# University of Canberra

# 11482 - Pattern Recognition and Machine Learning

# Tutorial Class Thursday 0930 (Semester 2 / 2024)

# Assignment Stage 1 Part A

## Max 2 pages (excluding title, contents and references)

## Group 101 (Individual)

## James McGuinness

By submitting this assignment, I declare that this assignment is solely my own work, except where due acknowledgements are made. I acknowledge that the assessor of this assignment may provide a copy of this assignment to another member of the University, and/or to a plagiarism checking service whilst assessing this assignment. I have read and understood the University policies in respect of Student Academic Honesty.

James McGuinness (u3196600)

# Part A1

**1. Provide a description of the problem, motivation and characterisation.**

I’ve initially chose the Breast cancer Wisconsin dataset available through SciKit Learn. I believe this dataset will demonstrate key concepts to me such as classification, confusion matrices and metrics like false positives and false negative.

First and foremost, this is a supervised learning problem as I will be providing the samples and labels to a model where the task is to classify samples as either benign or malignant based on features.

**2. Identify and describe the dataset.**

Dataset description below:

|  |  |
| --- | --- |
| Classes | 2 |
| Samples | 569 |
| Features per sample | 30 |

Features of a sample likely display patterns that correlate with the presence or absence of malignancy. The dataset contains breast scans with various measurements of tissue properties. Each sample is labelled as either benign or malignant.

I planned to include a high-level visualization of the dataset but removed it due to the 2 page limit. I will explore this further in Stage 2.

# Part A2

**1. Identify questions to be investigated.**

Initially I plan to:

1. Identify the independent and dependant variables
2. Use regression to identify a relationship between the independent and dependant variables
3. Investigate features that help in model accuracy and prediction

I believe further questions will become evident throughout the semester while working on this assignment as my knowledge improves about the dataset and pattern recognition techniques.

**2. How would this model be used in decision making in the problem domain?**

In the real world I believe a pattern recognition model could assist radiologists or any other professional in predicting outcomes and assisting in decision making. From my initial readings, the model could in theory highlight with a high probability of becoming malignant and reduce the chance of misdiagnosis.

For each new patient, the model would take the scan measurements as input and

predict whether the tumour is likely to be malignant or benign.

**3. How would you use the model to draw outcomes for new cases?**

At a high level, the model is first trained and then it’s validated using a separate test dataset that the model has not seen during training. This will show the models ability to generalise.

We then have the options to either create new samples manually or investigate another dataset with similar features and samples.

# Part A3

**1. Explain why it’s a PRML problem, with reference to design steps to PRML solution.**

The task is to classify breast scans into malignant or benign categories. This involves recognizing patterns in the scan features that correlate with the malignancy of tumours.

Relevant features are extracted from the scan data, such as mean radius, texture, and perimeter. These features serve as inputs for the machine learning algorithms.

*Design Pipeline:*

Data Collection: Obtain and preprocess breast scan data.

Feature Extraction: Identify and extract relevant features from the scans.

Model Selection: Choose appropriate machine learning models.

Training: Train the models using labelled data.

**2. Discuss what PR will be involved and how ML will be used in the design.**

The dataset is split into training and testing sets. A model is trained on the training set to learn the patterns and is evaluated on the test set to measure its performance.

I believe Principle Component Analysis should assist with visualization of the data.

**3. Identify 3-4 algorithms to investigate for the problem, explain why are they suited?**

*Homework about the general idea of different models is expected to archive high marks*

The four algorithms I have chosen to investigate are; Logistic and Ridge Regression, potentially K-Nearest Neighbour, Decision Trees and potentially Neural Networks to see what patterns can be determined in a non-linear realm.

These algorithms are chosen because they handle classification problems effectively and can accommodate both linear and non-linear relationships in the data. They also offer various ways to deal with potential overfitting and model complexity.

The relationship between features and the outcome may be linear or non-linear.

Depending on the linearity of the data, parametric models (like Logistic Regression) or non-parametric models (like Deep Neural Networks) may be used.

In reality, the choice of model may depend on the computational resources available and the complexity of the data. However, I don’t believe this is a consideration for this assignment.

**4. Explain why the proposed project qualifies as a RPML problem?**

# References

Müller, A.C. and Guido, S. (2017), *Introduction to Machine Learning with Python, O'Reilly Media, Inc., Sebastopol, CA*. Available at: <http://safaribooksonline.com>

Scikit-Learn (2024), *sklearn.datasets.load\_breast\_cancer, Scikit-Learn*. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html>

Yasserh (2024*), Breast Cancer Dataset, Kaggle.* Available at: <https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset>

Devasena K and Shana J (2021) *Building Machine Learning Model for Predicting Breast Cancer Using Different Regression Techniques*. Available at:

<https://doi.org/10.1088/1757-899X/1166/1/012029>

**End of document**