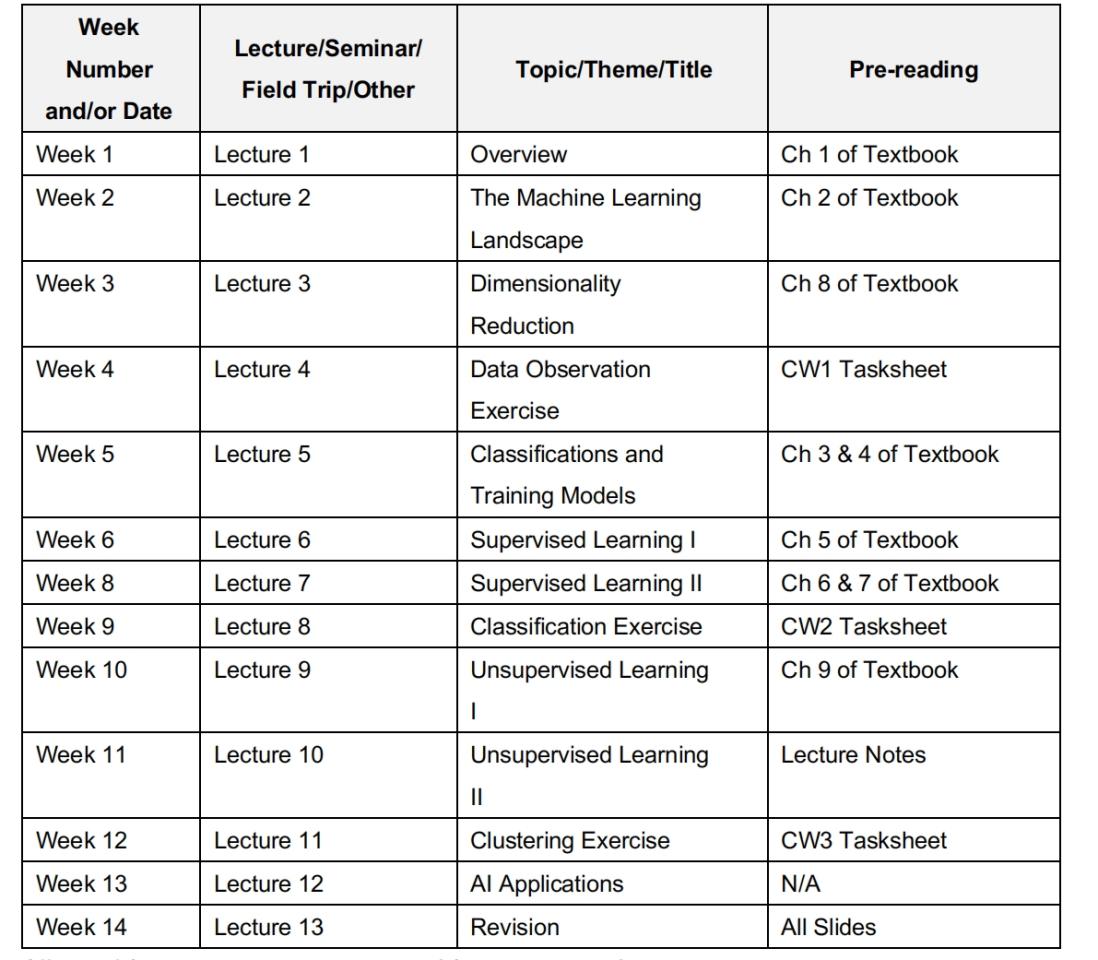
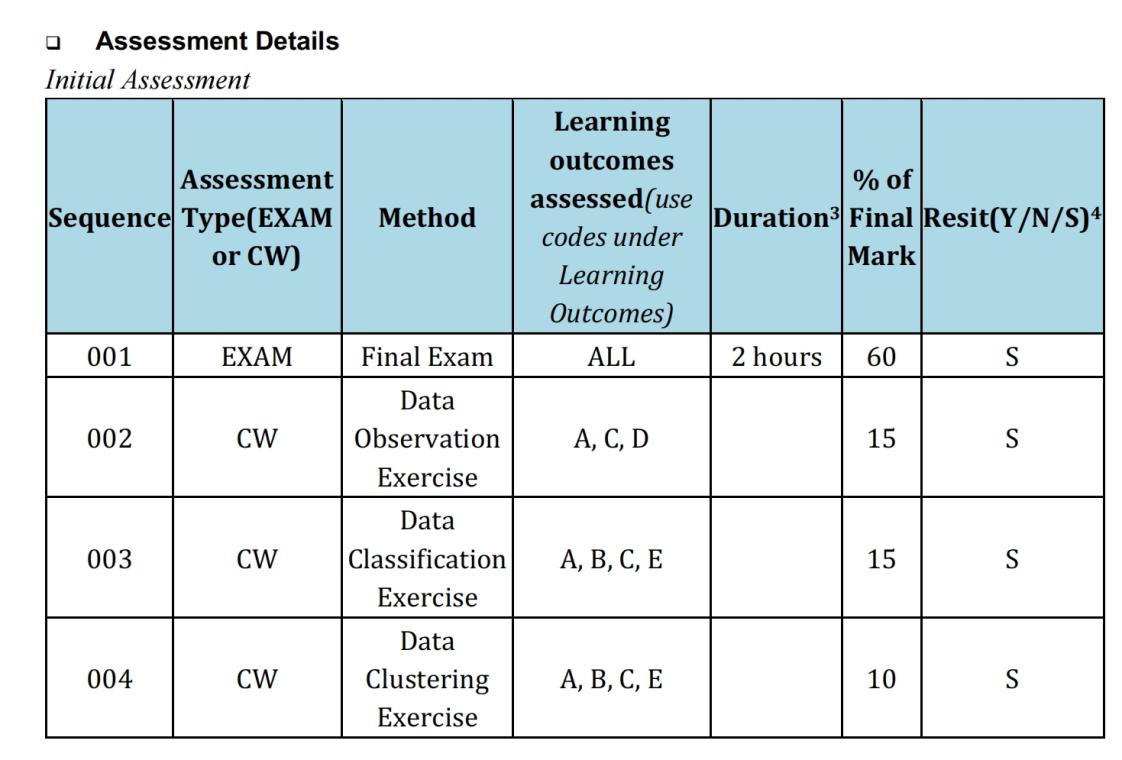
***INT104: Introduction to Artificial Intelligence***

*1 Introduction*

*1.1 Teaching Plan*

**

*1.2 Assessment*

**

*About Final exam*

*----Open book exam (2 hours)*

*---- MCQ: 54 Marks*

*---- Filling blanks: 24 Marks*

*---- Computation: 14 Marks*

*---- Essay Question: 8 Marks*

*1.3 Machine Learning : If a computer program improves its performance on task T measured by score function P , then we say its has learned some experience E.*

*1.3.1 Types of ML*

*Supervised Learning -- label available in training*

*Unsupervised Learning -- label unavailable in training*

*Semi-supervised Learning -- label partially available in training*

*Reinforcement Learning -- Use (s,a,s’,r ) from Bellman to maximize*

*1.3.2 How to select a model with good performance on task T ?*

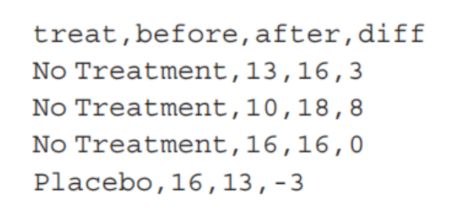
*Train: Training Datasets /Testing Datasets*

*Score: Validation Datasets----Training datasets should not overlap with validation datasets*

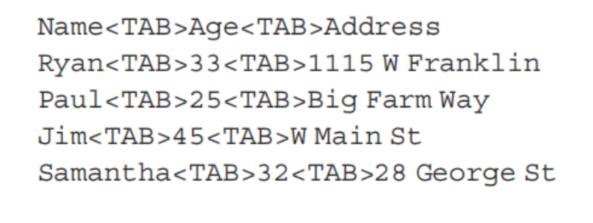
*2 Data preprocessing*

*2.1 Data Storage and Presentation*

*2.1.1 CSV (Comma Separated Values)* 逗号作为字段之间的分隔符

**

*2.1.2 TSV (Tab Separated Values)* Tab作为字段之间的分隔符

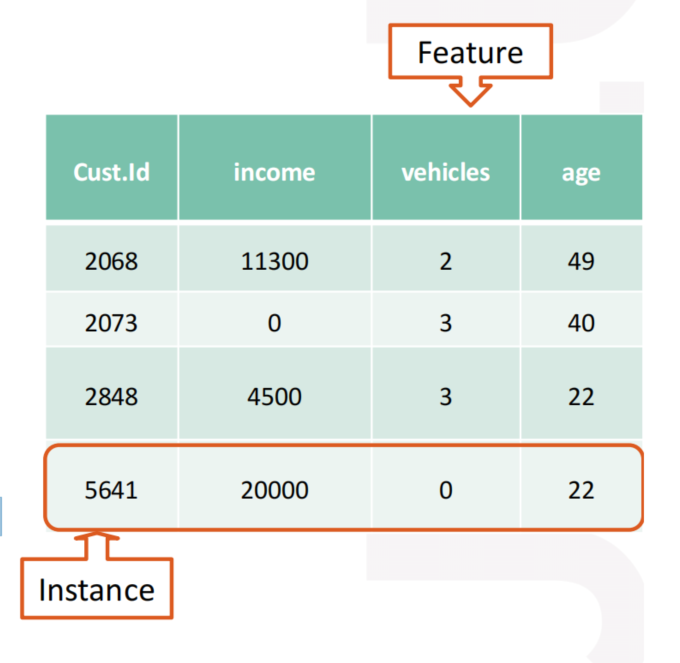
**

*2.1.3 XML (Extensible Markup Language)* 使用标签来标识数据，并且具有层级结构，可以表示复杂的数据关系。XML 可以通过定义自定义标签来适应不同的数据结构，但相对于 JSON，在表示数据时更为冗长。

*2.1.4 JSON (JavaScript Object Notation):* 使用键值对的方式来存储数据，并支持数组和嵌套对象。

*2.2 Data Preprocessing*

*Datasets: A dataset is a collection of instances, features, and target variables that are used to train and test machine learning models.*

**

*The following is the data preprocessing step in order.*

*2.2.1 Data Cleaning and Data Integration*

*• remove the corresponding instance.*

*• remove the whole column.*

*• set missing values to some value (zero, the mean, the median, etc.).*

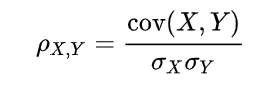
*• remove the outliers* 异常值

*• solve the consistency in data structure and format*

*2.2.2 Feature Selection*

*--- Filter Methods: features are selected and ranked according to their correlations with the target .*

*Pearson’s r correlation: ( no direct relationship with information entropy )*

**

*---- Wrapper Methods: use experiments to search for well-performing combinations of features*

*---- Embedded Methods: do feature selection when the model is training*

*2.2.3 Feature Extraction*

文本数据：对文本数据进行词袋模型、TF-IDF（词频-逆文档频率）计算等方式进行特征提取，也可以利用词嵌入（如Word2Vec等）将文本转换为密集向量表示。

图像数据：使用卷积神经网络（CNN）提取图像的特征，通常通过预训练的模型（如VGG、ResNet等）获取图像的特征向量。

时序数据：对时序数据进行滑动窗口统计、傅里叶变换等方式提取时序特征。

音频数据：对音频数据进行短时傅里叶变换（STFT）、梅尔频率倒谱系数（MFCC）提取声学特征。

*2.2.4 Data Transformation*

*• Handling text and categorical attributes Eg: [‘cat1’], [‘cat2’], [‘cat3’], [‘cat4’]*

*---- Ordinal encoder : [0], [1], [2], [3]*

*---- One-hot encoder : [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]* 消除了类别之间的大小顺序关系，每个类别都被表示为一个独立的维度，避免模型错误地认为类别之间存在顺序关系

*• Normalization*

*---- Min/max normalization*

*x’ =*

*---- Z-score normalization (Normalizing every value in a dataset such that the mean of*

*all of the new values is 0 and the standard deviation is 1 )*

*x’ =*

*---- Normalization by decimal scaling*

*X’ =*

*2.2.5 Data Reduction*