CS342 Machine Learning: Lab #4
Naive Bayes and Logistic Regression

Labs on February 4 & 5, 2016

Week 4 of Term 2

Office Hours:
CS 3.07, Monday & Friday 10:00-11:00

Instructor: **Dr Theo Damoulas** (T.Damoulas@warwick.ac.uk)
Tutors: **Helen McKay** (H.McKay@warwick.ac.uk), **Shan Lin** (Shan.Lin@warwick.ac.uk)

In the fourth Lab we will explore the use and implementation of Naive Bayes and Logistic Regression together with the use of ROC curves and the AUC metric (end of Lecture 6). None of these models/implementations are really fully Bayesian so if you are interested in such models check out BayesPy, pyMC, and Stan (`http://mc-stan.org`). Refer to the module slides and the Rogers & Girolami book for supporting material (module website) with respect to Naive Bayes and LogReg.

1 Classification with Naive Bayes and Logistic Regression

***Note**: Despite its name, Logistic Regression is a classifier.

The material here builds primarily on lectures 10-12. We will be providing the Python (unoptimised) "solutions" a week after each Lab. We are here to assist with your learning experience so if you need help ask us.

Download the breast-cancer-wisconsin (**Not** the wdbc or wpbc) dataset from `http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`. Our goal is to classify images of cell nuclei from tumors as malignant or benign (binary classification). Features from the images have been pre-computed and make up our attributes. See details on these and the history of this dataset at `http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names`

→ As usual import your data into pandas data frame(s) and do any necessary pre-processing.

→ Fit a Naive Bayes model with a Gaussian likelihood to a 60% random subset of your data.

→ Predict on the remaining 40% of your data and get the probabilities for class membership for each observation (use `predict_proba` function).

→ Construct and plot the ROC curve and report the AUC metric. You can use built in functions such as `sklearn.metrics.roc_curve` and `sklearn.metrics.roc_auc_score`.

→ Repeat the analysis on the same subsets with Logistic Regression from `sklearn.linear_model.LogisticRegression` with an L_2 penalty and default values for rest of the options.

→ For both models report what threshold you would use on the output probabilities to classify an observation as malignant if you want to have less then 5% FP rate (where "positive" is a prediction for malignant tumor). What do you observe?