# Exam Answer Sheet

A.D. van der Merwe
Department of Computer Science
University of Stellenbosch
24923273
24923273@sun.ac.za

October 28, 2024

## Section A: Introduction to Machine Learning

### Question 1

Ockham's razor is a principle that states if there are multiple ways to explain a phenomenon, the simplest explanation that still effectively explains the phenomenon should be used.

Regarding predictive models, Ockham's razor is crucial in model selection and development. A simple model that is less computationally expensive or a model with fewer parameters and assumptions should be favoured if the prediction accuracy is sufficient enough for the task at hand.

For example, if a classification decision tree model achieves similar results to a more complex random forest classification model, the simple classification tree is preferred to perform predictions for the specific task. Simpler models are also easier to interpret, are less prone to overfitting on the training data, and are much less computationally expensive.

### Question 2

Variance in a model refers to the extent to which a prediction varies when different test sets are used to construct the model [1]. It is almost impossible to find real-world data where there is no variability in the dataset. Therefore there will always be variability in each training dataset and the prediction of a model will almost always result in different values. Ideally should the prediction of an instance not vary too much between training sets. High variance indicates that the model may be overfitting, meaning it captures noise and specific patterns from the training data and generalises poorly to unseen data. This sensitivity to small changes in the training data can result in significant fluctuations in predictions. To calculate the variance of a regression model, the following equation is used.

$$Var(\hat{y}) = \mathbb{E}\left[(\hat{y} - \mathbb{E}\left[\hat{y}\right])^2\right] \tag{1}$$

where $\hat{y}$ is the predicted value of the regression model and $\mathbb{E}\left[\hat{y}\right]$ is the expected predicted value across different training sets.

Bias in a model refers to the error introduced by approximating a real-world problem, which may be extremely complicated, by a much simpler model. A model with high bias can lead to inaccurate predictions by underfitting the data as a simple model might not be able to capture the complex relationships between the descriptive features and the target feature. A high bias indicates that the model's predictions deviate significantly from the actual values, often due to an overly simplistic model that cannot capture the underlying complexities of the data. To calculate the bias of a regression model, the following equation is used.

$$Bias(\hat{y})^2 = (\mathbb{E}[\hat{y}] - y)^2 \tag{2}$$

where $y$ is the actual value that the model aims to approximate.

The expected test mean squared error (MSE) can be decomosed into the sum of three fundamental quantities which is given in the following equation.

$$\mathbb{E}\left(y - \hat{y}\right)^2 = Var(\hat{y}) + Bias(\hat{y})^2 + \sigma^2 \tag{3}$$

where $\sigma^2$ is the irreducible error, which is the noise in the observations that cannot be reduced by any model.

From equation (3) it is clear that both the variance and the bias of a model has to be small values to minimise the test MSE. The bias-variance dilemma arises from this. The more complex a model becomes to capture relationships between the descriptive features and the target feature, the more susceptible the model becomes to overfitting on the data, therefore the reduction in bias may lead to an increase in variance. Conversely, if a model is kept simple to prevent overfitting, the model may struggle to capture complex relationships between the descriptive features and the target feature, therefore the reduction in variance may lead to an increase in bias.

## Question 3

### (a)

Underfitting occurs when a predictive model is too simplistic to capture the underlying relationships between the descriptive features and the target feature resulting in poor predictive performance for both the training and test datasets.

### (b)

Overfitting occurs when the predictive model is overly complex, leading it to become sensitive to noise and fluctuations in the training dataset. This results in the model fitting too closely to the training data, where the complex underlying relationships between the descriptive features and target features are captured, but the model learns the noise present in the training dataset. This then results in a good predictive performance for the training set, but the generalisation too the unseen test dataset will be poor.

### (c)

When a model is underfitting, it fails to learn underlying relationships between the descriptive features and the target feature, thus resulting in a high bias and a low variance. Conversely, when a model is overfitting, it accurately learns the underlying relationships between the descriptive features and the target feature but also captures noise within the dataset. This leads to the model becoming sensitive to small fluctuations in the training dataset, thus resulting in a high variance and a low bias.

## Question 4

When a dataset contains outliers, it is generally a bad idea to impute missing numerical values by use of the mean as the mean is influenced by the outliers that leads to an inaccurate representation of the central dendency of the data. It is generally better to use the median of the numerical values of a feature when outliers are present in the feature values. The median is less affected by outliers and provides a better measure of central tendency of the feature values if these values are not normally distributed.

Additionally considering more advanced imputation methods that accounts for the relationships between the descriptive features can yield better results. This is especially the case for when the data structure is complex. A k-nearest kneighbours (KNN) classifier technique can be used to identify the k-nearest instances to the observation with a missing value based on the other features similarity. The missing value is then imputed using the mean or median of the feature values from the k-nearest instances, which often results in a more robust and representative imputed value.

When a classification task is considered, it is generally better to impute the missing value of a feature with only the mean or median features reletad to the class which contains the missing value. This approach leverages the conditional relationships between the descriptive features and the target feature, which could lead to more accurate imputations.

## Question 5

Feature selection is a crucial step when preprocessing data to construct a predictive machine learning. Reducing the number of features in the dataset will improve the performance of the model, reduce overfitting, result in faster training times, create a simpler model which is more interpretable and improve the data quality.

The model's performance improves, because only the most relevant descriptive features are chosen to perform predictions of the target feature, which often leads to improved accuracy and predictive power. Additionally, overfitting is reduced as the reduction in dimensionality of the original dataset leads to models that are less complex and reduces the chances of capturing noise in the dataset and increases the chance of the model learning the underlying relationships between the remaining descriptive features and the target feature. The dimensionality reduction also leads to faster training times as the model is constructed on less data, which is very beneficial when large datasets are used and the models are easier to interpret, which is beneficial when a model is used where the decision-making process is important to understand. Lastly, the quality of the data is improved as features with noise or irrelevant features are removed.

## Question 6

### (a)

Stationary data is data that has statistical properties which stays consistent over time, such as the mean and variance. This means that the behavior of the dataset does not significantly change over time and stays more or less constent. Conversely non-stationary data has statistical properties which changes over time. Non-stationary data may show increasing or decreasing means and changing variances, due to trends or even seasonal effects, making it less predictable.

### (b)

A real world problem that exhibits non-stationary data is stock market data. This is because stock market data exhibit upward and downward trends over time, which is influenced by different factors such as economic growth and some stocks may exhibit seasonal patterns, such as retail stocks that perform better during the holiday.

### (c)

The main implications are that the models can become unstable over time, the model needs to consantly learn and the models might overfit past historical patterns.

The first implications is that the model may become unstable overtime and lead to invalid predictionsm, as the relationships between the descriptive features and target feature may change over time. This makes previously learned patterns unreliable and justifies the need for the second implication, which is that the model needs to consantly be updated. The model has to incorporate some form of continuous learning or regular retraining to ensure that the models remain relevant and can adapt as soon as the underlying patterns between the descriptive features and the target feature changes over time. The models may also overfit to past historical, which leads to poor generalisation. This may lead to high error rates when the model is applied to current or future data.

## Question 7

When binning is used to convert a numeric feature into a categorical feature, what are the consequences of too few bins versus too many bins?

The consequences of using too few bins is that important information may be lost with respect to the distribution of values in the original continuous feature. This loss of information may lead to reduced predictive power of the model, as the small number of bins reduces the feature's ability to differentiate between classes or predict outcomes accurately. A small number of bins can also potentially hide variations or patterns that could be useful for modeling.

Conversely, the consequences of using too many bins, is that each bin will have a small number of instances contained within them, where some bins may end up having no instances. Having too many bins may lead to a more computationally expensive model. Additionally, too many bins may increase the complexity of the model, such as a classification tree that has to evaluate a greater number of categories.

## Question 8

### (a)

TODO

### (b)

The distribution of ID (d) indicates a feature that should be removed, as the feature contains unique values. This feature has no predictive information relevant to the target feature and therefore it should be removed.

### (c)

TODO

## Section B: Information-based Learning

### Question 1

### Question 2

**(a)**

**(b)**

### Question 3

### Question 4

### Question 5

### Question 6

**(a)**

**(b)**

### Question 7

### Question 8

## Section C: Similarity-based Learning

## Section D: Error-based Learning

## Section E: Unsupervised Learning

## Section F: Kernel-based Learning

## Section G: Ensemble Learning

## Section H: Reinforcement Learning

## References

[1] G. James, D. Witten, T. Hastie, R. Tibshirani *et al.*, *An introduction to statistical learning*. Springer, 2013, vol. 112.