Report 02 22/04/2022

TECHNICAL UNIVERSITY OF DENMARK

02450 Introduction to Machine Learning and Data Mining F22

Groupe 115

Bashar Bdewi S183356 Malaz Alzarrad S180424

ABSTRACT

The objective of this project is to apply the methods we have learned in the second section of the course, " Supervised Machine Learning methods, Classification and Regression" on the iris data set to get a basic understanding.

Split of responsibility	
Mandatory section	50 % Bashar 50% Malaz
Problem Definition	50 % Bashar 50% Malaz
Regression Part A & B	Bashar
Classification	Malaz
Discussion	50 % Bashar 50% Malaz

1 …… Mandatory Section

Q1, answer for question 1 is C, to get the ROC curve, we have to calculate (FPR) and the (TPR). If we consider FPR = 0.5, the prospect TPR for predicting A are 0.5, for predicting B are 0.5, 0.75, and 1, for predicting C are 0.25, 0.5, and 0.75, and for predicting D, 0.25 and 0.5. So C matches the ROC curve.

Q2, answer to question 2 is C. we have to use the classification error function to solve this question by using this form ClassError(v) = 1 − maxP(c|v)
We have to find totally samples 33 + 4 + 28 + 2 + 1 + 30 + 3 + 29 + 25 = 135 in x7=2 we can see only one 1 sample in class 2 and no sample in other classes, that means we have only one sample in branch one and 134 sample in other branch.To find maxP(c|v) we divide 134 with 135, as stated by classification error function above, the impurity should be 1/135
The impurity gain of the split x7 = 2 is $\Delta \approx 0.0074$.

Q4, answer to question 4 is D. By using the structure of the decision tree,step C is the key step to solve this question. because it is self-evident that if step C is correct, Congestion level 4 will be split out and when b1>-0.16 that means step tree is correct so the answer should be D.

Q5,answer to question 5 is C. For logistic regression and ANN, the inner fold K2 = 4, the outer fold K1 = 5, and there are 5 and 5 h. In addition, the generalisation error for the optimal parameters must be calculated one more time in each outer fold. As a result, the total time is $4 \times 5 \times 5 \times 25 + 4 \times 5 \times 5 \times 9 + 5 \times (25+9) = 3570$ ms.

2 …… Problem Definition

The Iris Dataset contains four features (length and width of sepals and petals) of 150 samples, 50 samples each class. There are three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms.

This data set consists of the physical parameters of three species of flower — Versicolor, Setosa and Virginica. The numeric parameters which the dataset contains are Sepal width, Sepal length, Petal width and Petal length. In this data we will be predicting the classes of the flowers based on these parameters. The data consists of continuous numeric values which describe the dimensions of the respective features. We will be training the model based on these features.

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. species: – Iris Setosa – Iris Versicolour – Iris Virginica

- Classification problem

The current variables X and y represent a classification problem, in which a machine learning model will use the sepal and petal dimensions (stored in the matrix X) to predict the class (species of Iris, stored in the dependent variable y).

- Regression problem

Since the variable we wish to predict is petal length, petal length cannot any longer be in the data matrix X. The first thing we did is storing all the information we have in the other format in one data matrix. The petal length corresponds to the third column in the data matrix and therefore our new y variable is petal length.

Similarly, our new X matrix is all the other information but without the petal length (since it's now the new y variable) and since the iris class information (which is now the last column in the matrix of features X) is a categorical variable, we will do a one-out-of-K encoding of it.

Now, X is of size 150x6 corresponding to the three measurements of the Iris that are not the petal length as well as the three variables that specifies whether a given observations is or isn't a certain

type. We need to update the attribute names and store the petal length name as the name of the target variable y for regression.

We standardize the dataset by subtracting the mean and dividing by standard deviation for each attribute, so each attribute has a mean of 0 and a variance of 1.

3 …… Regression Part A

In this part, we use X1, X2, X3 and X4 as features to predict the value of Y (petal length). We apply the following linear regression model:

$$y_i = f(x_i , \omega) = \tilde{x}_i\omega$$

And regularization parameter λ influence the performance of our regression model and we get the relationship of the weights and λ:

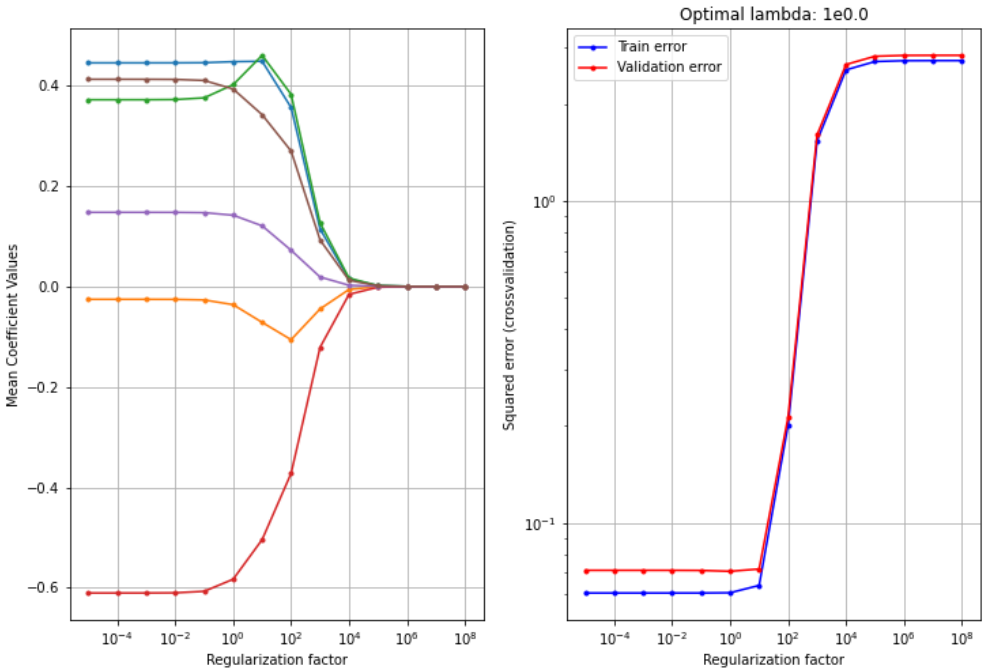
$$\omega = (\hat{X}^T \hat{X} + \lambda I)(\hat{X}^T \hat{y})$$

When the λ becomes larger, the weights are getting smaller,and as consequence the features would have a smaller influence on the model, so there is more likely to lead to high bias and low variance or overfitting.

The rang of λ: $\lambda \in [10^{-2} : 10^{8}]$

1 level cross validation is applied to train the linear regression model on the dataset and K-fold cross validation will be applied. The iris dataset is small and so K = 5 will be applied. In each fold, the squared loss per observation will be computed based on different λ as the performance evaluation for both training set and test set. 2 lists of average squared loss will be returned ,the first for training error and second for test error.

The λ corresponds to the least mean squared error among the test errors will be selected as the optimal λ.



Squared error with different regularization factors

And so, we can get the weights for all the features with optimal regularization parameter λ

<u>Petal Length in last fold:</u>

OffsetSepal Length 3.68

OffsetSepal Width 0.49

OffsetPetal Width -0.08

OffsetIris-setosa 0.47

OffsetIris-versicolor -0.52

OffsetIris-virginica 0.15

Linear regression without feature selection:

- Training error: 0.06580392069166542

- Test error: 0.07146326149759359

- R^2 train: 0.9786957083879935

- R^2 test: 0.9764446520557583

Regularized linear regression:

- Training error: 0.06584056054585148

- Test error: 0.07127174897728546

- R^2 train: 0.9786838460835897

- R^2 test: 0.9765077774149569

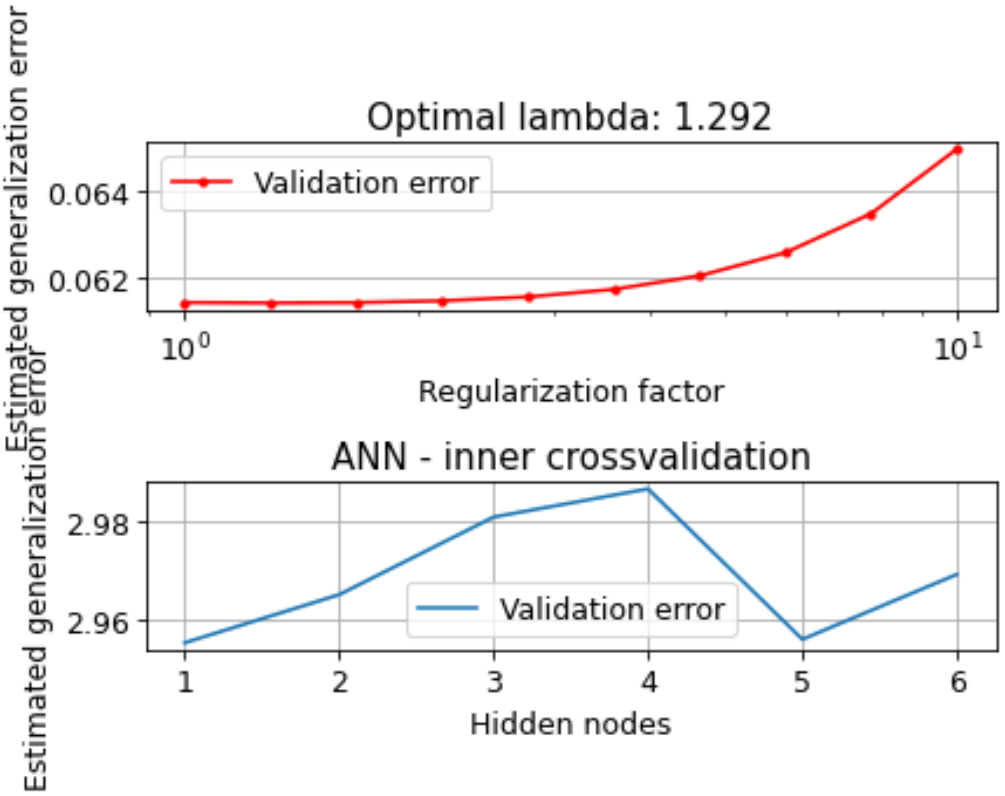
4 …….. Regression Part B

Linear regression, ANN (Artificial Neural Network) and baseline models are evaluated by applying two level cross validation. The inner folds is to select the optimal complexity controlling parameter, **λ** for linear regression model and **h** for ANN model. The outer folds find the squared loss based on the optimal complexity controlling parameter. K1 = K2 =5 because iris dataset is relatively small.

In the linear regression model, the λ range λ ∈ [1 : 10]

In ANN model, there is only one hidden layer, and the range of hidden unit h is: h ∈ [1 : 6]

The baseline is linear regression model with no features, it computes the mean of y on the training dataset, and then predict every sample in the test dataset to that value.



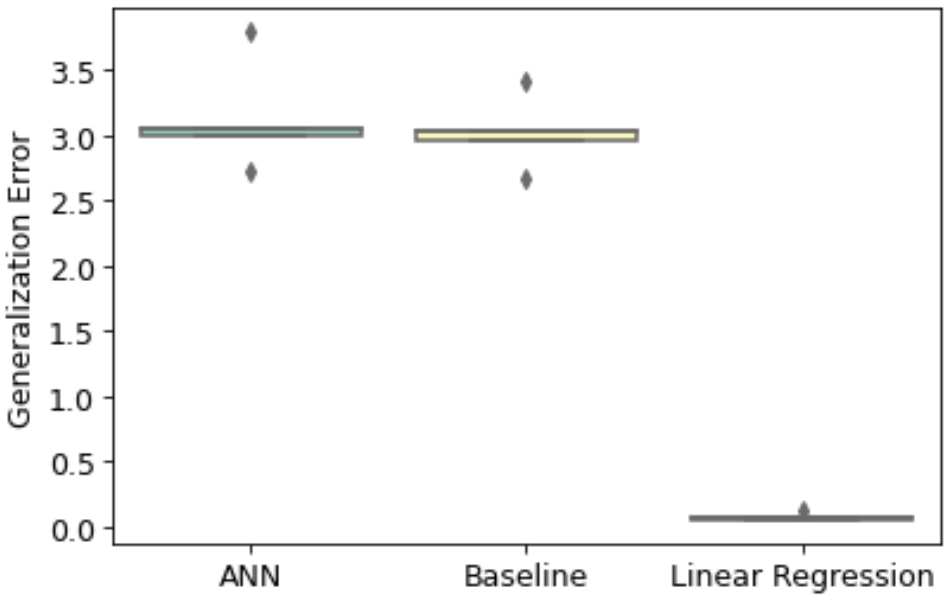
The optimal complexity controlling parameters (λ , h)

Below is a summary of two-level CV for the three regression models :

Errors ANN	Errors Baseline	Errors Linear
2.993	2.967	0.121
3.797	3.418	0.076
3.056	3.04	0.061
3.055	3.04	0.054
2.725	2.663	0.053

According to the result, the best test error of the Linear regression model is 0.053, the best test error of the ANN model is 2.725, and then the best test error of the Baseline is 2.663.

From boxplots here, we can see generalization errors of the different three models and we conclude that the linear regression is the best model in this problem because it has the least generalization error.



Boxplot of generalization errors from outer level of cross-validation for Linear Regression, ANN and the baseline model

According to the above results, it can be concluded that the linear regression has the best performance among the three models, the performance of ANN model and Baseline model are similar , but the baseline is better than ANN. We conclude also that when λ is small, the weights are large indicating high variance and low bias.

We apply now setup I to evaluate performance and statistically comparing the performance of the models as paired t-test applied with $\alpha = 0.05$. we can see below that linear regression is the best one and the two p-values for the the paired t-test are both very low indicating strong evidence against the models being similar.

A : Baseline B : LIN

CI: [2.57 3.34] p: 2.0489824530524218e-32

A : Baseline B : ANN

CI: [-0.29 0.09] p: 0.15203445797270254

A : LIN B : ANN

CI: [-3.45 -2.66] p: 1.4286187526919566e-32

5 …… Classification

We want now to classify the Y (species of Iris) given the features X1, X1, X3,and X4 (length and width of sepals and petals) into three classes, which are Iris Setosa – Iris Versicolour – Iris Virginica.

The classification problem is multi-class classification problem. We do classification using logistic regression, K-Nearest-Neighbors (KNN) and baseline.

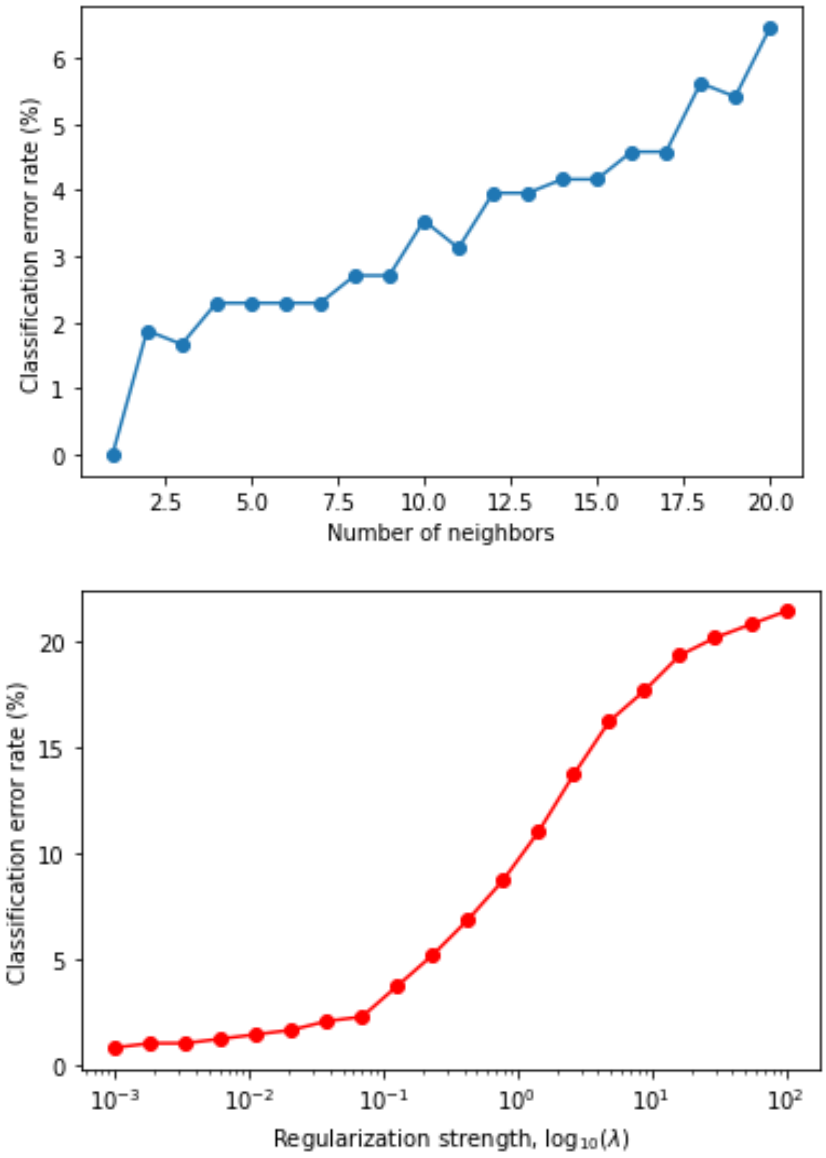
In Logistic regression model, One-vs.-rest strategy is used for reducing the problem of multiclass classification to multiple binary classification problems. One-vs.-rest strategy includes training a single classifier per class, with the samples of that class as positive samples ,and so all other samples as negative. The regularization factor λ controls the complexity and after several rounds of run, λ is selected in the range: $\lambda \in [-3, 2]$

In K-Nearest-Neighbors model,k is the complexity controlling parameter. After several rounds of trial run, this range of K is selected: $K \in [1, 20]$

The baseline computes the largest class on the training data, and predict everything in the test-data as belonging to that class.

Two level K-fold cross validation is used with K1 = K2 = 5. In the inner folds, the optimal complexity controlling parameters will selected,which are **λ** for logistic regression and **K** for KNN. In each outer fold, the error rate will be calculated based on the selected parameter of its inner folds.

E = Number of misclassified observations / N_{test}



Classification test error rate for complexity controlling parameters (λ, k)

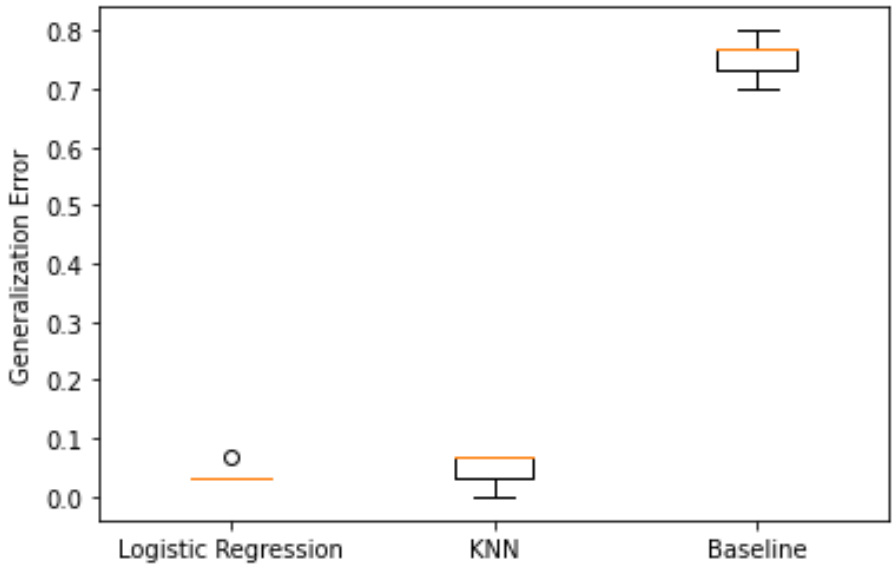
From this figure we get the optimal regularization parameters λ and the optimal k-nearest neighbor K.

Below are the error rates calculated based on the selected parameters (λ, k):

Errors KNN	Errors baseline	Errors LOG-REG
0.0	0.8	0.03
0.03	0.73	0.03

0.07	0.77	0.03
0.07	0.7	0.03
0.07	0.77	0.07

we can see that Baseline model has the worst performance with average classification error rate of ≈ 70%. The logistical regression and KNN models perform better with an average classification error rate of ≈ 7% and ≈ 3%, respectively. Boxplot of all the errors are plotted here for a more description.



Boxplots of generalization errors for Logistic Regression, KNN and the baseline model

Baseline method is resulting in more errors, but we still need to check which algorithm is performing better than other. Therefore, we do a comparison of algorithms by applying setup I to statistically evaluate if there is a significant performance difference between the three models pairwise.

The McNemar's test was used to estimate the difference in performance $\theta = \theta_A - \theta_B$ between model A (M_A) and model B (M_B). If $\theta > 0$, then model M_A is preferable over model M_B

Here is Pairwise statistical evaluation of the three classification models:

A : Baseline B : KNN

Result of McNemars test using alpha= 0.05

Comparison matrix n

[[34. 3.]

[109. 4.]]

Approximate 1-alpha confidence interval of theta: [thetaL,thetaU] = (-0.7814473445130594, -0.6231058903677106)

p-value for two-sided test A and B have same accuracy (exact binomial test): p= 9.022943270855311e-29

theta: -0.71 CI: [-0.78 -0.62] p: 0.0

A : Baseline B : Logistic Regression

Result of McNemars test using alpha= 0.05

Comparison matrix n

[[33. 4.]

[111. 2.]]

Approximate 1-alpha confidence interval of theta: [thetaL,thetaU] = (-0.7894497197300749, -0.6277973035975222)

p-value for two-sided test A and B have same accuracy (exact binomial test): p= 3.4507445333269973e-28

theta: -0.71 CI: [-0.79 -0.63] p: 0.0

A : KNN B : Logistic Regression

Result of McNemars test using alpha= 0.05

Comparison matrix n

[[141. 2.]

[3. 4.]]

Approximate 1-alpha confidence interval of theta: [thetaL,thetaU] = (-0.03576362182409454, 0.022435858191499403)

p-value for two-sided test A and B have same accuracy (exact binomial test): p= 1.0

theta: -0.01 CI: [-0.04 0.02] p: 1.0

We can see that there is a relatively large difference in performance between Logistic regression, KNN and baseline model. The performance difference theta is estimated to be around (theta = -0.71) and we inspected that zero is not in the confidence interval and p-value is zero, which is is very strong statistical evidence that the Logistic regression and KNN models are better than the baseline model.

For the KNN and Logistic regression model, the confidence interval contains zero which shows a weak evidence towards that Logistic Regression has higher accuracy than KNN and the p-value is relatively high therefore, we do not have sufficient evidence to conclude Logistic regression is better than KNN.

6 …… Discussion

We started with the most elementary model, namely linear regression.We solved a relevant regression problem for iris data and statistically evaluated the result. We noticed that the weights of optimal λ is same as we got from PCA analyse at the last project.

In the regression part B, we have compared three models: the regularized linear regression model , that we have done, an artificial neural network (ANN) and a baseline. And we evaluated whether one model better than the other. Linear regression had the best performance and the performance of ANN was as poor as baseline , which is out of our expectation and maybe we have to try feature transformations or feature selection before applying ANN.

In the classification problem, the logistical regression and KNN models perform better than baseline model.

References

02450 Exercises

02450 Toolbox

02450 Text Book …Introduction to Machine Learning and Data Mining (Tue Herlau, Mikkel N. Schmidt and Morten Mørup)