Assignment 1

SENG 474 / CSC 578D
Data Mining – Spring 2018
Total: 60 marks
Worth 5% of your final grade

Due: January 23rd

Submit a pdf through ConneX before 11:55pm. I will accept submissions up to 24 hours late for a 10% reduction in your mark. After that point, no submissions will be accepted.

Show your work for all questions.

Different marking schemes will be used for undergrad (SEng 474) and grad (CSc 578D) students. Undergrad students do not have to answer the grad questions.

All code questions use the python and/or the sckit-learn library. You may install it, along with the NumPy and SciPy libraries on your own computer. Alternatively, you can work in the lab.
http://scikit-learn.org/stable/install.html
http://www.numpy.org/

1: Decision Trees (SEng 474: 20 points; CSc 578D: 15 points)
a) (SEng 474: 15 points; CSc 578D: 10 points)
By hand, construct the root and the first level of a decision tree for the contact lenses data (attached to this assignment on connex) using the ID3 algorithm. Show the details of your construction and all your calculations; no points will be given for solutions only.

b) (SEng 474: 5 points; CSc 578D: 5 points)
Using the tree.DecisionTreeClassifier module from python's sckit-learn, fit a tree using the contact-lenses data using *criterion='entropy'*.

Compare the entropy values obtain in part a) with the ones calculated by the sklearn.tree module. Explain in detail why the trees are not the same. You may find the documentation for decision trees helpful:
http://scikit-learn.org/stable/modules/tree.html

Note: You can import the data directly from the 'contact-lenses.arff' file using the *Arff2Skl()* converter from *util2.py* provided with this assignment, using these lines of code:

```
from util2 import Arff2Skl

cvt = Arff2Skl('contact-lenses.arff')
label = cvt.meta.names()[-1]
X, y = cvt.transform(label)
```

2: Classifier Accuracy (SEng 474 and CSc 578D: 10 points)
Assume you were given a dataset built from random data, where attributes values have been randomly generated with no consideration to the class labels. The dataset has three classes: "red", "blue" and "yellow". You were asked to build a classifier for this dataset, and told that 50% of the data will be used for training, and 50% for testing. The testing set is balanced, so you can assume it has the same distribution as the training set. Because you are smart, you will start by establishing a theoretical baseline for your classifier's performance.

a) (2 points)
Assume the data is equally split between the three classes (33.3% "red", 33.3% "blue" and 33.3% "yellow") and your classifier systematically predicts "red" for every test instances, what is the expected error rate of your classifier? (Show your work)

b) (3 points)
What if instead of always predicting "red", the classifier predicted "red" with a probability of 0.7, and "blue" with a probability of 0.3. What is the expected error rate of the classifier in this case? (Show your work)

c) (2 points)
Now lets assume that the data is not split equally, but has half (1/2) of its data labeled "red", one-fourth (1/4) labeled as "blue", and one-forth (1/4) labeled as "yellow". What is the expected error rate of the classifier if, as in question a), the prediction is "red" for every test instances.

d) (3 points)
With this dataset (half (1/2) labeled "red", one-fourth (1/4) labeled "blue", and one-forth (1/4) labeled "yellow") What is the expected error rate of the classifier if, as in question b), it predicted "red" with a probability of 0.7, and "blue" with a probability of 0.3. (Show your work)

3: MLE and MAP estimates (SEng 474: 10 points; CSc 578D: 15 points)

a) Let $\theta = P(X=T)$. Calculate the MLE for θ for the following dataset by finding the maximum of $P(D|\theta)$. Show your work. [4 marks]

D={T, T, T, T, T, T, T, F, F, F}

b) Recall the PDF for a Beta random variable is proportional to

$$\Theta^{(\beta 1-1)} * (1-\theta)^{(\beta 2-1)}$$

with parameters $\beta 1$ and $\beta 2$. Let's say you have evidence from previous studies that $P(X=T) = ½$. Let $\beta 1 = 4$. Find $\beta 2$ and then calculate the MAP estimate $P(\theta | D)$ for θ with a Beta($\beta 1, \beta 2$) prior and the dataset above. Show your work. [6 marks]

c) (578D students only) In class we used the mode to find $\beta 1$ and $\beta 2$. Repeat part b using the mean of the beta distribution instead. [5 marks]

4: Gradient Descent (SEng 474: 20 points; CSc 578D: 20 points)

We have a new dataset where the input data has 2 continuous variables (x_1 and x_2), and the task is to predict a continuous value as the output y (i.e. regression). We have reason to believe the following new model for regression is a better fit to the problem domain:

$$\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2}^4$$

A) Write down the error function for this model. You should use the sum of squared error, as in class: [5 marks]

$$E(X) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

B) Derive the gradient descent update for w_0 using learning rate κ [5 marks]

C) Derive the gradient descent update for w_1 using learning rate κ [5 marks]

D) Derive the gradient descent update for w_2 using learning rate κ [5 marks]