# University of Canberra

# 11482 - Pattern Recognition and Machine Learning

# Tutorial Class Thursday 0930 (Semester 2 / 2024)

# Assignment Stage 1 Part A

# 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.

2. Identify and describe the dataset.

The problem is to develop a model that identifies malignant tumours in breast scans based on the provided dataset.

I chose the cancer dataset because it demonstrates key concepts like classification, confusion matrices and evaluation metrics like false positives (FP) and false negatives (FN).

This is a supervised learning problem where the task is to classify scans as either benign or malignant based on features extracted from the data.

Features of a scan likely display patterns that correlate with the presence or absence of malignancy, making this a typical pattern recognition problem.

The dataset contains breast scans with various measurements of tissue properties. Each sample is labelled as either benign or malignant.

The dataset includes features such as mean radius, mean texture and mean perimeter. These features are crucial for the model to identify patterns that correlate with malignancy.

<Insert plot showing a linear relationship?>

# Part A2

1. Identify questions to be investigated.

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

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

Initially I plan to investigate the features that have the largest impact on the model’s performance. Then investigate which features can help in refining the model and improving accuracy. I do believe further questions will become evident throughout the semester while working on this assignment.

In the real world I believe a ML model could assist radiologists or any other medical professional. The ability to highlight cases with a high probability 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.

The model will be validated using a separate test dataset that the model has not seen during

training. Common strategies include cross-validation or a train-test split, ensuring the model’s

performance is robust and not overfitted.

By applying the model to new cases, we can assess its ability to generalize. A high performance on the test data would indicate that the model can reliably predict outcomes for new patients.

# Part A3

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

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

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

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

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.

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.

**Can I put PCA anywhere?**

**Some homework about the general idea of different models is expected to archive high marks – might be hard in just two pages**

**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.

The pattern recognition involved focuses on extracting meaningful patterns from the data to make accurate classifications.

This problem is a supervised learning task, where the model learns from labelled data, for example scans with known malignancy status; to make predictions on new, unseen data.

The four algorithms I have chosen are; Logistic Regression, K-Nearest Neighbour, Decision Tree and potentially Neural Networks.

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. Visualization and statistical analysis are used to determine the nature of these relationships.

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.

# 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 [Accessed 23 August 2024].

Yasserh (2024), Breast Cancer Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset [Accessed 23 August 2024].

**End of document**