INTERNATIONAL MEDICAL UNIVERSITY
MALAYSIA

HIA302: Health Data Collection and Preparation

Workshop – Part 03

2 hours

Learning Outcomes:

1. Able to demonstrate the concepts of data collection and preparation in a healthcare environment
2. Able to deal with data collection, storage, and processing of data.

Brief description of the project:

A tumour is an abnormal lump or growth of cells. **Tumours can be benign (noncancerous) or malignant (cancerous)**. Benign tumours tend to grow slowly and do not spread. Malignant tumours can grow rapidly, invade and destroy nearby normal tissues, and spread throughout the body. Specific types of **benign tumours can turn into malignant tumours**.

```
Features are computed from a digitized image of a fine needle
        aspirate (FNA) of a breast mass.  They describe
        characteristics of the cell nuclei present in the image.
        A few of the images can be found at
        http://www.cs.wisc.edu/~street/images/

        Separating plane described above was obtained using
        Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree
        Construction Via Linear Programming." Proceedings of the 4th
        Midwest Artificial Intelligence and Cognitive Science Society,
        pp. 97-101, 1992], a classification method which uses linear
        programming to construct a decision tree.  Relevant features
        were selected using an exhaustive search in the space of 1-4
        features and 1-3 separating planes.

        The actual linear program used to obtain the separating plane
        in the 3-dimensional space is that described in:
        [K. P. Bennett and O. L. Mangasarian: "Robust Linear
        Programming Discrimination of Two Linearly Inseparable Sets",
        Optimization Methods and Software 1, 1992, 23-34].
```

Instructions:

In this workshop, we going to go through the data exploration and prepare the data for machine learning.

1. Download the sample dataset from eLearn Portal.
2. Ensure you have Google Colab to work on this dataset.
3. The instructor will guide you the steps during the workshop.

Dataset

This is a breast cancer dataset with number of instances of 569 data.

There are 32 attributes in total, which includes (ID, diagnosis and 30 real-valued input features)

In details:

Attribute Information as follows:

1. ID number
2. Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of grayscale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (“coastline approximation” — 1)

Why there are a total of 32 fields in this file:

The mean, standard error and worst or largest (mean of the three largest values) of *these features were computed for each image*, resulting in **30 features**.

For instance,

- field/column 3 is Mean Radius,
- field 13 is Radius Standard Error,
- field 23 is Worst Radius.

Exercises:

1. Visit the following Data Preparation Crash Course lesson.

   https://machinelearningmastery.com/data-preparation-for-machine-learning-7-day-mini-course/

2. Read through the non-medical dataset data exploration with Python and steps that heavily relaying on the descriptive statistics checking.

   https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

3. You may visit the original dataset. It is extracted from UCI Machine Learning Repository.

   https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29

Helpful information for your project
https://machinelearningmastery.com/what-is-data-preparation-in-machine-learning/