CS342 Machine Learning: Assignment #1
Plants vs Animals

Deadline: February 12, 2018 at 12:00pm

Week 3 of Term 2

Instructor:
Dr Theo Damoulas (T.Damoulas@warwick.ac.uk) Tutors:

Helen McKay (H.McKay@warwick.ac.uk), **Joe Meagher** (J.Meagher@warwick.ac.uk)
Karla Monterrubio-Gomez (K.Monterrubio-Gomez@warwick.ac.uk),
Jevgenij Gamper (J.Gamper@warwick.ac.uk)

UK SPACE AGENCY

You are a data scientist working for the United Kingdom Space Agency (UKSA) and you have been summoned to analyse some data from the secret space mission "Nereus". Two UKSA space probes have recently arrived at the planet Nereus to collect data on the existing extraterrestrial life-forms living under water.

The first probe (probe A) collected data on 1000 life-forms measuring 4 chemical compounds for each life-form {*cryptonine, mermaidine, posidine, neraidine*} each at 3 different chemical resolutions plus a further genetic attribute called *TNA*. The second probe (probe B) unfortunately malfunctioned during data transmission but before doing so we received a further dataset on 1000 life-forms **without** the TNA measurements.

UKSA brought in biology researchers from the University of Warwick and classified all the life-forms from the probe A dataset into plants (class 0) and animals (class 1). You are now being asked by UKSA to analyse the data and perform the following tasks:

Task 1: Submit[1] a python script *predictClass.py* that reads the original csv files *probeA.csv*, and *probeB.csv*, and outputs another csv file called *classB.csv* with your class predictions (probabilities for class 1) for the probeB data. [5 Marks]

Task 2: Submit a python script *predictTNA.py* that reads the original csv files *probeA.csv*, and *probeB.csv*, and outputs another csv file called *tnaB.csv* with your TNA predictions for the probeB data. [5 Marks]

Task 3: Which attributes are most predictive for whether the life-form is a plant or an animal? Can you comment on the nature of the relationship between inputs and target variables? What steps did you take to explore these relationships? Submit your answer as a pdf file named *answer.pdf*, which should support and present concisely your reasoning and justification. [5 Marks]

Rules of the game & Hints

- The final marks will be a function of how well you predict in Task 1 (in terms of AUC) & Task 2 (in terms of R^2) and how good is your reasoning and analysis in Task 3 plus quality of your code including commenting it.
- We will do our best to run your codes but put every effort to make sure they run! We are using Python 2.6/2.7 version so make sure everything runs ok in 2.6/2.7 before submitting.
- **You can only employ models that we have covered so far in the Lectures**. These include: Linear regression with OLS, Ridge regression, Lasso, Decision Trees, k-NN.
- You can do any pre-processing and any feature engineering/expansion you want.
- Your pdf report should be **max 2 pages** including any figures/appendices/references/poems etc.
- Do not send us the data back with your scripts and report. **Your codes should read in the data in the original format given to you**.
- When reading the data inside your scipts assume **the data lives one level up in the directory**. For example probeA.csv is at *../probeA.csv* from where your python scripts are.
- Write your output to the current directory

[1] See FAQ at the end of this document

- Your submitted scripts just need to be able to produce the **final models** and use them to predict the class or TNA. Do any hyper-parameter tuning with e.g. cross-validation **outside** these scripts.
- You can try out different models but you only need to submit the ones giving you the best performance
- Your *answer.pdf* for Task 3 should be concise and to the point. We will reward information, not amount of text. There is a 2 page limit.
- Monitor the ML Forum from the module website for useful hints and further Q&A.