10-601 Machine Learning: Assignment 1

- The assignment is due at 3:00pm (beginning of class) on **Monday, January 28**.
- Write your name at the top right-hand corner of each page submitted.
- Each student must hand in a writeup. See the course webpage for the collaboration policy.
- For the programming portions of the assignment, you can use whatever programming language you are comfortable wth. Instructors recommend Matlab for its built-in math and plotting libraries.

1 Q1: Entropy and information gain [10 pts]

Use the following (fictional) dataset for this problem. It is the records of 12 hypothetical patients, with attributes Sex, age Over 60, Diabetic, high Pulse rate, abnormal EKG; and classification HasArrhythmia.

Patient	Sex	Over60	Diabetic	Pulse	EKG	HasArrhythmia
1	M	+	+	-	-	-
2	M	-	-	+	+	+
3	M	-	+	+	-	-
4	M	+	-	-	+	+
5	M	+	+	+	-	+
6	M	-	+	+	-	+
7	F	-	-	+	-	-
8	F	+	+	+	+	+
9	F	-	+	-	+	+
10	F	+	-	-	-	-
11	F	+	+	-	-	-
12	F	+	-	+	+	+

1. Calculate the conditional entropy, $H(Arrhythmia|Sex = Female)$.
2. Refer to section 3.7.5 of Mitchell. Under the attribute selection measure $\frac{Gain^2(S,A)}{Cost(A)}$, what would be the first split in the tree? Assume that $Cost(Sex) = Cost(Over60) = 1$, $Cost(Diabetic) = 3$, $Cost(HighPulse) = 2$, $Cost(AbnormalEKG) = 5$.
3. Suppose that, on a different set of patients, we knew their exact ages. Ages of positive examples are: {40, 60, 62, 64, 70, 74, 75, 82} and negative examples are: {33, 35, 42, 45, 49, 52, 58, 59, 80}. Suppose that all other attributes in the data set are poor predictors, so we want to split the tree at $Age = k$ by dividing the continuously-valued data points into two groups, $Age \geq k$ and $Age < k$. What division might we choose, based on information gain?

2 Q2: Growing and pruning decision trees [40 pts]

1. Implement ID3 to train a decision tree for the German Credit Approval Data Set (see course webpage for train/validation sets and description). As you are building the tree, record the error rates on both the training set and the test set, making a plot similar to Fig. 3.6 in Mitchell. Show the resulting tree at completion.

2. Now implement reduced-error pruning to prune the tree. Record the error rate for the test set as the tree is being pruned, to produce a plot similar to Fig. 3.7 in Mitchell. Show the resulting tree at completion.

3. Interpret your plots.

4. Using your **node**-based implementation of pruning in the previous question as a starting point, describe what you would need to modify in order to implement a **rule**-based pruning routine. Specifically, describe which sections/routines would need to be changed, and, using pseudocode, demonstrate how. For this data set, do you expect this rule-based method to perform better/worse/the same as the node-pruned tree? Which types of datasets would do better under a rule-based pruning regime, and vice-versa? Please justify your answers.

5. Instructions for submitting code will be announced.

3 Q3: Probability basics [25 pts]

1. Using only the axioms of probability, prove $P(A|A,B) = 1$.

2. Using only the axioms of probability, show that $P(A,B|C) = P(A|C)P(B|C)$ only holds when A or B is marginally independent of C– that is, $P(A|B,C) = P(A|C)$ or $P(B|A,C) = P(B|C)$.

3. You have two bags. The first bag contains 12 white marbles and 7 black marbles. The second contains 9 white marbles and 4 black marbles. One bag is chosen uniformly at random, and then a marble is selected from the bag. The marble selected is white. What is the probability it came from the first bag?

4 Q4: Maximum likelihood estimators [25 pts]

1. Let $X_1, X_2, ..., X_n \backsim Uniform(a,b)$ where a and b are parameters, $a < b$. Find the MLE $\hat{a}$ and $\hat{b}$.

2. While MLE's can sometimes be found analytically, for complicated likelihood functions it may need to be computed using numerical methods. One method is the Newton-Raphson algorithm, which iteraively finds a sequence of $\theta^0, \theta^1, ...$ that (under ideal conditions) converges to the MLE $\hat{\theta}$.

   Note that one may expand the derivative of the log-likelihood function around θ^j:

   $$0 = l'(\hat{\theta}) \approx l'(\theta^j) + (\hat{\theta} - \theta^j) l''(\theta^j)$$

   Solving for $\hat{\theta}$ gives

   $$\hat{\theta} \approx \theta^j - \frac{l'(\theta^j)}{l''(\theta^j)}$$

This leads to an interative scheme where

$$\hat{\theta^{j+1}} = \theta^j - \frac{l'(\theta^j}{l''(\theta^j)}$$

For the binomial sampling function, pdf $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, find the MLE using Newton-Raphson, starting with an estimate $\theta^0 = 0.1$, $n = 100$, $x = 8$. Show the resulting θ^j until it reaches convergence ($\theta^j - \theta^j < .01$). (Note that the binomial pdf may be calculated analytically– you may use this to check your answer.) Submit code as in question 2.4.