TECHNICAL UNIVERSITY OF DENMARK

02450 Introduction to Machine Learning and Data Mining

Project 2 - Supervised learning: Regression and classification

Group 127

Gabriela Penarska - s223289
Shane Lim Zi En - s236971
Ying Qi Tiffanie Leong - s236763

Section	Contribution
2. Regression a	Gabriela: 50% Shane: 30% Tiffanie: 20%
3. Regression b	Gabriela: 20% Shane: 50% Tiffanie: 30%
4. Classification	Gabriela: 30% Shane: 20% Tiffanie: 50%
5. Discussion	All 100%
6. Exam Questions	All 100%

Table of Contents

1. Introduction

In the second project we will continue the analysis of the detection and diagnosis of breast cancer dataset originated from Diagnostic Wisconsin Breast Cancer Database [1]. This iteration will focus on the implementation of the supervised learning techniques - regression and classification. In the regression problem we want to predict the mean radius of the tumour based on all the other attributes. In the classification problem we aim to develop a predictive model that can accurately classify new breast mass cases into one of the two classes: benign or malignant.

2. Regression Part A

In the previous project we applied data visualizations techniques and PCA analysis in order to gain insights into the dataset. We removed the columns *id* and *Unnamed: 32*, due to the fact that *id* does not contain any valuable information and *Unnamed: 32* contains only nulls. The features were also scaled to a similar range by applying standardisation (we subtracted the mean and divided by the standard deviation) to ensure that all the attributes contribute equally to the model. To handle the categorical variables we encoded them into numerical format by using one-of-K coding where M (malignant) is 1 and B (benign) is 0 in the class label *diagnosis*. The classes are equally represented, where the count ratio is close to 0.6 so it is acceptable.

In the regression problem we want to predict a variable *radius_mean* as *y*, which is the mean of distances from center to points on the perimeter. We can gain insights into the size and growth pattern of breast tumours, which is crucial for the treatment. For the prediction of this attribute we selected all of the other attributes as independent variables X based on the previous analysis.

In linear regression there are two potential issues. First if X contains many linearly dependent rows or when the linear regression model can overfit, therefore a regularization factor λ is introduced to control the complexity of the model and stabilise the coefficients. To avoid underfitting and overfitting and maintain a balance between the variance and bias we selected the range of λ between $\lambda_1 = 10^{-6}$ and $\lambda_{15} = 10^{9}$which shows the area with the lowest generalization error. Generalization errors corresponding to the both range boundaries stabilize in those points. We wanted to find the range of values of the parameter λ where the generalization error drops and then increases, however we can notice that the E_{gen} does not decrease at all for the low values. A noticeable rise occurred starting from $\lambda = 10^{-1}$, but the lowest E_{gen} was found in $\lambda = 10^{-2}$. The generalization error results in 0.0032.

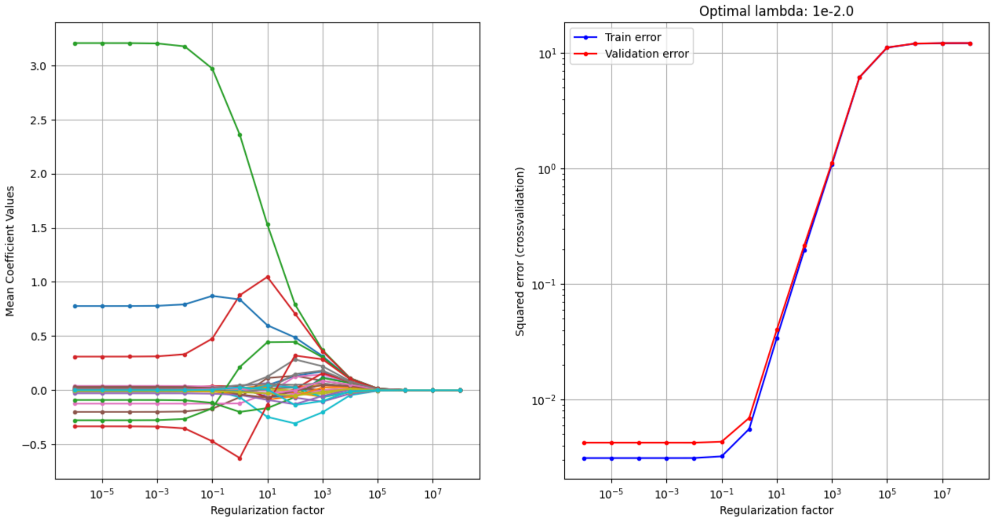

Figure 1. Estimated generalization error for different values of λ

To evaluate how the output y is predicted for a given input X, of the linear model with the lowest generalization error, it is necessary to assess the corresponding weight associated with the attributes with the regularization parameter $\lambda = 10^{-2}$ (Table 1). The magnitude of the weight indicates the strength of the relationship between the corresponding attribute and the target variable. Larger magnitudes imply a stronger influence of the attribute on the target. The sign of the weight (positive or negative) indicates the direction of the relationship. A positive weight means that as the value of the attribute increases, the predicted value of the target variable also increases. A negative weight means that as the value of the attribute increases, the predicted value of the target variable decreases. Attributes with larger magnitude weights are typically considered more important in the prediction process. In our analysis for some of the attributes the calculated weights seem acceptable and compatible with the results from the previous project. The strongest correlation and biggest influence we can see for *perimeter_mean,* for which that was also a case in the conducted data visualiztion. Moreover, the attributes such as perimeter and area of the tumour can directly affect its size. This suggests that tumours with larger boundaries and occupied areas tend to have larger mean radii, indicating potential tumour growth and size. This is reflected in the model. It is also expected that *radius_worst* correlates positively with the target variable. What is surprising is that *perimiter_worst* and *area_worst* correlate negatively, which is contrary to the previous analysis and general assumptions. While *concave_points_mean* gave previously high correlation, now it doesn't seem to play a significant role in determining tumour size.

Table 1. The weight of each attribute in linear regression model with a regularization parameter $\lambda = 10^{-2}$

diagnosis -0.01	texture_mean 0.0	perimeter_mean 3.26	area_mean 0.33	smoothness_mean 0.02
compactness_mean -0.2	concavity_mean -0.13	concave points_mean 0.01	symmetry_mean 0.01	fractal_dimension_mean 0.01
radius_se 0.03	texture_se 0.01	perimeter_se -0.09	area_se 0.02	smoothness_se 0.01
compactness_se -0.01	concavity_se 0.04	concave points_se 0.02	symmetry_se 0.0	fractal_dimension_se -0.01
radius_worst 0.8	texture_worst -0.01	perimeter_worst -0.31	area_worst -0.34	smoothness_worst -0.02
compactness_worst 0.06	concavity_worst 0.0	concave points_worst 0.0	symmetry_worst -0.01	fractal_dimension_worst -0.01

3. Regression Part B

In this part three models will be compared to determine if the other models perform better than the baseline model.

3.1.

The complexity of an ANN is often controlled by the number of hidden units, denoted as h and the regularization strength, denoted as λ. This section outlines the process and rationale for selecting the range of values for these hyperparameters.

To identify a suitable range for h and λ, preliminary test runs were conducted, which included $h = 1$ as a starting point to assess the model's ability to learn from the data. The range for h was selected based on the performance of the model, gauged by the mean squared error (MSE) during cross-validation. The objective was to identify a range where the model starts to improve in performance without showing signs of overfitting. For λ, the goal was to strike a balance between model flexibility and the prevention of overfitting. The regularization term penalizes the complexity of the model, thus limiting the risk of overfitting to the training data.

After the initial exploratory phase with varying hidden layers and regularization strengths, the following observations were made:

A single hidden unit (h = 1) resulted in high MSE values, indicating that such a simplistic model cannot capture the complexity of the data. Increasing the number of hidden units to h = 5 and h = 10 improved the performance but did not reach optimal levels, suggesting that the model complexity was still insufficient. The model with h = 50 hidden units demonstrated a significant decrease in MSE, suggesting an improvement in learning capability. With h = 100 hidden units, the model achieved the lowest MSE, indicating a strong learning capacity without apparent overfitting. Accordingly, the selected range for h was set from 1 to 100, recognizing that further refinement might be necessary depending on additional model diagnostics and validation. For λ, a wide range of values from 0.0001 to 1 was tested. It was found that as λ increased, the MSE initially decreased and then plateaued, suggesting that an optimal balance was found. Therefore, the chosen range for λ was also from 0.0001 to 1.

Table 2. Comparison of the test errors of each model corresponding to each outer fold

Outerfold	ANN		Logistic Regression		Baseline
i	h_i	E_i^{test}	λ_i	E_i^{test}	E_i^{test}
1	100	59.847653	0.001	0.003304	11.951937
2	100	57.318406	0.001	0.0055	15.763366
3	100	59.038105	0.001	0.004957	14.267581
4	100	61.227092	0.001	0.005122	10.994816
5	100	53.047357	0.001	0.003929	9.905257

Based on the preliminary testing, a reasonable range for the number of hidden units h was established to be from 1 to 100. For the regularization strength λ, the selected range was from 0.0001 to 1. These ranges provide a balance between model accuracy and complexity, ensuring that the ANN can generalize well to unseen data.

3.2.

This section presents a comparative analysis of three predictive models: a regularized linear regression model, an artificial neural network (ANN), and a baseline model which predicts the mean of the training data. We employed two-level cross-validation to evaluate model performance and to discern any statistical significance in their predictive capabilities. The assessment was based on the squared loss per observation.

Two-level cross-validation with K1=K2=5 folds was implemented to estimate the generalization error of each model. The baseline model predicted based solely on the mean of y from the training data, serving as a simplistic benchmark. The ANN's complexity was controlled by adjusting the number of hidden units h, with a preliminary run conducted to determine a sensible range for h and the regularization parameter λ. The selection criterion for the optimal h and λ was the minimization of the mean squared error (MSE) on the validation set within each fold's inner loop.

The results from the two-level cross-validation are summarized in Table 2. It depicts the optimal values of the number of hidden units (h) and regularization strength (λ) for each fold, alongside the corresponding generalization errors (E_i^{test}).

The best-performing ANN model, with h = 100 and λ =1 achieved a MSE of 0.0742, signifying a substantial improvement over the baseline's MSE of 12.5766. The regularized linear regression model exhibited consistent performance, achieving a Mean MSE of 0.0046, which is an order of magnitude lower than that of the baseline model, thereby indicating superior prediction accuracy.

Statistical tests comparing the models' MSE distributions confirmed that both the ANN and linear regression models significantly outperformed the baseline model (p-values: 8.59×10^{-6} and 3.04×10^{-4}, respectively). The linear regression model also statistically outperformed the ANN (p-value: 2.08×10^{-6}).

In conclusion, the linear regression model with feature regularization emerged as the most effective model in this analysis, demonstrating both high accuracy and robustness across different data partitions. The ANN provided significant predictive improvements over the baseline but did not reach the performance of the regularized linear regression model. These findings suggest that for the given dataset and problem, linear regression with optimal regularization is the preferred predictive model.

3.3.

This section provides a detailed statistical analysis comparing the performance of three predictive models: an artificial neural network (ANN), a regularized linear regression model, and a baseline model that predicts the mean of the dependent variable. Utilizing the mean squared error (MSE) obtained from a two-level cross-validation procedure, we assessed whether significant differences exist between the models' predictive accuracy.

We applied a paired t-test approach to evaluate the pairwise differences in the models' MSE across cross-validation folds, as per the methodology described in Method 11.3.4. We used a 95% confidence interval to infer the average differences in MSE between the models.

The calculated confidence intervals for the pairwise comparisons are described in Table 3.

Table 3. Statistical comparisons

Pairwise Comparison	Confidence Interval	p-value
Linear Regression vs. ANN	(-0.098, -0.033)	2.08×10^{-6}
ANN vs. Baseline	(-15.48, -9.53)	8.59×10^{-6}
Linear Regression vs. Baseline	(-15.55, -9.59)	3.04×10^{-4}

In terms of Confidence intervals:
Linear Regression vs. ANN: The confidence interval for the difference in MSE between the Linear regression and ANN models is (-0.098, -0.033). The negative interval indicates that the Linear model significantly outperforms the ANN model.

ANN vs. Baseline: The confidence interval for the difference in MSE between the ANN and Baseline model is (-15.48, -9.53). he negative interval indicates that the ANN model significantly outperforms the baseline model.

Linear Regression vs. Baseline: The confidence interval for the difference in MSE between the Linear Regression and Baseline model is (-15.55, -9.59). This negative interval reaffirms the Linear Regression model's superiority over the baseline model.

In terms of p-value:
The Linear regression model exhibits a statistically significant improvement over the ANN and baseline models, corroborated by both the confidence intervals and the associated p-values:

These p-values indicate a high statistical significance in the performance differences, leading to robust conclusions about the relative efficacy of the models.

Based on the statistical analysis, the regularized linear regression model is recommended for predictive tasks on this dataset due to its significantly lower MSE compared to the baseline models. The ANN can also be considered as the MSE and is only marginally lower than the linear regression model. However, due to its complexity and cost intensive nature it is important to consider if an ANN model will be necessary.

Further, it's advisable to extend the model validation beyond MSE as a sole metric, exploring alternative measures of performance to gain a more comprehensive understanding of the models' predictive capabilities.

4. Classification

In the classification task, we aim to predict the diagnosis of breast cancer, which is a categorical attribute indicating whether a breast mass is benign (B) or malignant (M). The classification problem described is a **binary** classification problem. This is because the target attribute, which is the diagnosis of breast cancer, has only two possible classes: benign (B) or malignant (M). Therefore, it is not a multi-class classification problem where there would be three or more distinct classes to predict.

We have fitted the data to a logistic regression as well as an ANN (Artificial Neural Network). We have chosen to use an optimal $\lambda = 0$ as a regularisation term is not necessary in this case. This is because the classification task is of a binary nature and there is a considerable number of observations present in the dataset. Next, for the range of hidden units in the ANN, we have decided on using $h_i = 3$ based on the test runs conducted. Lastly the data was also fitted to a baseline model that predicts every area value in the test data to be the training data's target attribute's mode. Since the mode for our data is benign (B), this model predicts every observation as benign (B).

In order to observe the differences in the performance of these 3 models, two-level cross-validation is used. The results of this can be seen in the table below. In the table, E_i^{test} refers to the test errors of each model corresponding to each outer fold (the misclassification rate for the *i*-th fold of cross-validation).

Table 4. Comparison of the test errors of each model corresponding to each outer fold

Outerfold	ANN		Logistic Regression		Baseline
i	h_i	E_i^{test}	λ_i	E_i^{test}	E_i^{test}
1	3	0.070175438596491	0	0.01754385964912	0.38488576449912
2	3	0.035087719298245	0	0.01754385964912	0.3655536028119
3	3	0.052631578947368	0	0.03508771929824	0.34270650263620
4	3	0.035087719298245	0	0.03508771929824	0.35852372583479
5	3	0.105263157894736	0	0.07017543859649	0.37609841827768
6	3	0.105263157894736	0	0.03508771929824	0.39718804920913
7	3	0.035087719298245	0	0.01754385964912	0.38664323374340
8	3	0.210526315789473	0	0.01754385964912	0.4305799648506
9	3	0.105263157894736	0	0.05263157894736	0.40070298769771
10	3	0.125	0	0.10714285714285	0.35325131810193

It is quite clear from the table that the highest test error values recorded came from the baseline model which is in accordance with our expectations and knowledge. Next the highest values were noted for ANN model and lastly for Logistic Regression.

Next, we proceed to perform a statistical evaluation of the above 3 models. We chose to use Setup I - the McNemar Test in order to compare the performance of each model. The comparisons were done pairwise to facilitate comparison. In the table below, we included the p-values and confidence intervals for the three pairwise tests. In the last column, we have the θ value where $\theta = \theta_A - \theta_B$. This θ value compares if model A is better than B. A positive θ value will indicate that model A is better than B while a negative θ value will indicate that model B is better than model A.

Table 5. Statistical evaluation of the 3 models

Pairwise Comparison	Confidence Interval	P-Value	θ
Baseline (A) vs Logistic Regression (B)	(-0.359298, -0.276304)	7.56535003614067e-44 ≈ 0.00	-0.318101933
Baseline (A) vs ANN (B)	(-0.344827, -0.262699)	1.2334209947574997e-41 ≈ 0.00	-0.304042179
Logistic Regression (A) vs ANN (B)	(-0.0059769, 0.0340908)	0.229481013	0.0140597539

The p-value for the two-sided test comparing model A (baseline) and model B (logistic regression) is approximately 7.57e-44 ≈ 0.00, which is significantly less than the significance level alpha of 0.05. Therefore, we reject the null hypothesis that the two models have the same accuracy. We can conclude that one model is indeed better than the other. The θ value is negative, indicating that logistic regression is a better model than the baseline. Based on these findings, it is recommended to use model B (logistic regression) over the baseline model A, as it demonstrates statistically significant improvements in accuracy. Similarly, the p-value for the two-sided test comparing model A (baseline) and model B (ANN) is approximately 1.23e-41 ≈ 0.00, which is significantly less than the significance level alpha of 0.05. Therefore, we also reject the null hypothesis that the two models have the same accuracy and can similarly conclude that the models are statistically different. The θ value is negative, indicating that ANN is a better model than the baseline. Hence, model B (ANN) should be used over baseline model A.

Next, we compare the two-sided test between model A (logistic regression) and model B (artificial neural network). The calculated p value is relatively high (0.229), suggesting that there is no significant difference in accuracy between the two models. This p value is greater than the alpha level of 0.05 hence we do not reject the null hypothesis that the two models have the same accuracy. There is therefore insufficient evidence to conclude that one model is better than the other. However if we want to consider the θ value, it is also possible to argue that the logistic regression model (model A) is slightly better than the ANN model (model B).

Given input features X, logistic regression computes a linear combination of these features weighted by coefficients (similar to linear regression), then applies the logistic function. The relevance of features in logistic regression is determined by the magnitude and sign of their coefficients. Larger absolute values indicate a stronger effect on the model's output. To see if the same features are considered relevant as in the regression analysis, we will compare the coefficients' magnitudes and signs.

Table 6. Evaluation of the selected features in Logistic and Linear regression

Feature	**Logistic Regression Coefficient**	**Linear Regression Coefficient**
concave points_mean	2.6313	-0.0195
concavity_mean	-1.2949	-0.0726
concave points_mean*	2.6313	-0.0195
perimeter_mean	-2.7652	3.4879
area_mean	-3.9696	0.1510
compactness_mean	-0.9585	-0.1790
radius_worst	-3.5963	0.8105
perimeter_worst	-0.0930	-0.5602
area_worst	16.9911	-0.2265
concavity_worst	1.5764	0.0263
concave points_worst	0.2766	0.0289
fractal_dimension_mean	-0.2005	0.0185

Magnitude Differences: The logistic regression coefficients show larger magnitudes for several features compared to the linear regression model. This suggests a stronger reliance on those features for classification.

Directionality: For most features, the signs of the coefficients differ between models, indicating a divergence in how each model interprets the feature's effect on the outcome. Notably, perimeter_mean and radius_worst show opposite signs, suggesting contrasting interpretations of these features' relationships with the target variable.

Significant Features:
Logistic Regression: The area_worst feature stands out with a significantly high coefficient of 16.9911, indicating a strong positive association with the probability of the target class being 1 (malignant).

Linear Regression: The perimeter_mean feature shows a notably high coefficient (3.4879), suggesting its significant role in predicting the continuous target.

5. Discussion

5.1.

In evaluating the performance of the regression models, particularly the linear regression, we not only considered the Mean Squared Error (MSE) but also conducted a thorough statistical analysis to ensure the robustness of our recommendations. The confidence intervals (CIs) and p-values associated with the MSE differences between the models offer compelling evidence of the superiority of the linear regression model over the baseline models. Specifically, the significantly lower MSE of the regularied linear regression model was not only statistically significant, as indicated by a p-value well below the conventional threshold of 0.05, but the confidence interval for the difference in MSE. This further reinforces our confidence in the linear regression model as the preferred choice for predictive tasks on this dataset.

While the ANN provided significant predictive improvements over the baseline, it did not reach the performance of the regularized linear regression model. However, the ANN can also be considered as the MSE of the ANN model is only marginally worse than the linear regression model.

Ultimately, given the close performance between the models, the linear regression model's simplicity and ease of interpretation make it a preferred choice for medical decision-making contexts.

Further, it's advisable to extend the model validation beyond MSE as a sole metric, exploring alternative measures of performance to gain a more comprehensive understanding of the models' predictive capabilities.

Based on statistical analysis of the various classification models, while both logistic regression and ANN outperform a naive baseline, their close performance invites a deeper deliberation on model selection criteria beyond mere accuracy. Factors such as interpretability, computational efficiency, and ease of integration into clinical workflows become paramount. The marginal preference for logistic regression, suggested by the θ value in its comparison with the ANN, might tilt the balance in its favour given its straightforward interpretability and lower complexity—a vital aspect in medical decision-making processes.

5.2.

We referenced the Abdulkareem study: “An Evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique” [2]. The referenced study highlighted the effectiveness of ensemble classifiers, specifically Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), in analyzing the Wisconsin Breast Cancer Dataset (WBCD). Their analysis, which applied Recursive Feature Elimination (RFE) for feature selection, demonstrated that XGBoost, with a reduced feature set, achieved the highest accuracy of 99.02%. This finding aligns with our study's emphasis on the efficacy of ensemble methods and feature selection techniques in enhancing model performance.

Both studies underscore the importance of feature selection in improving model accuracy. The cited research employed RFE to identify a subset of features that resulted in optimal model performance. Our study's reliance on optimal feature selection methods reinforces the value of this approach in achieving high predictive accuracy.

While the Abdulkareem study recommended XGBoost as the most effective model based on their analysis, the authors state that MLP and Logistic Regression had better performance in comparison to the other classifiers that were utilized with an accuracy of 97.9%. The decision to recommend logistic regression in our analysis was based on its balance of accuracy and interpretability, particularly valuable in medical applications where understanding model decisions is crucial.

There is an ongoing need for research in machine learning applications for medical diagnosis, particularly in the early detection of breast cancer. The Abdulkareem study suggests further exploration of ensemble classifiers and feature selection techniques, a recommendation that aligns with our call for continued investigation into various machine learning models and methodologies to refine predictive accuracy.

6. Exam questions

6.1. Question 3 - Answer A

There are 7×10=70 weights associated with connections between the input and hidden layers. There are 10 bias terms, so the total parameters between the input and hidden layers are 70+10=80.

There are 10×4=40 weights associated with connections between the hidden and output layers. There are 44 bias terms, so the total parameters between the hidden and output layers are 40+4=44.

Adding the parameters from both layers, we get 80+44=124.

6.2. Question 4 - Answer B

Examining the connection between the output layer and the graph reveals the initial separation occurs at b1=−0.76. Consequently, only options B and D remain as potential answers. In option D, the subsequent separation occurs around b2, preceded by the necessity of separation along b1 as in option C. Therefore, answer B is correct.

6.3. Question 5 - Answer C

k_1 is the number of the outer loops. The time taken in the two processes can be calculated by the formula: $(t_{log.reg} + t_{ANN}) * k_1 * k_2 * \lambda + (t_{log.reg} + t_{ANN}) * k_1$ results in 3570ms.

6.4. Question 6 - Answer B

Class y = 4 is selected only when P(y = 4|$\hat{y}$) > P(y = k|$\hat{y}$) for k ≥ 3. This condition is met when $e^{\hat{y}} < 1$, indicating that $\hat{y}k$ must be less than 0. Given $\hat{y}k$ = [1b]w_k, we assess all combinations of b = (-0.6, -1.6) is the only criterion that is met.

References

[1] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

[2] Abdulkareem, S. A., & Abdulkareem, Z. O. (2021). An Evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique.
https://gssrr.org/index.php/JournalOfBasicAndApplied/article/view/12300