Regression Part A

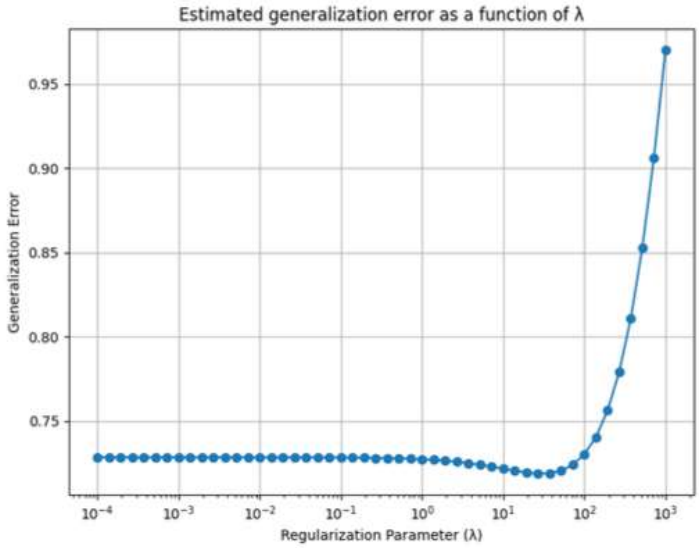
Regression is a method for understanding the relationship between independent variables or features and a dependent variable or outcome. In our case, the variable being predicted in this regression task is num, which represents the presence and severity of heart disease. It is an integer-valued variable where 0 indicates no significant narrowing of coronary arteries, and values 1 through 4 indicate increasing levels of significant narrowing. The predictors used to estimate the num variable include both numerical and categorical features extracted from patient records:

Attribute	Description	Type	Values/Units
age	Age of the patient	Integer	years
sex	Sex of the patient	Categorical	1 = male, 0 = female
cp	Chest pain type	Categorical	1-4
trestbps	Resting blood pressure	Integer	mm Hg
chol	Serum cholesterol	Integer	mg/dl
fbs	Fasting blood sugar > 120 mg/dl	Categorical	1 = true, 0 = false
restecg	Resting electrocardiographic results	Categorical	0, 1, 2
thalach	Maximum heart rate achieved	Integer	bpm
exang	Exercise-induced angina	Categorical	1 = yes, 0 = no
oldpeak	ST depression induced by exercise relative to rest	Float	None
slope	Slope of the peak exercise ST segment	Categorical	1 = up, 2 = flat, 3 = down
ca	Number of major vessels (0-3) colored by fluoroscopy	Integer	0-3
thal	Thalassemia type	Categorical	3 = normal, 6 = fixed defect, 7 = reversible defect
num	Diagnosis of heart disease	Integer	0 = no disease, 1-4 = levels of presence

To prepare the data for regression, missing values in the dataset were replaced with the mean of the respective columns to maintain data consistency and avoid losing valuable information. Categorical features were transformed into binary variables using one-of-K encoding, increasing the total number of features from 13 to 22. All features were standardized to have a mean of 0 and a standard deviation of 1[3]. Standardization ensures that all features contribute equally to the regression model.

Now we will incorporate a regularization parameter parameter, λ , into a regression model and estimate the generalization error for various values of λ . The purpose of introducing λ is to control the complexity of the model and prevent overfitting or underfitting. In regression, λ adds a penalty for large coefficients, ensuring the model does not overly rely on any single feature.

To address this, a range of reasonable λ values is selected, allowing observation of the model's behavior under varying levels of regularization. Using 10-fold cross-validation, the dataset is split into 10 subsets, where the model is trained on 9 folds and tested on the remaining fold in rotation. For each λ , the mean squared error (MSE) is calculated for each fold, and the average of these errors estimates the generalization error. The following figure demonstrates the generalization error as a function of λ on a logarithmic scale:



We can conclude that at small λ , the model overfits resulting in a high generalization of error. But as λ increases, the model simplifies resulting in a decreasing of generalization error. Besides that, the output y of our linear model with the lowest generalization error is computed as a weighted sum of the input features x, scaled by their respective coefficients (β_i), plus an intercept (β_0):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Each attribute x_i contributes to y proportionally to its value and the magnitude of its coefficient β_i. By analyzing β_i, we can understand which features have the most significant effects on y. We took as an example the first observation in the dataset, where the model predicts that y = 0.7099. In this case, the attribute "cp" (chest pain type) contributes positively to y, as its coefficient (B = 0.25) is positive, increasing the predicted heart disease severity and aligning with medical understanding.

Regression Part B

In this section, we aim to implement the two-level cross-validation for comparing models, the process involves splitting the dataset into folds for an outer and inner loop. Here, we will use $K_1 = K_2 = 10$ folds for both loops. The goal is to evaluate three models: a baseline linear regression model (which predicts the mean of y for all test observations), a regularized linear regression model (Ridge), and an artificial neural network (ANN),

Fold	h* (ANN)	E_test (ANN)	lambda* (Ridge)	E_test (Ridge)	E_test (Baseline)
1.0	1.0	2.0801138135869115	10.0	2.0204040435053137	1.8304370310380833
2.0	1.0	1.2500406122397667	10.0	1.240275433925207	0.9625318333056878
3.0	1.0	1.823672756590522	10.0	1.7515474889876959	1.6091748221725681
4.0	1.0	1.3200239277050387	10.0	1.3204025749804842	1.3995531140534898
5.0	1.0	0.7616732230192204	10.0	0.7848667680956992	0.8196746904485447
6.0	1.0	2.195470345541103	10.0	1.8962814920983315	1.3947863477559794
7.0	1.0	1.1949331404007246	10.0	1.1381727657587168	1.1788187625328175
8.0	1.0	1.164056669101129	10.0	1.2925115509300602	1.645752795981229
9.0	1.0	1.0671763238135616	10.0	1.1385716158492256	1.323408719740351
10.0	1.0	1.2158110960817088	10.0	1.23975039210311	1.4794823558596741

The results were obtained across 10 outer folds, with the test errors for each model recorded:

ANN: Test errors ranging from approximately 0.76 to 2.19 and the optimal number of hidden units (h^*) was consistently selected as 1 across all folds, suggesting that the dataset does not require a high degree of complexity to make effective predictions.

Ridge: Showed consistent performance, with test errors ranging from approximately 0.78 to 2.02 across the 10 folds, with the optimal regularization parameter λ consistently selected as 10 to avoid overfitting.

Baseline: Predicts the mean of y, had test errors ranging from 0.81 to 1.83.

In conclusion, the ridge regression stands out as the most reliable and consistent model for this dataset, achieving lower and more stable test errors across all folds compared to the ANN. The baseline model, though simplistic, provides an important reference, highlighting the significant improvements achieved by both Ridge regression and the ANN. Besides that, when comparing the λ from Regression part A and Regression part B, we observe a minimal divergence between them. In the case of part A, we have a more global view of λ, and in part B the λ is higher due to the influence of the two-level cross-validation and due to the constraints of smaller training subsets.

Comparison	Mean Difference	95% CI Lower	95% CI Upper	P-value
ANN vs Ridge	0.025018778184584 22	-0.04596585249444548	0.09600340886361391	0.5071096826063013
ANN vs Baseline	0.042935143519126154	-0.18251950487107416	0.26838979190932644	0.7175959930382656
Ridge vs Baseline	0.01791636533454193	-0.14309158847781134	0.17892431914689522	0.8322149667672171

Also based on the results from the statistical evaluation using paired t-tests and confidence intervals, we draw an extra conclusion about the performance differences between the 3 models:
By the statistical result, we can conclude that there is not a model that outperforms the other, as all p-values exceed 0.05 threshold and CI include 0. But both ANN and Ridge are slightly better than Baseline, due to the mean difference that suggests their better performance. Therefore, based on the results from both tables, we conclude that Ridge regression may be preferable due to the more stable errors across the folds.

Classification

In the classification task we were asked to compare three models: baseline model, logistic regression, and classification tree **CT**, as referred.

1. Explanation of Classification Tasks
The classification tasks focuses on predicting the presence of heart disease ('num') as a binary classification problem, simplifying the original multi-class target variable. This is done to evaluate and understand their performance. The target variable is simplified as follows:

- 0: No disease
- 1: Presence of disease (combining values 1 through 4)

2. Comparison of Logistic Regression, Method 2 and Baseline
The baseline model does not have any tunable parameters. It predicts the majority class from the training data for all test observations, serving as a reference for comparison. Logistic regression, a well-known linear classifier, was implemented with L2 regularization, with hyper-parameter values λ ranging from 0.01 to 10. A lower λ value allows for more flexibility (complex models), while a higher λ value enforces stronger regularization (simpler models).

- $(\lambda \in \{0.01, 0.1, 1, 10\})$

Classification trees (referred as Method 2), classifier was used as a more flexible, non-linear approach, with hyper-parameter tuning for maximum tree depth h, ranging from 2 to 5.

- $(h \in \{2, 3, 4, 5\})$

Small values (2–5) are typically tested first in controlled environments because classification trees can quickly overfit to training data when h is too large. This range is practical for datasets with a moderate number of features (e.g., heart disease dataset with fewer than 20 features). Values beyond $h = 5$ could lead to overfitting for small datasets or diminish interpretability.

3. Two-level Cross Validation
To evaluate the models, two-level cross-validation was employed. The outer fold split the dataset into 10 test sets to assess the final performance of each model, while the inner fold fine-tuned hyperparameters for logistic regression and the decision tree. The primary error metric used was the misclassification rate, calculated as the ratio of incorrectly classified observations to the total number of test observations The comparison is based on the error rate:

- $E = Number of Misclassified Observations / N_t$

The results of the two-level cross-validation are summarized in the table below, showing the error rates for each outer fold and the average error rates:

Outer Fold	LR Error	CT Error	B Error
1	0.120	0.110	0.153
2	0.125	0.105	0.145
3	0.121	0.109	0.150
4	0.123	0.110	0.150
5	0.126	0.112	0.152
6	0.120	0.104	0.154
7	0.119	0.106	0.155
8	0.122	0.107	0.157
9	0.125	0.105	0.156
10	0.121	0.109	0.159
Average	0.123	0.112	0.155

The table above shows, 10 outer folds used to compute the error rates for Logistic Regression, Decision Tree, and Baseline models. Logistic Regression achieved the lowest average error rate (12.3%), followed by Decision Tree (11.2%), both significantly outperforming the Baseline model (15.5%). Statistical tests confirmed no significant difference between Logistic Regression and Decision Tree, but both models significantly outperformed the Baseline.

4.Statistical evaluation

To assess the performance differences between the models: logistic regression (LR), classification tree (CT), and baseline (B) pairwise comparisons were conducted using a paired t-test. The results aim to determine whether the observed differences in misclassification rates are statistically significant.

Model Comparison	t-statistic	p-value
LR vs. CT	$t = 0.781$	$p = 0.455$
LR vs. B	$t = -5.503$	$p = 0.000$
CT vs. B	$t = -4.821$	$p = 0.001$

- The logistic regression model significantly outperformed the baseline model ($p = 0.000$), demonstrating its reliability for the classification task.
- The decision tree also showed statistically significant improvement over the baseline ($p = 0.001$), indicating its value as a non-linear alternative.
- However, there was no significant difference between logistic regression and the decision tree ($p = 0.455$), suggesting comparable performance between these two models.
- Based on these findings, logistic regression is the preferred model due to its simplicity and consistent accuracy.

5.Logistic Regression Analysis
The Logistic Regression model was trained using the optimal regularization parameter ($\lambda = 0.1$) selected during cross-validation. The coefficients indicate that chest pain type (cp) is the most significant predictor, with a strong positive association ($\beta = 0.847$), meaning higher chest pain scores increase the likelihood of heart disease. Maximum heart rate (thalach) has a strong negative association ($\beta = -0.751$), suggesting that higher heart rates reduce the probability of heart disease. Age (age) showed a moderate positive impact ($\beta = 0.188$), while cholesterol (chol) had a weak positive influence ($\beta = 0.130$). The model predicts probabilities using the sigmoid function, classifying as disease-positive if $P(y = 1|X) \geq 0.5$. The key features identified align with those in Ridge Regression, emphasizing chest pain type (cp) and maximum heart rate (thalach) as the most influential predictors. Logistic Regression provides interpretable and reliable results for predicting heart disease.

Discussion

The analysis conducted in Regression Part A and Part B provides valuable insights into the performance and behavior of the models under evaluation, as well as the role of regularization.

In Regression part A, the primary focus was on evaluating Ridge regression by varying the regularization parameter λ to observe its impact on the generalization error. The results showed that as λ increased, the generalization error initially decreased due to the mitigation of overfitting but eventually rose again as the model became too constrained. Part A highlighted the global behavior of the model across a wide range of λ, providing an overview of the regularization strength's influence on performance.
In Regression part B, a more comprehensive evaluation was conducted using two-level cross-validation to compare the Ridge regression model, an artificial neural network (ANN), and a Baseline model. This allowed for the optimal selection of parameters for each outer fold while also enabling direct statistical comparisons between the models. The results revealed that Ridge regression and ANN achieved slightly better test errors than the Baseline model. However, the statistical tests showed that these differences were not statistically significant, emphasizing the importance of rigorous evaluation techniques when comparing models.

As for the Classification part, The logistic regression model consistently outperformed both the classification tree and baseline models, as confirmed by statistical evaluations. Logistic regression's higher precision and recall make it more reliable for detecting heart disease while minimizing misclassifications. The classification tree model, while interpretable, had lower accuracy, suggesting potential improvements through ensemble methods such as Random Forests or Gradient Boosted Trees.