Project 1 report

Rasmus Clausen (s203279), Andras Kurucsai (s240067),
Leonardo Rodovero (s240095)

29 February 2023

Section	Rasmus	Andras	Leonardo
Section 1	30%	30%	40%
Section 2	40%	30%	30%
Section 3	30%	40%	30%
Section 4	40%	30%	30%
Exam Questions	33%	33%	33%

1 Introduction

1.1 Dataset

The area of interest pertains to understanding how socio-economic (GDP, schooling, income composition), health (immunization rates, adult mortality, HIV/AIDS prevalence), and lifestyle (alcohol consumption, BMI) factors explain life expectancy among and between countries. In this project we have considered data [1] from year 2000-2015 for 179 countries for further analysis and it consists of 21 columns and 2864 rows.

1.2 Previous works on the original dataset

We have considered a 'corrected' dataset because in the original dataset there are a lot of missing values and corrupted data (see 2.1). However, to have an idea about the previous work on the original dataset [2] which has data from 193 countries between 2000 and 2015 we have taken into consideration two notebooks.

The first notebook [3] looks at health indicators like immunization rates and economic conditions from sources like the WHO (World Health Organization) and the UN (United Nations). The main goal is to see how various factors affect life expectancy, focusing on changes over time but also on differences between countries. The data set used has details from many countries, showing improvements in health over 15 years, especially in developing countries. The analysis pays special attention to health statistics like adult and child mortality

rates and the presence of diseases, alongside economic measures like how well countries use their resources, which seems to greatly affect health outcomes.

In the second notebook [4] there is an analysis of how health-related and socio-economic factors affect life expectancy worldwide.The study looks at life expectancy and health data for 193 countries from the World Health Organization (WHO) and economic information from the United Nations, covering the years 2000 to 2015. It highlights significant health sector improvements, especially in developing countries. The analysis also focuses on the Income Composition of Resources (ICOR) to assess how efficiently countries use their resources. ICOR is strongly linked to life expectancy, indicating that effective resource use is crucial for better health outcomes. In the corrected dataset that we are using however does not contain the ICOR and and Government expenditure properties.

1.3 Goal of the project

The goal for this project is to do classification and regression on the corrected WHO dataset. Here we plan on doing classification for the Developing/Developed attribute and regression on Life Expectancy. From this we hope to find which factors affect the Development and Life expectancy in a country, and what effect does immunization, health and economic factors have on these.

In order to meet our goal, we plan to transform the dataset in the ways our analysis sees fit, possibly removing attributes if we deem them unnecessary. Here it is also possible to project our data using PCA, in a way that reduces the dimensionality of our model. For the next project, we plan to train different models and evaluate their performance on our regression and classification tasks. We will then choose the right model based on our evaluation and the requirements for the task.

References

[1] https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

[2] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

[3] https://www.kaggle.com/code/lonnieqin/life-expectancy-prediction

[4] https://www.kaggle.com/code/hazalyl/proje-final

2 The attributes of the data

The dataset has 21 properties, which are listed below in Table 1.

Name and description	*Numerical Type*	*Type*
Country Country in which the measurement took place. (179 countries in total)	Discrete	Nominal
Region 179 countries are distributed in 9 regions. E.g. Africa, Asia, Oceania, European Union, Rest of Europe and etc.	Discrete	Nominal
Year Years observed from 2000 to 2015.	Discrete	Interval
Infant deaths Number of infant deaths per 1000 population.	Continuous	Ratio
Under-five deaths Number of under-five deaths per 1000 population.	Continuous	Ratio
Adult Mortality Number of adult deaths between 15 and 60 years per 1000 population, for both sexes.	Continuous	Ratio
Alcohol Alcohol consumption per capita(15+), in litres of pure alcohol.	Continuous	Ratio
Hepatitis B HepB immunization coverage among 1-year-olds (%).	Discrete	Ratio
Measles Number of reported cases per 1000 population.	Discrete	Ratio
BMI Average Body Mass Index of entire population.	Continuous	Interval
Polio Polio (Pol3) immunization coverage among 1-year-olds (%).	Discrete	Ratio
Diphtheria Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).	Discrete	Ratio
HIV/AIDS	Continuous	Ratio

Incidents of HIV per 1000 population aged 15-49.		
GDP Gross Domestic Product per capita in USD.	Discrete	Ratio
Population Total population in millions.	Continuous	Ratio
Thinness 10 - 19 Prevalence of thinness among adolescents aged 10-19 years. BMI < -2 standard deviations below the median (%).	Continuous	Ratio
Thinness 5 - 9 Prevalence of thinness among children aged 5-9 years. BMI < -2 standard deviations below the median (%).	Continuous	Ratio
Schooling Number of years of Schooling.	Continuous	Ratio
Economy status: Developed Indicating if a country is developed.	Discrete	Nominal
Economy status: Developing Indicating if a country is developing.	Discrete	Nominal
Life expectancy Given as age in years.	Continuous	Ratio

Table 1: The properties of the dataset

2.1 Data issues

The dataset we are using is a modified version of another dataset. The decision to use the modified version was made because there were several issues found in the original. The original dataset contained a large amount of missing values (see Appendix 10.), and a lot of values for certain attributes were found faulty considering common sense. E.g. there were values larger than 1000 for attributes that were measured "per 1000 population".

In the corrected dataset, we have checked and verified that there are exactly 16 entries (corresponding to the 16 years) for each of the 179 countries. The validity of the one hot key encoding for the Developed/Developing properties have also been checked. For attributes given in percentage or in per 1000 population, it has been checked if they lie in the correct interval, namely [0, 100] and [0, 1000] respectively. After running these tests, all data entries were found to be valid.

2.2 Summary statistics

For summary statistics for the dataset we chose to calculate the **Mean**, **Mode**, **Variance**, **Standard Deviation** and **Range**. They can be seen in table 2 below. Here it is important to note that for the *Mode* of 'Year', there is a data point for every country with the years 2000-2015. When looking at the ranges and standard deviations of the data, we see that it will be necessary to standardize the dataset by subtracting the mean and dividing by the standard deviation.

Attribute	Mean	Mode	Std. Deviation	Range
Year	2007.5	-	4.6106	[2000, 2015]
Infant deaths	30.364	3.3	27.5381	[1.8, 138.1]
Under_five deaths	42.938	4.1	44.5700	[2.3, 224.9]
Adult mortality	192.252	91.946	114.9103	[49.384, 719.361]
Alcohol consumption	4.821	0.0	3.9819	[0.0, 17.87]
Hepatitis_B	84.293	99	15.9955	[12, 99]
Measles	77.345	64	18.6597	[10, 99]
BMI	25.033	26.0	2.1939	[19.8, 32.1]
Polio	86.500	99	15.0804	[8, 99]
Diphtheria	86.272	99	15.5342	[16, 99]
Incidents HIV	0.894	0.08	2.3814	[0.01, 21.68]
GDP per capita	11540.925	554	16934.7889	[148, 112418]
Population	36.676	0.11	136.4859	[0.08, 1379.86]
Thinness 10-19 years	4.866	1.0	4.4382	[0.1, 27.7]
Thinness 5-9 years	4.900	0.9	4.5252	[0.1, 28.6]
Schooling	7.632	8.4	3.1716	[1.1, 14.1]
Eco. st.: Developed	0.207	0	0.4050	[0, 1]
Life expectancy	68.856	72.6, 73.1	9.4056	[39.4, 83.8]

Figure 1: Summary statistics for the WHO life expectancy dataset.

3 Data visualization

3.1 Outliers

To detect outliers, the box plots are very efficient graphs that can be used because they provide a concise visual summary of the distribution of a dataset, making it easy to spot any values that lie far outside the typical range. The box plots below, before (Appendix 12.) and after standardization (Figure 2.), will provide a useful graphical representation of the characteristics that pertain to the presence of outliers in the "Life Expectancy" dataset. The original box plot, which is not adjusted for scale, portrays the distribution of raw data across the attributes. This will be made possible to see the outliers as individual points

that fall outside the whiskers for the boxes. The main problem in interpreting outliers here lies in the different scales applied to the attributes (e.g 'GDP' attributes in Appendix 12.).

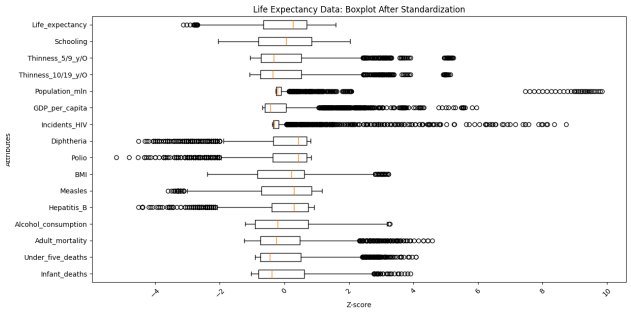

Figure 2: Outliers after standardization.

Attribute	Value
Incidents_HIV	77
Diphtheria	63
Thinness_5/9 y/O	55
GDP per capita	54
Polio	53
Thinness_10/19 y/O	53
Adult_mortality	49
Hepatitis_B	46
Measles	39
Under_five_deaths	28
Infant_deaths	14
BMI	13
Alcohol_consumption	5
Life_expectancy	2

Figure 3: Outliers

The value of the attributes are scaled such that all of them reside in a similar scale, where each attribute has zero mean and a standard deviation of 1. With outliers considered to be those points which fall outside the normal range of -2 to 2 standard deviations. After calculating and examining the z-scores for the standardized attributes in the current dataset, we observe the following regarding outliers (using an absolute z-score threshold of 3) (Figure 3.). Each outlier should ideally be evaluated in the context of the attribute it belongs to. E. g. high values in 'GDP-per-capita' might represent wealthier countries, while high ' Incidents-HIV' might indicate regions with severe HIV epidemics.

3.2 Normal Distribution

To visualize the attributes of the current dataset, histograms are a straightforward and effective tool for visually assessing the normality of attributes. They provide immediate insights into the distribution's shape. First of all, it represents the shape of the distribution of the entire dataset (Figure 4.) and the presence of some outliers are noticeable, so they have been removed to have a better visualization of the attributes distribution (Appendix 13).

We have made some general observations based on the histograms after the outliers removal (Figure 11). 'Infant Deaths', 'Under Five Deaths', and 'Adult Mortality' attributes show a skewed distribution, with a concentration of values at the lower end, indicating many countries have low rates of these indicators but a few have high rates, which is common in health-related data.

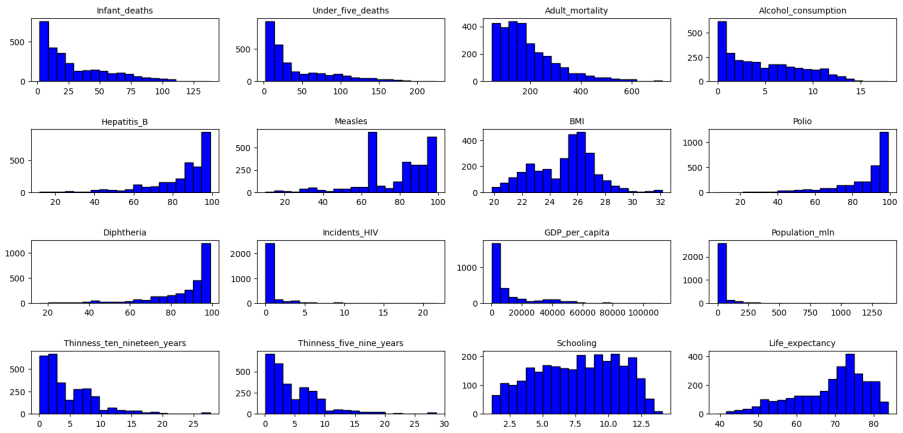

Figure 4: Histogram: distribution of the dataset

'Alcohol Consumption', 'Hepatitis B', 'Measles', 'BMI', 'Polio', 'Diphtheria', 'Incidents HIV', 'GDP per Capita', 'Population mln': These attributes might also show skewed distributions, typical for data where a range of values is possible but a significant number of observations cluster towards one end of the spectrum.

'Thinness Ten Nineteen Years', 'Thinness Five Nine Years', 'Schooling': Similar to the other health-related attributes, these might also not follow a normal distribution, potentially showing skewness based on the specific characteristics and conditions of the populations being studied.

'Life Expectancy': Although not visualized in the last set of histograms, life expectancy could potentially have a more symmetric distribution among different countries, but it can still be skewed due to variations in global health standards, access to medical care, and other socio-economic factors.

3.3 Data Correlation

It is important to analyze the correlation between attributes in our dataset. Since it can show patterns in the data which tells us the feasibility of our machine learning goals. Here we want to focus on which attributes have high correlation, meaning that they one can be explained by the other. And which attributes correlate with our regression and classification attributes, Life_expectancy and Development status. We have created a correlation matrix, which can be seen on figure 5.
From the correlations we can see some observations.

'Infant_deaths', 'Under_five_deaths' and 'Adult_mortality' all have high positive correlation with each other. This makes sense since they all are measures of deaths just in different age groups. Another observation is that all of these have high negative correlation with **Life_expectancy**, which is useful for

our regression tasks. This means that we could possibly explain a lot of the 'Life_expectancy' attribute using only one of the three others.

'Life_expectancy' has high positive correlation with immunization 'Polio', 'Diphtheria', 'Hepatitis_B' and also 'Schooling', and 'GDP_per_capita'. Here the only attributes which does not have a high correlation with 'Life_expectancy' is 'Population_mln' (0.026) and 'Year' (0.17). In general these two attributes do not have a strong correlation with any other attribute. For 'Year' the strongest are 'Under_five_deaths'(-0.18) and 'Hepatitis_B'(0.18). And it has no correlation with 'Economy_status_Developing'(0). This means that this attribute could be easily dropped. For 'Population' it's correlation with 'Life_expectancy' is 0.026, meaning that it could be dropped in this case, but with 'Economy_status' it has a medium negative correlation of -0.52, meaning that it is useful in our classification task.

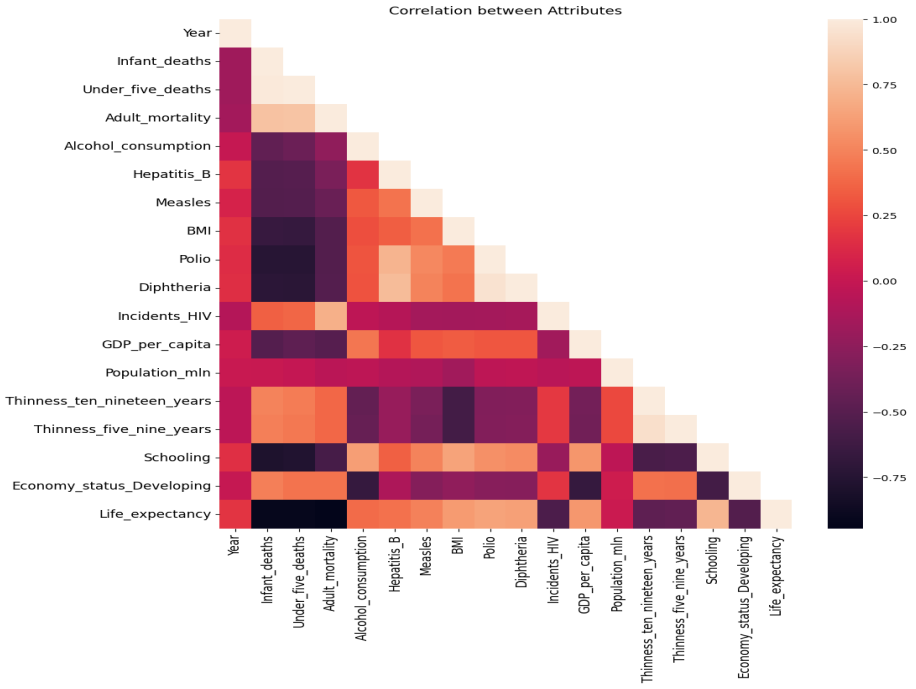

Figure 5: Correlation heat-map for attributes in the WHO Life Expectancy dataset

3.4 PCA

For the PCA analysis of the dataset, we only considered attributes that have a quantitative numerical meaning, therefore we ignored nominal attributes, namely the Country, Region and Economy status. We standardized the re-

maining attributes by subtracting their mean and dividing by their standard deviation.

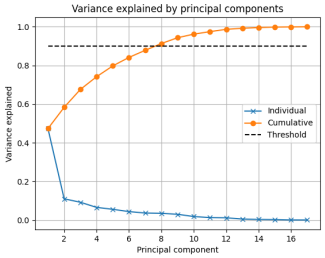

Figure 6: The amount of variation explained as a function of the number of PCA components included.

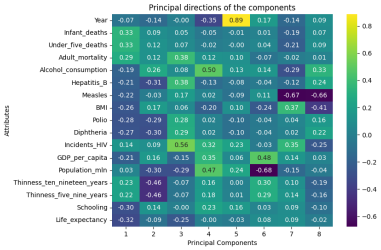

Figure 7: The principal directions of the considered PCA components.

It can be seen from Figure 6 that to be able to explain 90% of the variance, we need the first 8 principal components. We plotted these components in correspondence with the attributes of the data (see Figure 7). It can be seen that PC1 mainly splits the attributes into two categories: 1) Deaths, mortality, and thinness, and 2) Diseases, schooling, and life expectancy. It is also interesting to note that at PC5, most of the variance is explained by the year, which of course follows a uniform distribution including 16 years through the whole dataset.

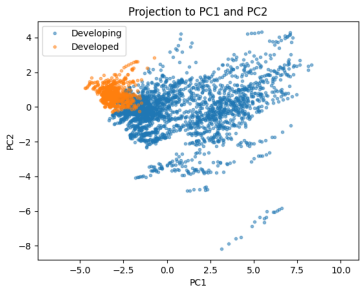

Figure 8: Projection of the data to the first two principal components, colored by the economy status.

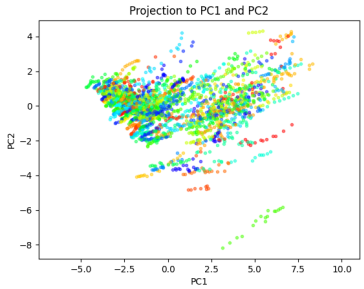

Figure 9: Projection to PC1 and PC2 colored by countries.

We chose to include the projection to the first two principal components. Since we are planning on performing classification on the economy status of the countries, we plotted the data with different colors for developed/developing countries on Figure 8. The first two PC-s cluster the developed countries together distinctively. We noticed linear correlations between certain data points

on the projection, and suspected that there correspond to the same country, only in different years. Therefore we changed the coloring to be based on the countries, and according to Figure 9 our assumption was correct. According to Figure 7, these linear "traces" introduce the biggest variance when reaching PC5.

4 Discussion

From our analysis of the dataset, we have found some important learning's that we will need in the coming modelling phase. Moreover we found that our original dataset contained a lot of missing and corrupted values. Here we found an updated dataset and confirmed that it only contains valid data. By calculating summary statistics and analyzing the distributions of our attributes, we found that it is necessary to standardize our data.

We found in our correlation analysis which attributes correlate with 'Life Expectancy' and 'Economic Status', which shows that our regression and classification goals are possible. This is also seen in our PCA analysis, where we see the principle projections cluster the data. With this we are ready to start analysing and evaluating different models for classification and regression.

5 Exam Questions

- Question 1: C is true
  - x1 ordinal, there is a certain order to the 30-minute intervals, here 8:30-9:00 can be thought of as less than 9:00-9:30, since it comes earlier in the day. This makes it ordinal
  - x6 (Traffic lights) is ratio since there can be some amount and no traffic light.
  - x7 (Running over) is the same as for x6, there can no run over accidents so it is ratio
  - y is ordinal, it is discrete with values 1-4, where Light congestion (2) is less than (<) Intermediate congestion (3).
- Question 2: A is true
  - The p-norm distance between two vectors $\mathbf{a}$ and $\mathbf{b}$ in R^n, for $p \geq 1$, is defined as:

$$\|\mathbf{a} - \mathbf{b}\|_p = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{\frac{1}{p}}$$

– For $p = \infty$, the distance is defined using the maximum norm. In this case the the maximum distance between the two vectors is 26-19 which is 7.

- Question 3: A is true
  The variances explained by the principal components can be seen in order in the diagonal of the matrix **S**. Considering this as a vector **s**, the following equation gives the proportion of the summed variance covered by principal components k:

$$\frac{\sum_{i \in k} \mathbf{s}_i}{\sum_j \mathbf{s}_j}$$

  Plugging in the numbers in **S** we can find out that only A is a correct statement.

- Question 4: D is true
  We can look at the attributes in question, what their principle value is and if the have high or low value.

Attribute	value	principle
Time of day	Low	-0.5
Broken Truck	High	0.23
Accident Victim	High	0.23
Defects	High	0.8

  We can see that Time of day is the only negative part of the component, with a low value it won't have a big effect. While Defects value is High and have 0.8 as it's component. This results in a positive projection.

- Question 5: A is true

  – Write the dictionary using all the words all the document (Max size = 20000)
  – Construct a boolean array for each document where 1 means that the word in the dictionary is in the current document and 0 is not.
    s1:

$$1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1$$

    s2:

$$0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1$$

  – Apply the Jaccard similarity definition to the arrays s1 and s2.

$$J(s1, s2) = \frac{|s1 \cap s2|}{|s1 \cup s2|}$$

- Question 6: B is true
  We can obtain the required probability by summing up all probabilities of the cases where $y = 2$ and $\hat{x}_2 = 0$. That includes two cases, therefore $0.81 + 0.03 = 0.84$

6 Appendix

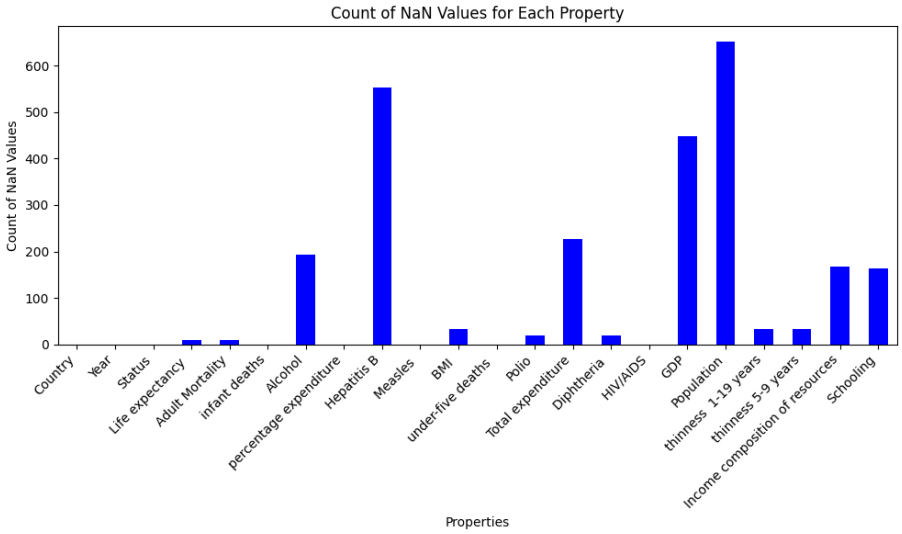

Figure 10: Missing values for the properties in the original dataset. The corrected dataset has no missing values at all.

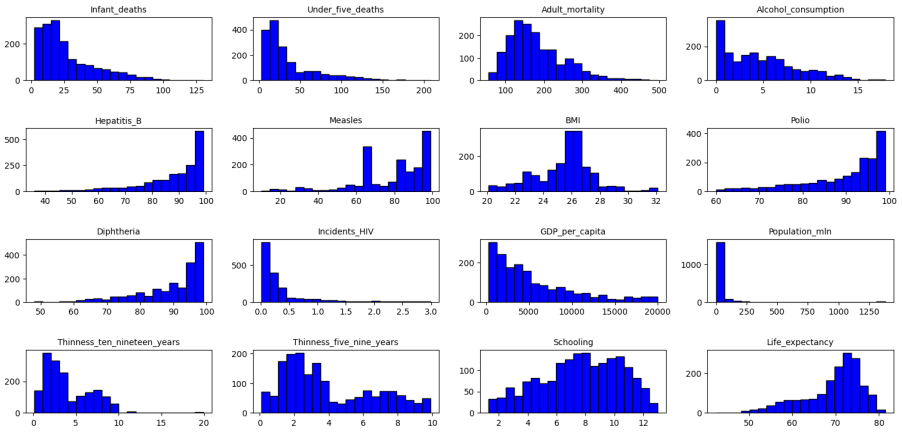

Figure 11: Histogram: distribution without outliers

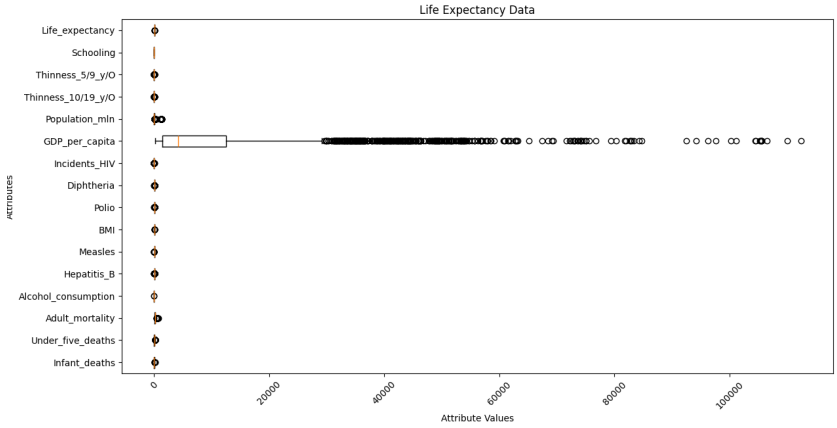

Figure 12: Outliers before standardization.

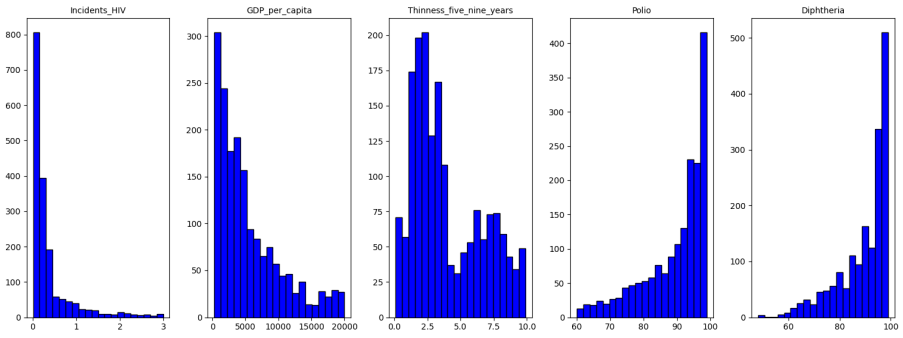

Figure 13: Enhanced histogram of some properties after removing the outliers.

Project description for report 2

Objective: The objective of this second report is to apply the methods you have learned in the second section of the course on *"Supervised learning: Classification and regression"* in order to solve both a relevant classification and regression problem for your data.

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 5 to 7 in order to see how the various tasks can be carried out.

Mandatory section

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

- The report must be submitted by a group of 3 students unless explicit permission has been given by the teachers to work alone. According to the DTU regulations, a student's contribution to the report must be clearly specified. Therefore, for each section, specify (in **a table on the frontpage**) who was responsible for it. **A report must contain this documentation to be approved**[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] Every team member is resonsible and must (ideally) contribute to all parts of the report. 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. For exam problems students are expected to contribute equally, and a student will not get any credit if they only contribute to the exam problems!.

[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.

1. Option $A/B/C/D$: To see this ...
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 assessed 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 **Deadline for handin is no later than Thursday 11 April 2024 at 17:00 CET via DTU Learn**. Late handins will not be accepted under normal circumstances.

Description

Project report 2 should naturally follow project report 1 on *"Data: Feature extraction, and visualization"* and cover what you have learned in the lectures and exercises of week 5 to 8 on *"Supervised learning: Classification and regression"*. The report

should therefore include two sections. A section on regression and a section on classification. The report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality.

Regression, part a: In this section, you are to solve a relevant regression problem for your data and statistically evaluate the result. We will begin by examining the most elementary model, namely linear regression.

1. Explain what variable is predicted based on which other variables and what you hope to accomplish by the regression. Mention your feature transformation choices such as one-of-K coding. Since we will use regularization momentarily, apply a feature transformation to your data matrix $\mathbf{X}$ such that each column has mean 0 and standard deviation 1[3].

2. Introduce a regularization parameter λ as discussed in chapter 14 of the lecture notes, and estimate the generalization error for different values of λ. Specifically, choose a reasonable range of values of λ (ideally one where the generalization error first drop and then increases), and for each value use $K = 10$ fold cross-validation (algorithm 5) to estimate the generalization error.

   Include a figure of the estimated generalization error as a function of λ in the report and briefly discuss the result.

3. Explain how the output, y, of the linear model with the lowest generalization error (as determined in the previous question) is computed for a given input $\mathbf{x}$. What is the effect of an individual attribute in $\mathbf{x}$ on the output, y, of the linear model? Does the effect of individual attributes make sense based on your understanding of the problem?

Regression, part b: In this section, we will compare three models: the regularized linear regression model from the previous section, an artificial neural network (ANN) and a baseline. We are interested in two questions: Is one model better than the other? Is either model better than a trivial baseline?. We will attempt to answer these questions with two-level cross-validation.

1. Implement two-level cross-validation (see algorithm 6 of the lecture notes). We will use 2-level cross-validation to compare the models with $K_1 = K_2 = 10$

[3] We treat feature transformations and linear regression in a very condensed manner in this course. Note for real-life applications, it may be a good idea to consider interaction terms and the last category in a one-of-K coding is redundant (you can perhaps convince yourself why). We consider this out of the scope for this report

Outer fold	ANN		Linear regression		baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	10.8	0.01	12.8	15.3
2	4	10.1	0.01	12.4	15.1
⋮	⋮	⋮	⋮	⋮	⋮
10	3	10.9	0.05	12.1	15.9

Table 1: Two-level cross-validation table used to compare the three models

folds[4]. As a baseline model, we will apply a linear regression model with no features, i.e. it computes the mean of y on the training data, and use this value to predict y on the test data.

Make sure you can fit an ANN model to the data. As complexity-controlling parameter for the ANN, we will use the number of hidden units[5] h. Based on a few test-runs, select a reasonable range of values for h (which should include $h = 1$), and describe the range of values you will use for h and λ.

2. Produce a table akin to Table 1 using two-level cross-validation (algorithm 6 in the lecture notes). The table shows, for each of the $K_1 = 10$ folds i, the optimal value of the number of hidden units and regularization strength (h_i^* and λ_i^* respectively) as found after each inner loop, as well as the estimated generalization errors E_i^{test} by evaluating on $\mathcal{D}_i^{\text{test}}$. It also includes the baseline test error, also evaluated on $\mathcal{D}_i^{\text{test}}$. Importantly, you must re-use the train/test splits $\mathcal{D}_i^{\text{par}}, \mathcal{D}_i^{\text{test}}$ for all three methods to allow statistical comparison (see next section).

   Note the error measure we use is the squared loss *per observation*, i.e. we divide by the number of observation in the test dataset:

$$E = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} (y_i - \hat{y}_i)^2$$

   Include a table similar to Table 1 in your report and briefly discuss what it tells you at a glance. Do you find the same value of λ^* as in the previous section?

[4] If this is too time-consuming, use $K_1 = K_2 = 5$

[5] Note there are many things we could potentially tweak or select, such as regularization. If you wish to select another parameter to tweak feel free to do so.

3. Statistically evaluate if there is a significant performance difference between the fitted ANN, linear regression model and baseline using the methods described in chapter 11. These comparisons will be made pairwise (ANN vs. linear regression; ANN vs. baseline; linear regression vs. baseline). We will allow some freedom in what test to choose. Therefore, choose either:

   setup I (section 11.3): Use the paired t-test described in Box 11.3.4

   setup II (section 11.4): Use the method described in Box 11.4.1)

   Include p-values and confidence intervals for the three pairwise tests in your report and conclude on the results: Is one model better than the other? Are the two models better than the baseline? Are some of the models identical? What recommendations would you make based on what you've learned?

Classification: In this part of the report you are to solve a relevant classification problem for your data and statistically evaluate your result. The tasks will closely mirror what you just did in the last section. The three methods we will compare is a baseline, logistic regression, and **one** of the other four methods from below (referred to as *method 2*).

Logistic regression for classification. Once more, we can use a regularization parameter $\lambda \geq 0$ to control complexity

ANN Artificial neural networks for classification. Same complexity-controlling parameter as in the previous exercise

CT Classification trees. Same complexity-controlling parameter as for regression trees

KNN k-nearest neighbor classification, complexity controlling parameter $k = 1, 2 \ldots$

NB Naïve Bayes. As complexity-controlling parameter, we suggest the term $b \geq 0$ from section 11.2.1 of the lecture notes to estimate[6] $p(x = 1) = \frac{n^+ + b}{n^+ + n^- + 2b}$

1. Explain which classification problem you have chosen to solve. Is it a multi-class or binary classification problem?

[6] In Python, use the `alpha` parameter in `sklearn.naive_bayes` and in R, use the `laplacian` parameter to `naiveBayes`. We do *not* recommend NB for Matlab users, as the implementation is somewhat lacking.

Outer fold	Method 2		Logistic regression		baseline
i	x_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	10.8	0.01	12.8	15.3
2	4	10.1	0.01	12.4	15.1
⋮	⋮	⋮	⋮	⋮	⋮
10	3	10.9	0.05	12.1	15.9

Table 2: Two-level cross-validation table used to compare the three models in the classification problem.

2. We will compare logistic regression[7], *method 2* and a baseline. For logistic regression, we will once more use λ as a complexity-controlling parameter, and for *method 2* a relevant complexity controlling parameter and range of values. We recommend this choice is made based on a trial run, which you do not need to report. Describe which parameter you have chosen and the possible values of the parameters you will examine.

   The baseline will be a model which compute the largest class on the training data, and predict everything in the test-data as belonging to that class (corresponding to the optimal prediction by a logistic regression model with a bias term and no features).

3. Again use two-level cross-validation to create a table similar to Table 2, but now comparing the logistic regression, *method 2*, and baseline. The table should once more include the selected parameters, and as an error measure we will use the error rate:

$$E = \frac{\{\text{Number of misclassified observations}\}}{N^{\text{test}}}$$

   Once more, make sure to re-use the outer validation splits to admit statistical evaluation. Briefly discuss the result.

4. Perform a statistical evaluation of your three models similar to the previous section. That is, compare the three models pairwise. We will once more allow some freedom in what test to choose. Therefore, choose either:

   setup I (section 11.3): Use McNemera's test described in Box 11.3.2)

[7] in case of a multi-class problem, substitute logistic regression for multinomial regression

setup II (section 11.4): Use the method described in Box 11.4.1)

Include p-values and confidence intervals for the three pairwise tests in your report and conclude on the results: Is one model better than the other? Are the two models better than the baseline? Are some of the models identical? What recommendations would you make based on what you've learned?

5. Train a logistic regression model using a suitable value of λ (see previous exercise). Explain how the logistic regression model make a prediction. Are the same features deemed relevant as for the regression part of the report?

Discussion:

1. Include a discussion of what you have learned in the regression and classification part of the report.

2. If your data has been analyzed previously (which will be the case in nearly all instances), find a study which uses it for classification, regression or both. Discuss how your results relate to those obtained in the study. If your dataset has not been published before, or the articles are irrelevant/unobtainable, this question may be omitted but make sure you justify this is the case.

The report itself should be maximum 10 pages long including figures and tables and give a precise and coherent account of the results of the regression and classification methods applied to your data.

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 13:

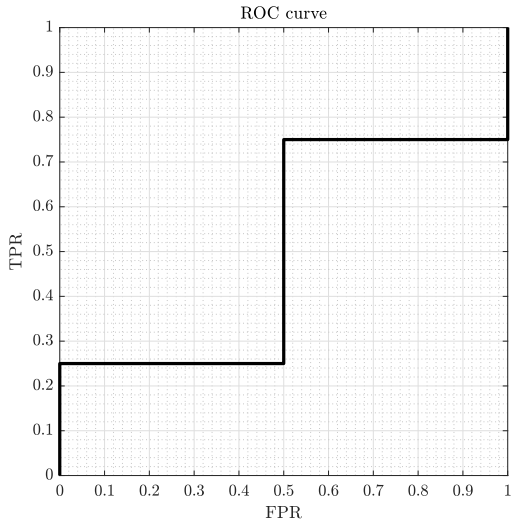

Figure 1: ROC curve for a neural network classifier, where the predictions and true class labels are one of the options in fig. 2.

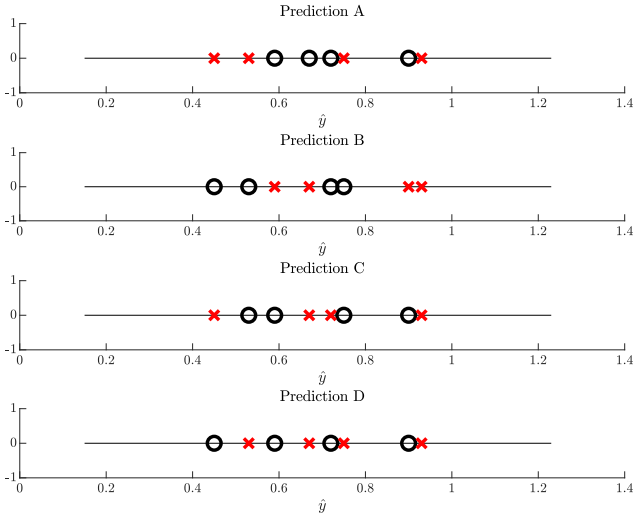

Figure 2: Four candidate predictions for the ROC curve in fig. 1. The observations are plotted horizontally, such that the position on the x-axis indicate the predicted value $\hat{y}_i$, and the marker/color indicate the class membership, such that the black circles indicate the observation belongs to class $y_i = 0$ and red crosses to $y_i = 1$.

A neural network classifier is trained to distinguish between two classes $y \in \{0, 1\}$ and produce class-probability $\hat{y}$ and the *reciever operator characteristic* (ROC) curve of the network when evaluated on a test set with $N = 8$ observations is shown in fig. 2. Suppose we plot the predictions on the $N = 8$ test observations by their $\hat{y}$ value along the x-axis and indicate the class labels by either a black circle (class $y = 0$) or red cross ($y = 1$), which one of the subplots in fig. 2 then corresponds to the ROC curve in fig. 1?

A Prediction A

B Prediction B

C Prediction C

D Prediction D

E Don't know.

Question 2. Spring 2019 question 15: Suppose we wish to build a classification tree based on Hunt's algorithm where the goal is to predict Congestion level which can belong to four classes, $y = 1$, $y = 2$, $y = 3$, $y = 4$. We consider binary splits based on the value of x_7, such that observations where $x_7 = z$ are assigned to the left branch and those where $x_7 \neq z$ are assigned the right branch. In table 3 we have indicated the number of observations in each of the four classes for the different values x_7 take in the dataset. Suppose we use the *classification error* impurity measure, which one of the following statements is true?

	$x_7 = 0$	$x_7 = 1$	$x_7 = 2$
$y = 1$	33	4	0
$y = 2$	28	2	1
$y = 3$	30	3	0
$y = 4$	29	5	0

Table 3: Proposed split of the Urban Traffic dataset based on the attribute x_7. We consider a two-way split where for each interval we count how many observations belonging to that interval has the given class label.

A The impurity gain of the split $x_7 = 2$ is $\Delta \approx 0.0195$

B The impurity gain of the split $x_7 = 2$ is $\Delta \approx 0.0178$

C The impurity gain of the split $x_7 = 2$ is $\Delta \approx 0.0074$

D The impurity gain of the split $x_7 = 2$ is $\Delta \approx 0.0212$

E Don't know.

Question 3. Spring 2019 question 18: We will consider an artificial neural network (ANN) trained on the Urban Traffic dataset described in table 4 to predict the class label y based on attributes $x_1, \ldots, x_7$. The neural network has a single hidden layer containing $n_h = 10$ units, and will use the softmax activation function (specifically, we will use the over-parameterized softmax function described in section 14.3.2 *(Neural networks for multi-class classification)* of the lecture notes) to predict the class label y since it is a multi-class problem. For the hidden layer we will use a sigmoid non-linearity. How many parameters has to be trained to fit the neural network?

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 4: 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 Network contains 124 parameters

B Network contains 280 parameters

C Network contains 110 parameters

D Network contains 88 parameters

E Don't know.

Question 4. Spring 2019 question 20:

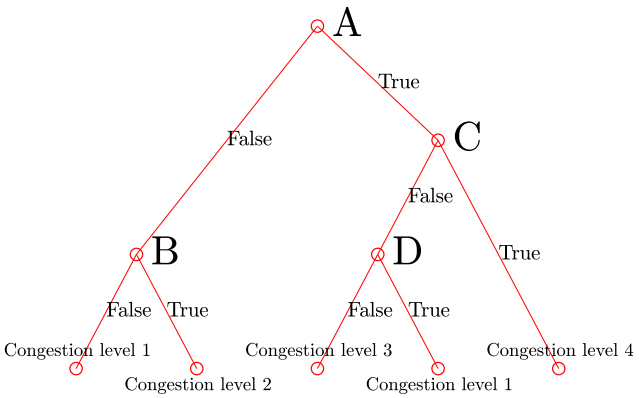

Figure 3: Structure of decision tree. The goal is to determine the splitting rules.

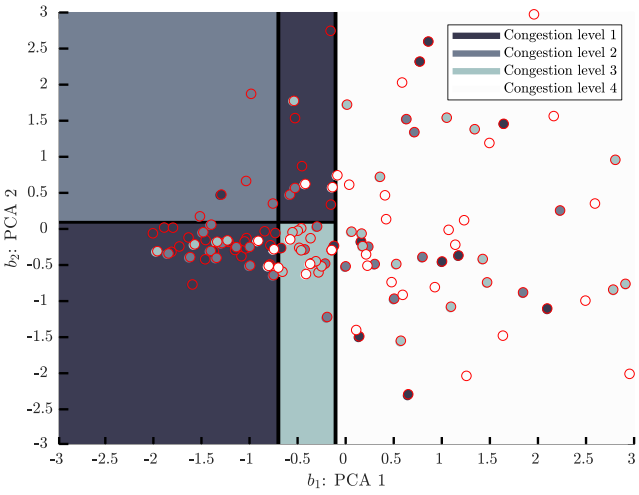

Figure 4: Classification boundary.

We will consider the Urban Traffic dataset projected onto the first two principal directions. Suppose we train a decision tree to predict which of the four classes an observation belongs to. Since the attributes are continuous, we will consider binary splits of the form $b_i \geq z$ for different values of i and z, where b_1, b_2 refer to the coordinates of the observations when projected onto principal directions. Suppose the trained decision tree has the form shown in fig. 3, and that according to the tree the predicted label assignment for the $N = 135$ observations are as given in fig. 4, what is then the correct rule assignment to the nodes in the decision tree?

A ***A***: $b_1 \geq -0.16$, ***B***: $b_2 \geq 0.03$, ***C***: $b_2 \geq 0.01$, ***D***: $b_1 \geq -0.76$

B ***A***: $b_1 \geq -0.76$, ***B***: $b_1 \geq -0.16$, ***C***: $b_2 \geq 0.03$, ***D***: $b_2 \geq 0.01$

C ***A***: $b_2 \geq 0.03$, ***B***: $b_1 \geq -0.76$, ***C***: $b_2 \geq 0.01$, ***D***: $b_1 \geq -0.16$

D ***A***: $b_1 \geq -0.76$, ***B***: $b_2 \geq 0.03$, ***C***: $b_1 \geq -0.16$, ***D***: $b_2 \geq 0.01$

E Don't know.

Question 5. Spring 2019 question 22:

	ANN		Log.reg.	
	n_h^*	E_1^{test}	λ^*	E_2^{test}
Outer fold 1	1	0.385	0.01	0.615
Outer fold 2	1	0.357	0.01	0.286
Outer fold 3	1	0.429	0.01	0.357
Outer fold 4	1	0.571	0.06	0.714
Outer fold 5	1	0.538	0.32	0.538

Table 5: Result of applying two-level cross-validation to a neural network model and a logistic regression model. The table contains the optimally selected parameters from each outer fold (n_h^*, hidden units and λ^*, regularization strength) and the corresponding test errors E_1^{test} and E_2^{test} when the models are evaluated on the current outer split.

Suppose we wish to compare a neural network model and a regularized logistic regression model on the Urban Traffic dataset. For the neural network, we wish to find the optimal number of hidden neurons n_h, and for the regression model the optimal value of λ. We therefore opt for a two-level cross-validation approach where for each outer fold, we determine the optimal number of hidden units (or regularization strength) using an inner cross-validation loop with $K_2 = 4$ folds. The tested values are:

$$\lambda : \{0.01, 0.06, 0.32, 1.78, 10\}$$
$$n_h : \{1, 2, 3, 4, 5\}.$$

Then, given this optimal number of hidden units n_h^* or regularization strength λ^*, the model is trained and evaluated on the current outer split. This produces table 5 which shows the optimal number of hidden units/lambda as well as the (outer) test classification errors E_1^{test} (neural network model) and E_2^{test} (logistic regression model). Note these errors are averaged over the number of observations in the the (outer) test splits. Suppose the time taken to train/test a single neural network model in milliseconds is

training time: 20 and testing time: 5

and the time taken to train/test a single logistic regression model is

training time: 8 and testing time: 1,

what is approximately the time taken to compose the table?

A 6800.0 ms

B 13600.0 ms

C 3570.0 ms

D 13940.0 ms

E Don't know.

Question 6. Spring 2019 question 26: Consider again the Urban Traffic dataset. We consider a multinomial regression model applied to the dataset projected onto the first two principal directions, giving the two coordinates b_1 and b_2 for each observation. Multinomial regression then computes the per-class probability by first computing the 3 numbers:

$$\hat{y}_k = \begin{bmatrix} 1 \\ b_1 \\ b_2 \end{bmatrix}^\top \boldsymbol{w}_k, \text{ for } k = 1, \ldots, 3$$

and then subsequently use the softmax transformation in the form:

$$P(y = k|\hat{\boldsymbol{y}}) = \begin{cases} \frac{e^{\hat{y}_k}}{1+\sum_{k'=1}^{3} e^{\hat{y}_{k'}}}. & \text{if } k \leq 3 \\ \frac{1}{1+\sum_{k'=1}^{3} e^{\hat{y}_{k'}}} & \text{if } k = 4 \end{cases}$$

to compute the per-class probabilities. Suppose the weights are given as:

$$\boldsymbol{w}_1 = \begin{bmatrix} 1.2 \\ -2.1 \\ 3.2 \end{bmatrix}, \boldsymbol{w}_2 = \begin{bmatrix} 1.2 \\ -1.7 \\ 2.9 \end{bmatrix}, \boldsymbol{w}_3 = \begin{bmatrix} 1.3 \\ -1.1 \\ 2.2 \end{bmatrix}.$$

Which of the following observations will be assigned to class $y = 4$?

A Observation $\boldsymbol{b} = \begin{bmatrix} -1.4 \\ 2.6 \end{bmatrix}$

B Observation $\boldsymbol{b} = \begin{bmatrix} -0.6 \\ -1.6 \end{bmatrix}$

C Observation $\boldsymbol{b} = \begin{bmatrix} 2.1 \\ 5.0 \end{bmatrix}$

D Observation $\boldsymbol{b} = \begin{bmatrix} 0.7 \\ 3.8 \end{bmatrix}$

E Don't know.