CSE446 Machine Learning, Spring 2017: Homework 1

Due: Thursday, April 20^{th}, beginning of class

Start Early! Also, typed solutions (specifically those in LaTeX) are preferred to hand-written solutions. Any illegible solutions will be counted wrong at the sole discretion of the grader. Please feel free to use the homework document as a template, putting your solutions inline. We will only accept answers in **.pdf** format. Solutions should be submitted at
https://catalyst.uw.edu/collectit/assignment/dhlee4/40217/160249

1 Probability Review [30 points]

1. You are just told by your doctor that you tested positive for a serious disease. The test has 99% accuracy, which means that the probability of testing positive given that you have the disease is 0.99, and also that the probability of testing negative given that you do not have the disease is 0.99. The good news is that this is a rare disease, striking only 1 in 10,000 people.

   a (5 points) Why is it good news that the disease is rare?

   b (10 points) What are the chances that you actually have the disease? Show your work.

2. A group of students were classified based on whether they are senior or junior and whether they are taking CSE446 or not. The folowing data was obtained.

	Junior	Senior
taking CSE446	23	34
no CSE446	41	53

   Suppose a student was randomly chosen from the group. Let J be the event that the student is junior, S be the event that the student is senior, C be the event that the student is taking CSE446, and $\bar{C}$ be the event that the student is not taking CSE446. Calculate the following probabilities. Show your work.

   a (5 points) $P(C, S)$

   b (5 points) $P(C|S)$

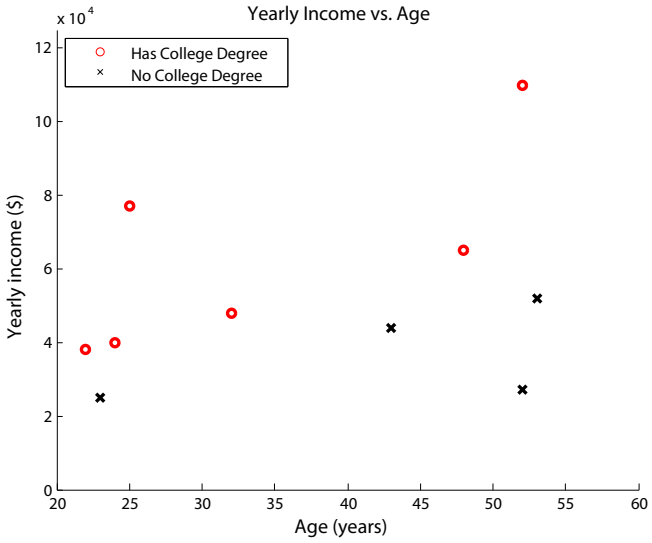
   c (5 points) $P(\bar{C}|J)$

2 Decision Trees [30 points]

For the first two problems, it would be helpful for you to draw the decision boundary of your learned tree in the figure.

1. *(14 points)* Consider the problem of predicting if a person has a college degree based on age and salary. The table and graph below contain training data for 10 individuals.

Age	Salary ($)	College Degree
24	40,000	Yes
53	52,000	No
23	25,000	No
25	77,000	Yes
32	48,000	Yes
52	110,000	Yes
22	38,000	Yes
43	44,000	No
52	27,000	No
48	65,000	Yes



Build a decision tree for classifying whether a person has a college degree by greedily choosing threshold splits that maximize information gain. What is the depth of your tree and the information gain at each split?

2. (12 points) A multivariate decision tree is a generalization of univariate decision trees, where more than one attribute can be used in the decision rule for each split. That is, splits need not be orthogonal to a feature's axis.

   For the same data, learn a multivariate decision tree where each decision rule is a linear classifier that makes decisions based on the sign of $\alpha x_{age} + \beta x_{income} - 1$.

   Draw your tree, including the α, β and the information gain for each split.

3. (4 points) Multivariate decision trees have practical advantages and disadvantages. List an advantage and a disadvantage multivariate decision trees have compared to univariate decision trees.

3 MLE [20 points]

This question uses a discrete probability distribution known as the Poisson distribution. A discrete random variable X follows a Poisson distribution with parameter λ if

$$\Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad k \in \{0, 1, 2, \dots\}$$

You are a warrior in Peter Jackson's The Hobbit: Battle of the Five Armies. Because Peter decided to make his battle scenes as legendary as possible, he's decided that the number of orcs that will die with one swing

of your sword is Poisson distributed (i.i.d) with parameter λ. You swing your sword eight times in the scene. Later, you go see the movie in theaters and record the number of orcs slain during each swing of your sword:

Sword Swing	1	2	3	4	5	6	7	8
Orcs Slain	6	4	2	7	5	1	2	5

Let $G = (G_1, \ldots, G_n)$ be a random vector where G_i is the number of orcs slain on swing i:

1. *(6 points)* Give the log-likelihood function of G given λ.
2. *(8 points)* Compute the MLE for λ in the general case.
3. *(6 point)* Compute the MLE for λ using the observed G.

4 Programming [100 points]

Go to https://github.com/dhlee4/C4.5-Homework