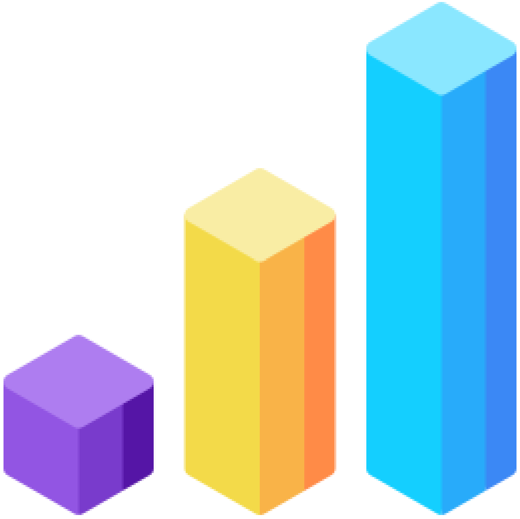


SIT720 Machine Learning

Task 6.1

Michael Rideout
Student Id: 225065259

1.1 Main Points Summary

Week 5 Summary

Week 5 introduced supervised learning, which is the machine learning task whereby models are trained with a known target or label. In mathematical terms it learns the mapping function of input variable(s) to a target variable. Examples of supervised learning algorithms are Decision Trees, Support Vector Machines, Linear Regression and Artificial Neural Networks.

Supervised Learning Process
The hypothesis space is the set of all possible functions that can be learnt by an algorithm. The object of the supervised learning algorithm is to find the within the hypothesis space, the function that best approximates the true underlying function

In order to judge the quality or validity of a function, a loss function provides the means to measure the error between the predicted and actual true target label. Using this, the empirical risk can be determined which is the average of the loss function results.

Model Complexity
It has been observed that the complexity of a model has a related impact on its predictive performance. A model that is too complex may lead to overfitting, i.e. the model learned the training data too well and doesn't perform well on unseen data. Conversely, if the model is too simple, underfitting may occur where the underlying patterns in the data may not be learnt.

Model Evaluation
Model evaluation is the method used to measure the model’s predictive performance. The type of evaluation method used usually depends on the type output datatype of the model (regression vs classification for example).

Classification metrics measure different aspects of the classification performances. A confusion matrix provides a grid summary of the correct and incorrect predictions. Accuracy measures the overall fraction of correct predictions. Recall measures the fraction of actual positive data points correctly identified. False Positive Rate measures the fraction of actual negative data points correctly identified. ROC Curve is a plot of true positive rate vs false positive rate so that imbalances may be discovered. AUC is a single number which summarises the ROC curve performance. The closer to 1 the better. 0.5 is equal to the model choosing randomly. The F1- measure is the harmonic mean of precision and recall, resulting in a score balancing both measures.

Regression metrics measure the accuracy of the model’s predictions. The mean squared error is the squared difference between the actual and predicted value. The root mean squared error is the square root of the mean squared error. Mean absolute error is the average absolute difference R squared is the proportion of the variance in the target variable predictable from the input variables.

Data Splitting

Splitting of a dataset is required in order to perform reliable model evaluation and selection.
The training set is used to train the mode, the validation set is used to tune hyperparameters and make model selections, while the test set is used for final evaluation of the model's performance on unseen data.

Methods utilised for splitting data are: random subsampling by choosing at random rows from the dataset to be in either the training or test sets. Stratified sampling ensures that the splitting of the data takes into account the target classes and ensures appropriate proportions are maintained. Cross validation partitions data in k equal folds then uses each train / test set from each fold for training and testing.

Hyperparameter Optimisation

Hyperparameters are settings that can modify how a ML algorithm performs during training. An example method of hyperparameter optimisation is to use a grid of parameter values, then iterate each combination of values and perform training and then use the validation set to evaluate performance. Internal cross validation is another approach that utilises cross validation and then internal train / validation splits.

Imbalance Data

Imbalanced data is where there is a significant over or under representation of classes in comparison to other classes. Some methods to address this are to use resampling of the data (under or oversampling of the data) or utilise an algorithm that can penalise misclassifications of the minority class.

Week 6 Summary

Week 6 covered the following topics:

Linear Regression

Linear regression is the machine learning algorithm that, given a set of input features, attempts to predict a single output scalar variable. Under the hood the algorithm learns the weights associated with each feature that minimises predictive errors. The measure of error is usually the mean squared loss, ie the average square of the difference between the actual and predicted target value.

Logistic Regression

Similarly to linear regression logistic regression maps a set of input features to a single output variable, however the output variable is a binary categorical (1 or 0). Instead of using mean squared error as the loss function, the maximum likelihood estimated is used.

Model Complexity

Model complexity is a topic that involves measures and mitigating the effects of model complexity (how simple or complex a model is). A machine learning model can fall between two extremes of complexity where the following occurs:

- Overfitting - An excessive number of variables or inappropriate training data produce a model that works well on the training set but performs badly on the test set.
- Underfitting - Happens when a trained model is too simplistic to be able to capture the underlying patterns in the dataset which results in poor evaluation scores.

The expected loss (Risk) is a measure of the fitness of a model. It has two components, bias and variance. Bias measures how accurate a model is with the lower the score the better. Variance is a measure of the model's sensitivity to changes in the data. The higher the variance the more complex a model is. There is a trade off inherent between these two measures, an increase in variance usually leads to a decrease in bias and vice-versa. The object is to find the best balance that produces the minimum risk

Regularisation

Regularisation aims to reduce the risk of overfitting by introducing penalty schemes into the loss function of the machine learning algorithm. Usually weights are added to coefficients which discourage excessively large values. The two most common regularisations are L1 (Lasso - encourages sparsity by performing variable selection) and L2 (Ridge - penalises large weights)
Regularisation generally increases bias and reduces variance.

1.2 Summary of Reading List Items

Week 5 Reading Items

How to Use Occam's Razor Without Getting Cut

This article explores Occam's Razor, the rule of thumb which states that when all things being equal, the simplest solution is preferable. It is motivated in the desire to reduce unnecessary complexity as much as possible. It has been particularly useful when applied to scientific theories as they are then easier to prove or falsify. The article invites caution when applying it as it isnt a rule, but a suggestion of how to approach complexity.

Standard Error of the The Estimate

A video that delves into measuring accuracy of a regression line by comparing predicted values to ground truths. It discusses mean square error and compares it to the R squared measure

Hyperparamter Optimization methods - ML

An article that describes methods for hyperparameter optimisation. The methods it describes are Grid Search (tries every combination of parameter), Random Search (samples random points in the parameter grid), Bayesian Optimisation (uses prior results to build a probabilistic model), Sequential model Based Optimisation (iterative layer on top of bayesian optimisation).

Week 6 Reading Items

Predicting Heart Disease Diagnoses with ML

The blog describes the process of making a SVM model to predict a binary classification of heart disease. It obtained the health data from the UCI Machine learning Repository.

Diagnosing Bias vs Variance

A video that investigates bias and variance issues in ML algorithms. It shows how to identify underfitting or overfitting based on training errors.

Linear regression - Regularisation

This video explains regularisation techniques for linear regression, covering ridge regression (L2) and Lasso (L1)

Semi-supervised Feature Selection via Rescaled Linear Regression

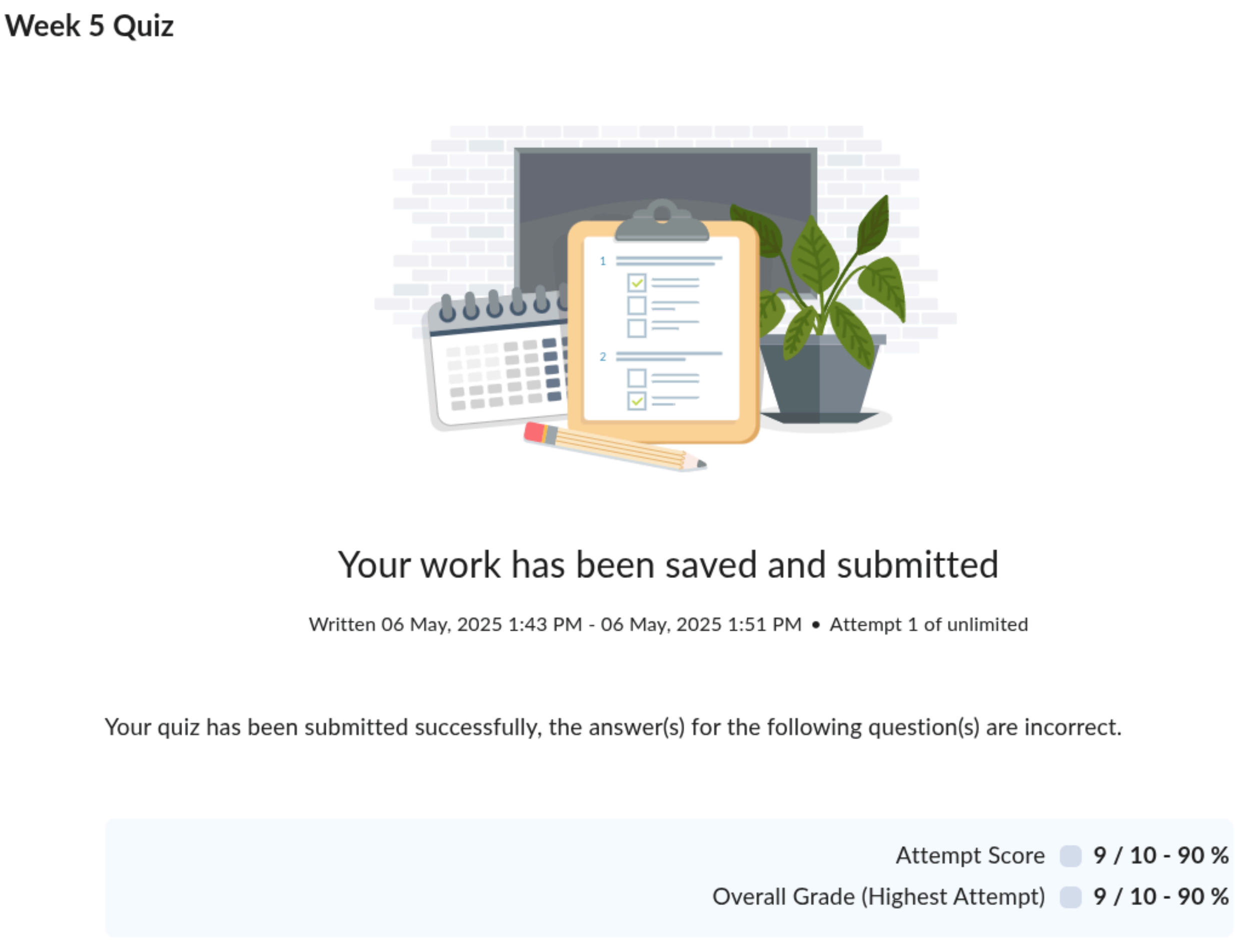
This paper describes a semi-supervised method for feature selection for a highly dimensional sparse dataset. It generates a linear regression model and then rescales coefficients.

1.3 Learning Reflection

In week 5 I learnt about supervised learning, how to evaluate the predictive performance of supervised learning models, data splitting and evaluation metrics. Fundamental concepts such as bias and variance and over and underfitting were described which are important to understand as these items have a huge impact on supervised learning model performance. Week 6 focused on linear regression, logistic regression and regularisation for linear regression models.

1.4 Quiz Results

Week 5 Quiz:



Week 5 Quiz

Your work has been saved and submitted

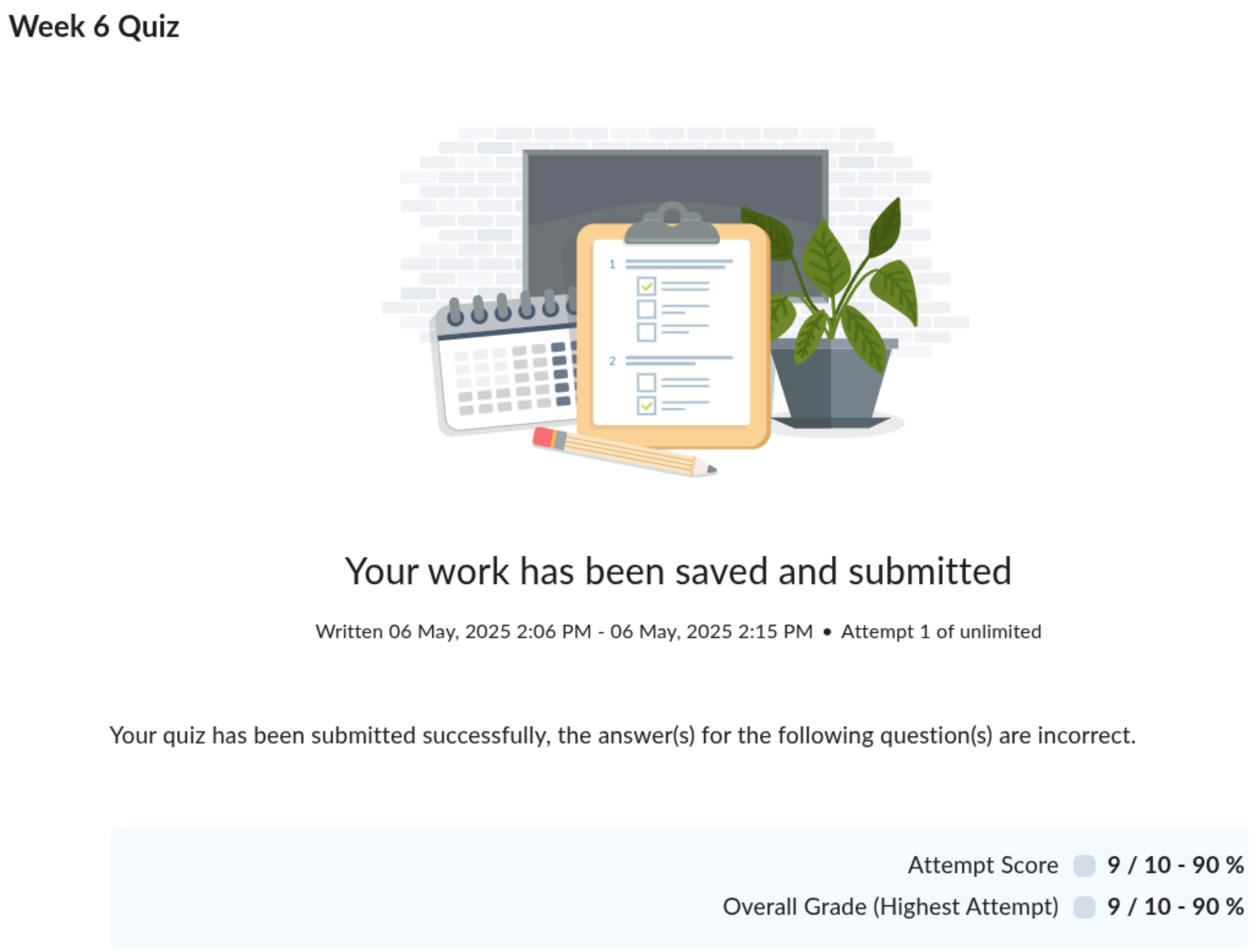
Written 06 May, 2025 1:43 PM - 06 May, 2025 1:51 PM • Attempt 1 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score **9 / 10 - 90 %**

Overall Grade (Highest Attempt) **9 / 10 - 90 %**

Week 6 Quiz:



Week 6 Quiz

Your work has been saved and submitted

Written 06 May, 2025 2:06 PM - 06 May, 2025 2:15 PM • Attempt 1 of unlimited

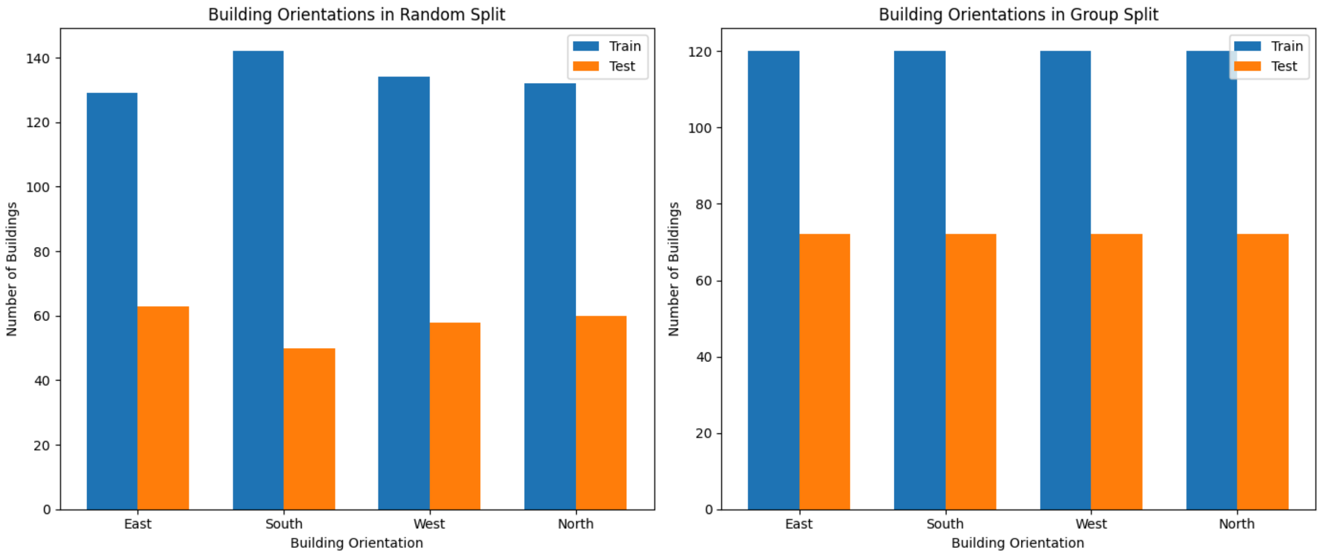
Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score **9 / 10 - 90 %**

Overall Grade (Highest Attempt) **9 / 10 - 90 %**

Jupyter Notebook Graph Outputs

Item2 Bar Chart


Building Orientations in Random Split
Train
Test
Number of Buildings
East
South
West
North
Building Orientation
Building Orientations in Group Split
Train
Test
Number of Buildings
East
South
West
North
Building Orientation