Mini-Workshop

Recent Trends in Feature Selection

23-24 July 2018
Bangor University, UK
School of Computer Science

PROGRAMME:

23rd July 2018

11:00 Welcome & coffee

11:15 L. Kuncheva: **Feature Selection (Our meandering journey thus far)**
The task of feature selection will be briefly introduced in the light of wide datasets. The problems of biased estimates and unstable feature sets will be explained and illustrated. We will share the insights which we gleaned from the literature and our own experiments, and will put on the table the questions which are still seeking to answer. The most important of these questions is whether we should be selecting features at all from a given wide dataset.

12:00 G. Brown: **On the Stability of Feature Selection Algorithms**
Feature Selection is central to modern data science, from exploratory data analysis to predictive model-building. The “stability” of a feature selection algorithm refers to the robustness of its feature preferences, with respect to small changes in the training data. An algorithm is “unstable” if a small change in data leads to large changes in the chosen feature subset. Whilst the idea is simple, quantifying this has proven more challenging - we note numerous proposals in the literature, each with different motivation and justification. We present a rigorous statistical framework for this issue. The conclusion suggests a new measure satisfying a number of useful properties, including (for the first time in the literature) hypothesis testing and confidence intervals on stability estimates.

13:00 Lunch

13:30 K. Sechidis: **Distinguishing prognostic and predictive biomarkers: an information theoretic approach**
The identification of biomarkers to support decision-making is central to personalized medicine, in both clinical and research scenarios. The challenge can be seen in two halves: identifying predictive markers, which guide the development/use of tailored therapies; and identifying prognostic markers, which guide other aspects of care and clinical trial planning, i.e. prognostic markers can be considered as

covariates for stratification. Mistakenly assuming a biomarker to be predictive, when it is in fact largely prognostic (and vice-versa) is highly undesirable, and can result in financial, ethical and personal consequences. We present a framework for data-driven ranking of biomarkers on their prognostic/predictive strength, using a novel information theoretic method. This approach provides a natural algebra to discuss and quantify the individual predictive and prognostic strength, in a self-consistent mathematical framework.

14:15 **Discussion**

15:30 Coffee

24th July 2018

10:30 Coffee

10:45 C. Matthews: **Error estimator bias for two-class, two-feature classifiers**
For the simple case of a binary classifier on a two-dimensional feature space, we investigate some theoretical properties of the error estimator bias. Specifically, we aim to show that the bias is always positive and to demonstrate the effect of correlation between features on the bias.

11:30 J. Rodriguez: **Feature selection in multilabel problems**

12:15 **Discussion**

13:15 Lunch out

Sponsored by Project RPG-2015-188:

Prototype Selection from Streaming, Drifting and Partly Labelled Data Using Classifier Ensembles

The Leverhulme Trust

Contact:
Ludmila Kuncheva masooa@bangor.ac.uk