Week 11: Recap on all Concepts

Over the past 10:11 weeks, you covered introductory requirements for Data mining. These are summarized as follows:

Week 1: Acquiring Datasets, and the various databases and data storage formats. We worked mainly with SQLite and xls files. We only experimented with JSON, XML, CSV, and nary XPT in the third week. However, there is no end to how many different database management systems, unstructured data sources, infinite file formats and libraries to read and interact with them. We also introduced Python as one of the most resourceful programming language for data mining, and Pandas library and its famous DataFrame data structures.

Week 2: The main focus was data munging, but we continued data acquiring by introducing big data platforms such as PySpark and the main different data structures "RDD" and how it can be used with DataFrames and Pandas. We resumed Data Munging by further exploring the Pandas DataFrame indexing functions, querying, filling missing values, merging, grouping, scaling (ratio scaling, interval scaling, ordinal scaling, and nominal scaling), creating pivot tables, working with dates and times.

Week 3: Data Visualization week was focusing on exploratory data analysis. We experimented with correlation analysis, histograms, means and variance, outliers, data linearity and non-linearity. We used mainly python matplot library, and some students suggested that seaborn python library offer better visualisations such as the pairplot. The lab material showed the significance of:

- height and width aspect ratio in the interpretation of the visualization,
- elimination Heavy grids,
- removing the background colour to increase the data-ink ratio,
- removing the bounding box,
- use of colour,
- use of charting styles based on objective of the visualisation,
- showing uncertainty,
- showing distributions and clusters,
- partitioning on categories,
- multiple plots and stacking,
- and using logs for ratios and power scales.

Week 4: Introduction to probability week emphasized Bayesian vs Frequentists approaches using probability distributions vs histograms. We experimented with Spearman and Pearson correlation and showed data examples in which one of them would be more meaningful than the other. Ratios are better presented using logs.

Week 5: MLE and MAP calculations and various distributions types week was focused on the expected value and how to calculate it from the various distributions, and how to predict it using MLE the frequentist way, or MAP the Bayesian way.

Week 7: Hypothesis testing week used scipy stats python library to calculate the t-test and the p value to prove or negate a hypothesis, and in example 4 we calculated them from scratch. Four examples were presented on how to represent the hypothesis and its null, and when to accept it or reject it.

Week 8: Model Building week emphasized the importance of accuracy measures and error distributions. We showed how to build a confusion matrix, to calculate the precision, recall accuracy, and the f score from scratch. Then using sklearn python library, we used the ROC curve metric to select the classifier threshold to trade-off precision and recall. Fitting the dataset using various estimators and checking the goodness of fit is explored using SVC and grid search. Also cross validation generators in sklearn python library are used. Linear models and non-linear models are illustrated in the chosen examples. Links to hierarchical models examples are given.

Week 9: Introduction to Machine Learning week was focused on regression and linear algebra. Some presented ideas include:

- Dimensionality Reduction techniques such as PCA and SVD are introduced with some linear algebra matrix operations such as covariance matrices, eigenvalues and eigenvectors.
- Data Normalisation was emphasized in high dimensional datasets.
- The kernel trick using dot product to get a similarity measure between two vectors to avoid calculating the complete distance matrix.
- Regression and SVM are supervised ML approaches that produce a decision boundary to classify data points from a labelled dataset.
- Non-Linear Regression can be performed by fitting the dataset to higher order curved by selecting which non-linear terms to add to the model such as: √x, lg(x), x^2, x^3, and 1/x and estimate each term weight and the bias.
- Logistic Regression uses the logit function to model class and work as classification (discrete classification prediction rather than continuous value prediction). This can be extended to multi-class easily.

Week 10 and Week 11: Advanced Machine Learning Topics weeks focused on presenting introductory material to what you will cover in COMM055 next semester. Distance Measures in the various norms are presented and the high dimensional space effects on the selection of the norm is emphasized. Classification using Nearest Neighbour, and clustering using K-Means are used as basic methods for supervised (availability of labelled data) vs unsupervised (no labelled data, but distance between data points decide if they belong to the same cluster or not). Other methods are introduced such as decision trees, Naïve Bayes, SVM, Boosting, Graphical Models and deep learning. Each of these need to be studied in detail and its suitability to various problems. A subjective comparison of these models is presented by Steven Skiena in his book

"The Data Science Design Manual" comparing them on the basis of power of expression, ease of interpretation and use, training and prediction speed. A number of concepts are presented with each of these approaches that only practice would emphasize their understanding.

As you might appreciate by now, data science is about finding evidence in dataset to support a claim or estimate a prediction. Plethora of tools, programming environments and languages, modelling approaches are existing already and new is introduced every day. To continue excelling you need practice first and foremost, literature review and good resources. The public domain is full of resources to excel in data science. 45 techniques are mentioned in the following article that include the ones we discussed in this module and more to explore on your own when required:

https://www.datasciencecentral.com/profiles/blogs/40-techniques-used-by-data-scientists

Good Luck with your data science projects,

The COMM054 teaching team.