table 2 and the prior probability of the congestion levels are

$$p(y = 1) = 0.274,\ p(y = 2) = 0.23,$$
$$p(y = 3) = 0.244,\ p(y = 4) = 0.252,$$

What is then the probability an observation had $\hat{x}_2 = 0$ given light congestion?

$p(\hat{x}_2, \hat{x}_7 \mid y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$
$\hat{x}_2 = 0,\ \hat{x}_7 = 0$	0.73	0.81	0.7	0.68
$\hat{x}_2 = 0,\ \hat{x}_7 = 1$	0.11	0.03	0.03	0.09
$\hat{x}_2 = 1,\ \hat{x}_7 = 0$	0.16	0.1	0.21	0.18
$\hat{x}_2 = 1,\ \hat{x}_7 = 1$	0	0.06	0.06	0.05

Table 2: Probability of observing particular values of $\hat{x}_2$ and $\hat{x}_7$ conditional on y.

A $p(\hat{x}_2 = 0 \mid y = 2) = 0.116$

B $p(\hat{x}_2 = 0 \mid y = 2) = 0.84$

C $p(\hat{x}_2 = 0 \mid y = 2) = 0.243$

D $p(\hat{x}_2 = 0 \mid y = 2) = 0.193$

E Don't know.