Technical University of Denmark

02450
INTRODUCTION TO MACHINE LEARNING AND DATA MINING

PROJECT 1
Data: Feature extraction and visualization

PROJECT 2
Supervised learning: Classification and regression

S182565 Miriam Mazzeo

Fall 2019

Section 1

Introduction

The overall project aims to address three main tasks:
- Finding a data set and perform useful transformation in order to prepare the data for analysis, such as filling the missing values and deleting redundant attributes.
- Analyzing and describing the data using principle component analysis and basic statistic.
- Solving both a relevant classification and regression problem for the data and statistically evaluate the results.

Section 2

Data-set Description [1] [2]

In this report, the "Countries of the World"[3] and "World Happiness Report"[4] data-sets have been combined and used for machine learning purposes. The two raw data-sets are composed as following:
- 'Country Data' data-set is composed by 227 instances and 20 attributes, including continuous, discrete, and nominal attributes. This dataset contains key statistical indicators of the countries: general information, such as population density, economic indicators, such as GDP, and social indicators, such as level of Literacy.
- 'Happiness Report' data-set is composed by 156 instances and 9 attributes, including continuous, discrete, and nominal attributes. "The data are scores based on answers to the main life evaluation question asked by the Gallup World Poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale." The scores are from nationally representative samples for the years 2013-2019 and use the Gallup weights to make the estimates representative. The columns, following the happiness score, estimate the extent to which each of six factors – economic production (GDP), social support, life expectancy, freedom, absence of corruption, and generosity – contribute to make life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. The scores differ for each country as they are the combination of 'residuals' with the estimate for life evaluations in Dystopia so that the score will always have positive values: the residuals reflect the extent to which the six variables either over- or under-explain average life evaluations, and have an average value of approximately zero over the whole set of countries. Although some life evaluation residuals are quite large, occasionally exceeding one point on the scale from 0 to 10, they are always much smaller than the calculated value in Dystopia, where the average life is rated at 1.85 on the 0 to 10 scale [4].

Motivation. Happiness is increasingly considered the proper measure of social progress and the goal of public policy. In a recent speech, the head of the UN Development Program (UNDP) spoke against measures of happiness based mainly on the GDP, arguing that what matters is the quality of growth."Paying more attention to happiness should be part of our efforts to achieve both human and sustainable development" she said.[1] Understanding the factors that affects the happiness score of a country, comparing economic and social factors, may be therefore an important quest to address in order to achieve a sustainable development of the countries that nowadays are classifies as "unhappy". This is also reported in the conclusion of the 'World Happiness Report': "broadening the focus from income to happiness greatly increases the number of ways of improving lives for the unhappy without making others worse off, and further, this can be achieved in more sustainable and less resource-demanding ways". The World Happiness Report found out that three-quarters of the differences among countries are accounted for by six key variables: GDP per capita, healthy years of life expectancy, social support (as measured by having someone to count on in times of trouble), trust (as measured by a perceived absence of corruption in government and business), perceived freedom to make life decisions, and generosity (as measured by recent donations)[1].
For this report, I've decided to investigate causes of happiness basing my variables on people scores and perspective and including other variables more related to statistical indicators, to have a complete overview on the different factors that may affect people's happiness perception in a certain country. The machine learning aim throughout the report will be to predict the happiness score or country and to classify a country as "happy" or "unhappy" based on 27 features.

Section 3

Data-set Attributes

This data set has 27 attributes, which contain both discrete and continuous attributes that fall into the nominal, ordinal, interval, and ratio categories. This is a list of all 22 attributes in this data set:

- 1: Region (score)
  - discrete, nominal
  - Possible values: names of 11 geographical areas such as Asia, Baltics, Northern Africa..
- 2: Social support (score)
  - continuous, ratio
  - possible values: floats between 0 and 1.85
- 3: Healthy life expectancy (score)
  - continuous, ratio
  - possible values: floats between 0 and 1.85
- 4: Freedom to make life choices (score)
  - continuous, ratio
  - possible values: floats between 0 and 1.85
- 5: Generosity (score)
  - continuous, ratio

    - possible values: floats between 0 and 1.85
- 6: Perceptions of corruption (score)
    - continuous, ratio
    - possible values: floats between 0 and 1.85
- 7: Pop. Density (ppl per sq. mi)
    - continuous, ratio
    - values: floats between 1 and 6.5K
- 8: Coastline (coast area/total area)
    - continuous, ratio
    - values: floats between 0 and 67.12%
- 9: Net migration (n.immigrants/n.emigrants)
    - continuous, ratio
    - values: floats between -11 and 24
    - if positive indicates that there are more people entering than leaving an area.
- 10: Infant mortality (deaths per 1000 births)
    - continuous, ratio
    - values: floats between 2 and 164
- 11: GDP (million $ per capita)
    - continuous, ratio
    - values: floats between 500 and 37.8K
- 12: Literacy (% of people ages 15 and above)
    - continuous, ratio
    - possible values: floats between 0 and 100%
- 13: Phones (per 1000 ppl)
    - discrete, ordinal
    - values: between 1 and 898
- 14: Arable (% of country total area)
    - continuous, ratio
    - possible values: floats between 0 and 100%
- 15: Crops (% of country total area)
    - continuous, ratio
    - possible values: floats between 0 and 100%
- 16: Climate
    - discrete, nominal
    - possible values: 1, 1.5, 2, 2.5, 3, 4
- 17: Birth rate (births per 1000 ppl)
    - continuous, ratio
    - values: floats between 7 and 51
- 18: Death rate (deaths per 1000 ppl)
    - continuous, ratio
    - values: floats between 2 and 30
- 19: Agriculture (% of economy stratification)

    - continuous, ratio
    - possible values: floats between 0 and 100%
  - 20: Industry (% of economy stratification)
    - continuous, ratio
    - possible values: floats between 0 and 100%
  - 21: Service (% of economy stratification)
    - continuous, ratio
    - possible values: floats between 0 and 100%
  - 22: Happiness Class
    - discrete nominal
    - this attribute is the 'output' for the classification problem, calculated taking into account the 'Happiness Score' attribute.
    - possible values: happy - if happiness score ¿ 5, unhappy - if happiness score ¿= 5
  - 22: Happiness Score
    - continuous, ratio
    - this attribute is the 'output' for the regression problem. It's a metric measured in 2019 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest"
    - possible values: floats between 0 and 10
  - Population, Total Area, Other (% of country total area), GDP per capita (score)
    - these are all continuous variable
    - due to the fact that these variables are redundant because other attributes account for them already (e.g. 'Population Density' accounts for both 'Population' and 'Total Area'), they have been excluded from the analysed data-set

The raw data, obtained by merging the two data-sets, is composed by the attributes listed above and 156 country instances. Approximately 20% of the instances contain missing values in the form of a "NaN". Moreover two attributes, 'Region' and 'Climate', contain categorical values, not suitable for PCA analysis. In order to deal with this issues:
- instances NaN are eliminated from the data-set, as estimating missing values would be unreliable and time-consuming;
- attributes, redundant/unnecessary, are eliminated as mentioned above;
- categorical variables are converted to one-out of-K coding;
- 'Region' attribute is eliminated as considered not so much suited for the analysis.

This leads to 25 attributes, including encoded 'Climate' columns, and 2 additional 'outcome' attributes, 'Happiness Score' and 'Happiness Class', that are going to be used respectively for regression and for classifications. For the statistic the classification variable is taken aside from the data-set, as this information will be the one predicted during the classification problem.

Data visualization and PCA

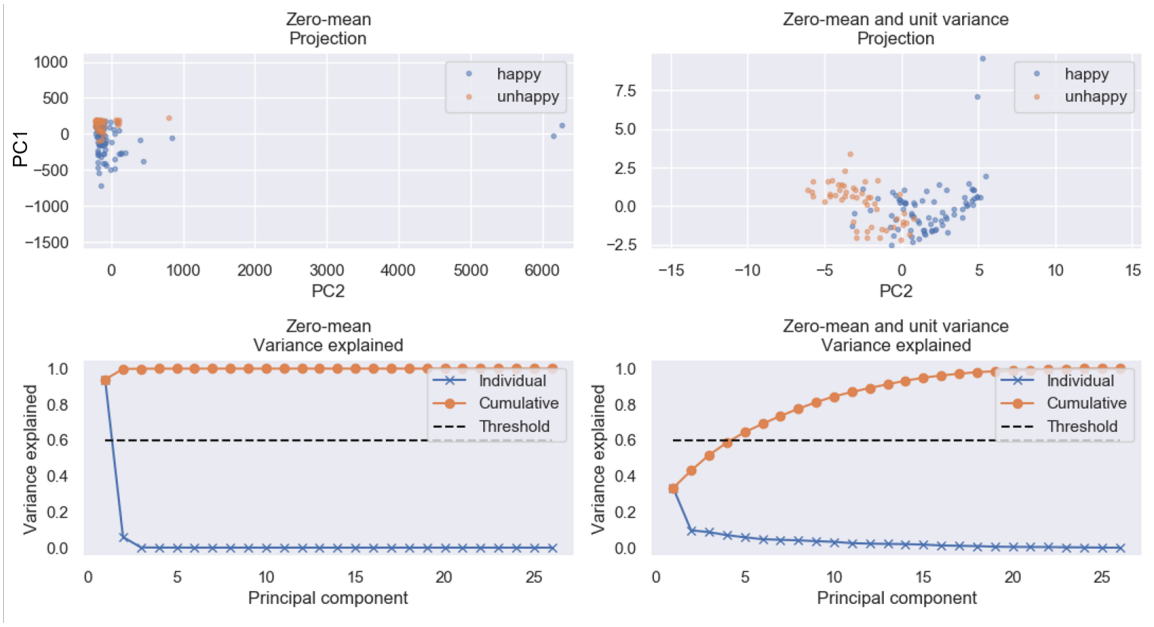


Figure 1: PCA plot and variance explained, with standardization on the right.

Standardization and Variance explained. The data-set presents variable different unit scales, therefore, it's necessary to standardize the data-set by scaling and centering, which ensures each attribute same mean and variance (in this case zero mean and unit variance). As shown in the scree plots (Fig 1 bottom) in the bottom graphs, the data-set variance is plotted against the principal components which describe it in order of importance. The blue line indicates the proportion of variance explained by each PCs, which is calculated by taking that principal component's eigenvalue divided by the sum of all eigenvalues; the orange line indicates the cumulative proportion of variance explained including all principal components up to that point. In the case with only centralization (Fig. bottom left), we can see that only 1 PCA explain all the data variance and that's clearly an abnormal result. The attributes have not being scaled yet and are not suited for comparison. The second figure (bottom right), on the contrary, needs almost 12 PCA to explain 90% of the variance. As usual, the first principal component is the most important, describing around 30% of the data variance. Looking at the threshold, we can observe that just four principal components are necessary to explain 60% of the variance.

PCA. The PCs are defined as a linear combination of the data's original variables, the coefficients are stored in a 'PCA loading matrix', which can be interpreted as a rotation matrix that rotates data and project each sample onto the dimension with greatest variance, along the first axis. In 1 upper left, data are represented after centralization and projection onto the first principal component

(PC1), in the PC1 - PC2 coordinate system. Some data points looks very distant from the area where most of the data are condensed, that may suggest presence of outliers and may also be suggested by plotting the attributes standard deviation histograms, where 'population density' and 'phones' standard deviation rockets. That's not due to real outliers present in the data but just an effect of comparing variables with different scales and units, producing results that don't have any statistical meaning. Standardization and, in some cases, normalization have to be applied to the data to investigate the attribute statistic. With standardised data, 1 upper right, the two classes, 'happy' and 'unhappy', are easier to separate. If we draw a vertical line in the middle of the plot, the majority of the points belonging to the class 'happy' would lay on the right, while the points belonging to the other class would mostly be condensed on the left. Only few data points have been misclassified and lie mainly in the middle of the two clusters and in the top-right of the figure. It can be argued that some form of correlation is shown among the points of the two classes and the first two principle components, which are very effective in separating the two data clusters. The attributes projected on the first two components, therefore, are likely to have high influence in the classification of the instances.

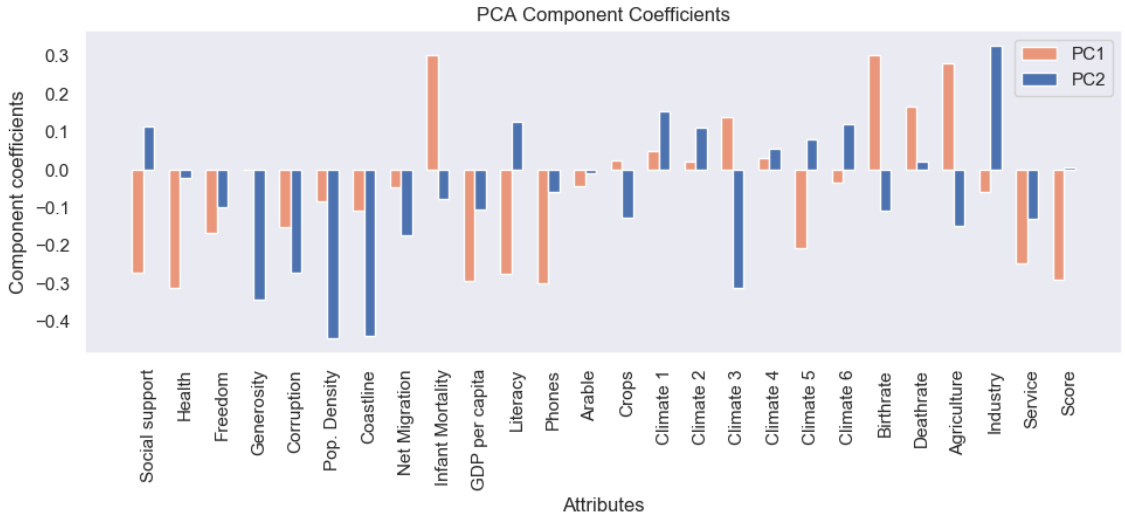
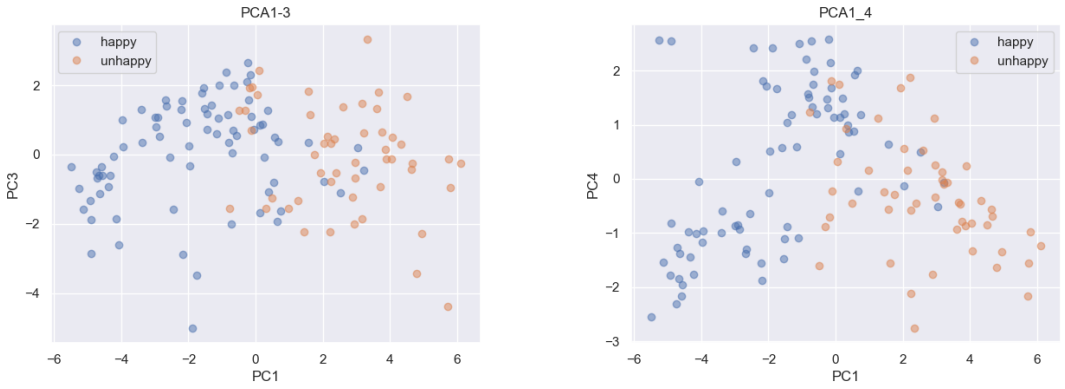


Figure 2: PCA Coefficient comparison

PCA coefficients. Looking at the PCA component coefficient, we can compare the contribution that each attribute is giving toward the first 2 Principle components. Fig 2, shows that 3 attributes are mainly positive contributing (with positive correlation) to the first principal components (Infant mortality, Deathrate, Agriculture), while mainly 7 attributes are negatively contributing to PC1 (Social support, Health, GDP, Literacy, Phones, Service and Happiness score). The second principal components appear instead mainly affected by Generosity, Corruption, Population density, Coastline, Climate 3 and Industry.

Other PCAs. In Figure 3a, the data doesn't appear to be normally distributed, but it's somewhat possible to distinguish different groups (the unhappy countries are on the right and happy ones on the left). However, this is not a super clear relationship due to some happy country samples present



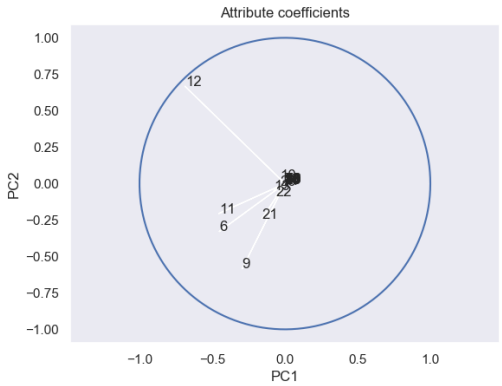
(a) PCA plot of PC1 and PC3

(b) PCA plot of PC1 and PC4

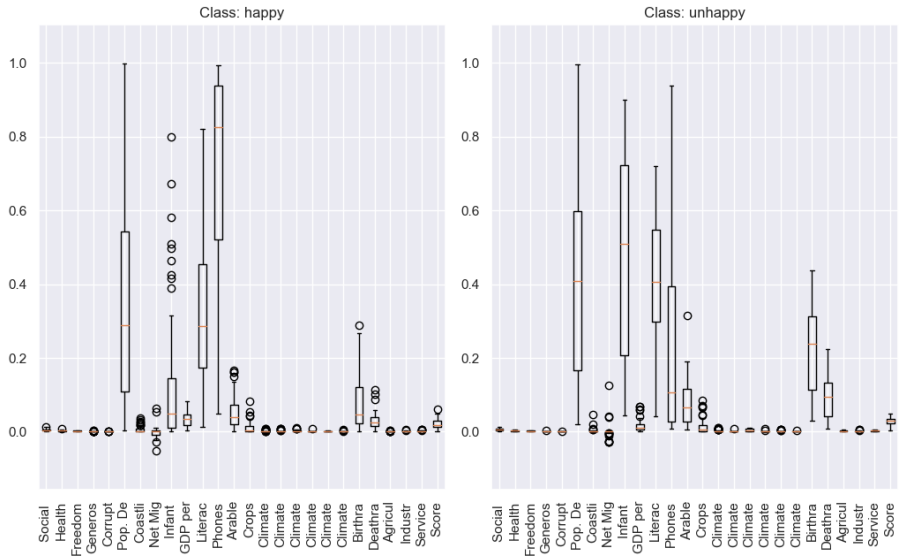
Figure 3

on the right side too but, anyway, in a similar number as the misclassifications in the plot pca1 vs pca2 1. Same pattern can be observed in Figure 3b, showing correlation between PC1 and PC4. A sort of correlation, therefore, may be inferred among the first four principal components, but not strong since data points appear very dispersed and not so much confined in a thigh space.

PC space and Attributes analysis. Another way of interpreting the principal directions is by plotting the coefficients as vectors in the PC1/PC2-space. In Fig 4, ones can interpret the relationship between principal components and attributes every vector represent the attribute coefficient, pointing in positive direction of PC1, for example, if the attribute has a positive projection onto PC1. In the plot shown in Figure 4a, the attributes' projected values on each PC show how much weight they have on that PC and the angles between the vectors show how they are correlated with one another. In Fig 4, normalization has been performed on the data to be able to visualize the attributes contribution to the principal components in the PC-space and the contribution of each attribute in determining a specific class 4b, excluding issues with different units that make the attributes unsuitable for comparison. The influence of some variable on PC1 and PC2 is easily recognizable in the plots: 'Phones' (12), 'Literacy' (11), 'Pop. Density' (6), 'Infant Mortality'(9) highly influence both PC1 and PC2 almost in the same way, as they have diagonal directions (in between the two components). Moreover, in the loading plot is possible to infer eventual correlations between attributes: attributes 11, 6, 9 and 21, present vectors very close to one another, showing positive correlation; this attributes, on the other hand, form an angle of around 90◦, showing negative correlation between each other.In Fig 4b, we can observe the attribute contribution in determining the class chosen. 'Phones' attribute seems to have a big impact in classifying a country as 'happy', as we can see the values range between 0.45 and 0.9 for happy country while between 0 and 0.4 for 'unhappy' ones. Also 'Infant mortality' seems to give one of the highest contribution to classification as its box in 'happy' countries lies under the value of 0.2 while for 'unhappy' ones lies completely above this value. 'Birthrate' and 'Deathrate' are also quite interesting value, with overall higher values for 'unhappy' countries, as expected.



(a) PC coefficients Loading plot after normalization



(b) Boxplot of attribute values per class plot after normalization

Figure 4

Section 5

Conclusion

To summarise analysis and results from the above section:
- Redundant data has been removed and a usable data frame has been created for future supervised learning analysis.
- The data don't present outliers but they cannot be compared without performing regularization as they present very different units.

- The data-set attributes present different scales therefore it's necessary to standardize the data by scaling and centering, which is ensuring each attribute has same mean and variance, especially to perform PCA and statistic analysis.
- Almost 12 principal components are necessary to explain 90% of the variance, the first principal component is the most important as it describes around 30% of the accumulated data variance; analysing the first four principal components appears to be good enough to explain data variance (around 60%).
- Looking at the scatter plot of PC1 against the other three principal components, the majority of the points belonging to the different classes ('happy' and 'unhappy' countries) appear mainly concentrated in different areas of the plot, it's almost always possible to divide the two groups with a straight line or to cluster in two main groups, with very few misclassification. This suggest correlation among the principal components and that the attributes will be feasible for the primary machine learning modeling aim, especially for the classification problem.
- To project the attributes onto the PC-space, normalization of the data is needed to eliminate the different scales of the variables. The longest vectors correspond to the contributions of attribute 'Phones' (12), 'Literacy' (11), 'Pop. Density' (6), 'Infant Mortality'(9), these cariables highly influence both PC1 and PC2 almost in the same way, having similar weights and directions.
- Analysing attribute vs classes (to predict), 'Phones', 'Infant mortality', 'Birthrate' and 'Deathrate' contribute the most to the classification of an 'happy' country.
- Considering above analysis, the data-set looks suitable for the study of regression and classification, as the attributes seems to be correlated and the classes looks to be easily to cluster in the PCA.

Section 1

Introduction

This project is a continuation of the first project and focuses on classification and regression for the data set in question. For this project I will explore two types of supervised learning, regression and classification. Since regression is for continuous outputs and classification is for discrete outputs, two different output variables have been chosen, as explained also previously.

Section 2

Regression A

A.1. Regression maps a data-function onto a continuous valued output. For this data set, I will attempt to predict the Happiness score of a country taking into consideration all the other variables (26 in total) of the dataset, as described in Section 4. Happiness score is a continuous variable so it's suitable to be predicted with regression. One-out-of-K encoding was performed to create binary variables for the multi-class variable 'Climate', which includes 6 different climate classification and therefore is disassembled in 6 different binary variables. The data for the regression part of this assignment were separated into input and output variables and were normalised in order to ensure that all the variables are considered with a comparable unit.

A.2. The simplest regression model is a linear regression model,to this model we introduce a regularization parameter, λ, for controlling model complexity. The range for λ should be in a range where the estimated generalization error first drops and then increases. This in order to find the optimal λ by testing a set of different lambda for each of the 10-fold (cross-validation) and determining the one minimizing the generalization error.

A.3. In Fig.5, for each fold of train-test set, λ has been varied between 10^{-5} and 10^{9} in order to find the optimal lambda value corresponding to the minimum generalization error, and, afterwards, being able to select the best model. The training and test error were normalized by the number of observations of the data-set, we see that the generalization error is stable till 100, having a minimum when λ is reaches its optimum value at 10, thereafter, as expected, the error increases as λ increases. The fact that minimum λ is large may signify that the model have high bias. On the other end, looking at the residuals in 6a, they don't differ too much from the ones for not regularised linear regression, they are overall smaller even if not so symmetric. Looking at the scatter plot, a linear trend is much more evident with linear regression than with regularised one, therefore, the regularization factor doesn't seem to improve the model and it does not make sense to use it in this case.

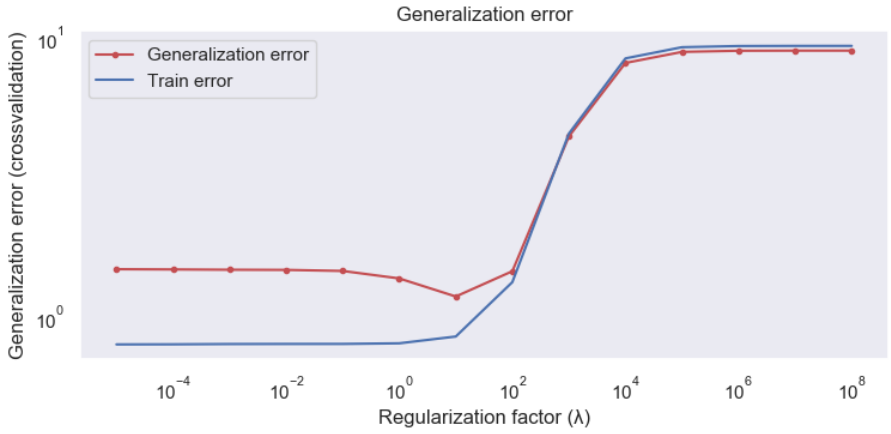
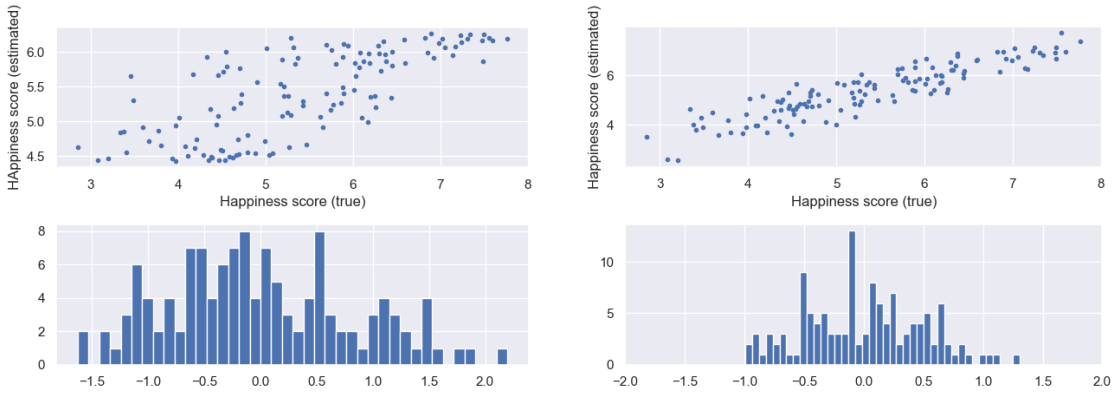


Figure 5: Generalization error vs regularization factor



(a) Ridge regression: Optimal λ for $K = 10$

(b) Linear Regression

Figure 6

Section 3

Regression B

In this section two-level 10-cross-validation (as in algorithm 6 of the lecture notes) has been implemented to compare three models: baseline model that computes the mean of y on the training data and use this value to predict y on the test data, regularised linear regression model that uses lambda as complexity-controlling parameter and artificial neural network model using the number of hidden units as complexity-controlling parameter. After few test-runs, a reasonable range of values for the complexity-controlling parameter has been set: for lambda, 50 values between 1e-10 and 1e-1, for h, integer values between 1 and 17. The data matrix, used to create the set of train and test

data, is composed by 25 columns, including all the attributes summarised in the previous report and excluding the attribute that we desire to predict with the regression model analysis. The variable used for prediction is the happiness score per country, a continuous rate. Looking at the regularised

Outer Fold	ANN		Ridge Linear Regression		Baseline
	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	7	0.48	1e-10	0.41	1.53
2	6	0.31	0.1	0.41	0.97
3	13	0.49	1e-10	0.18	0.79
4	8	0.11	0.1	0.44	1.25
5	15	0.31	0.1	0.23	2.23
6	1	0.29	0.1	0.36	1.41
7	6	0.54	0.1	0.19	1.24
8	1	0.47	1e-10	0.59	1.29
9	5	0.76	1e-10	0.09	1.67
10	11	0.15	1e-10	0.31	1.19
E gen	0.39		0.31		1.36

Table 1: Two-level cross-validation table used to compare a Regularised Linear Regression model (Ridge Regression), an Artificial neural network model and baseline model to address the regression problem.

linear regression results and compared to the previous model with just one level cross-validation, the optimal lambda values appear much smaller than before. For larger values of , the solution is dragged towards the x-axis because the bias becomes high. A model with a lambda of 0.1 could be a good candidate for describing the data, on the other hand too small lambda values, such as 1e-10, may reflect a solution with very high variance.

Statistical Evaluation. To estimate which method is the best suited for a particular problem, and therefore to make progress, we need a systematic method for comparing models. Looking at the generalization mean error for each model, giving indication on the capability of the model of making good predictions, linear regression seems to have the best performance. It has been decided to perform a McNemar's test among the models:

- Between Linear and ANN models the p value accounts for 0.08103335744678428 and the confidence interval for -0.197: 0.033. The low p value indicates gives evidence that the Linear regression model performs better than ANN but the confidence interval contains 0, gives weak evidence towards the first model being more accurate than the other.
- Between linear regression and baseline models, we have interval of -1.3 : -0.8 and p=1.5e-13, p-value is very small but the interval lies in the negative range therefore any conclusion can be made onto this two models at comparison. The same happens between baseline model and ANN.

Classification

This section aims to solve a binary classification problem: whether a country is 'happy' or 'unhappy'. Three classification methods have been compared: baseline, logistic regression, and k-nearest neighbor classification. Two-level 10-cross-validation (as in algorithm 6 of the lecture notes) have been implemented to compare three models. The baseline model compute the largest class on the training data, and predict everything in the test-data as belonging to that class; regularised logistic regression model uses lambda as complexity-controlling parameter and k-nearest neighbor classification model uses the number of nearest neighbors as complexity-controlling parameter. After few test-runs, a reasonable range of values for the complexity-controlling parameter has been set: for lambda, 50 values between 1e-10 and 1e-1, for K, 40 integer values between 1 and 40. The data matrix, used to create the set of train and test data, is composed by 25 columns, including all the attributes summarised in the previous report and excluding the attribute that we desire to predict with the classification analysis. The variable used for prediction is the "happiness class" of a country, a discrete nominal attribute (binary variable).

Outer Fold	Ridge Logistic Regression		KNN		Baseline
	λ^*	E (%)	K*	E (%)	E (%)
1	0.042955	0.15	4	0.23	0.38
2	0.0009	0.31	17	0.23	0.46
3	0.1	0.23	1	0	0.23
4	0.0655	0	17	0.08	0.46
5	0.0001	0.46	16	0.15	0.38
6	0.1	0.08	4	0.23	0.61
7	0.06551	0.23	1	0.15	0.08
8	0.1	0.25	26	0.25	0.58
9	0.0022	0.08	16	0.17	0.33
10	0.06551	0.25	1	0.17	0.5
Average	0.54	0.23	10	0.16	0.4

Table 2: Two-level cross-validation table used to compare a Regularised logistic Regression model (Ridge Regression), a k-nearest neighbor classification model and baseline model to address the classification problem.

Statistical Evaluation. To estimate which method is the best suited for the classification problem investigated, It has been decided to perform a McNemar's test among the models:
- for comparison between logistic and KNN we obtain low p=0.15 but negative confidence interval, therefore it's not possible to rule out if one model is more accurate than the other.
- for comparison between logistic and baseline positive confidence interval is calculated (0.09 : 0.25) and very low p-value p= 0.0045, this prove higher accuracy and better performance for the Ringe

Logistic classification over the baseline.
- comparing KNN to baseline, it is obtained CI = (0.188, 0.315) and p-value of p= 4.4e-07, this clearly indicate that KNN perform better than baseline as p-value is really small and confident interval is well clear from zero.

Section 5

Contribution

This is a project for a re-exam that I've decided to make anew on my own and need to be evaluated.

Bibliography

[1] Helliwell, J. & Layard, R. J. sachs (2017), world happiness report 2017. *Sustainable Development Solutions Network, New York, http://worldhappiness. report/ed* (2017).

[2] Book, C. F. The world fact book. *available online at: https.//www. cia. gov/library/8. world. fact publications/the-book/-cached (accessed 16 August 2013)* (2013).

[3] Kaggle.com. Countries of the world (2017). URL `https://www.kaggle.com/fernandol/countries-of-the-world`.

[4] Kaggle.com. World happiness report (2019). URL `https://www.kaggle.com/fernandol/countries-of-the-world`.