CMSC 726 : Machine Learning *First Last, UID, email*

HW1

0.1 Instructions

- Submit your assignment via ELMS, at `http://elms.umd.edu`.
- Your submission should include the PDF file only (no need to submit the MATLAB code for this assignment).
- You may use LaTex or Word to create your submission, but *you must submit in PDF format.*
- For your convenience, you can simply edit the included .tex file. To do so,
  1. Edit the header information with your own personal information (otherwise, I won't know whose submission it is).
  2. Uncomment the `\begin{solution} ... \end{solution}` blocks and fill in your solution.
- To add an image to this tex file, use the command `\includegraphics{filename}` (search the LaTeXdocumentation for additional arguments).

1 Basic Probability

(This problem is similar to PRML Problem 1.3, but *not identical.*)

Suppose that we have three colored boxes, r (red), b (blue) and g (green). Box r contains 4 apples, 3 oranges and 5 limes; box b contains 2 apple, 5 orange and 1 limes; box g contains 1 apples, 4 oranges and 7 limes. If a box is chosen at random with probabilities $P(r) = 0.2, P(b) = 0.3, P(g) = 0.5$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then,

1. What is the probability of selecting an apple?
2. If we observe that the selected fruit is in fact an an orange, what is the probability that it came from the green box?

2 MLE and MAP

2.1 Maximum Likelihood Estimation (MLE)

The Poisson distribution is a useful discrete distribution which can be used to model the number of occurrences of something per unit time. For example, in networking, packet arrival density is often modeled with the Poisson distribution. That is, if we sit at a computer, count the number of packets arriving in each time interval (say every minute, for 30 minutes) and plot the histogram of how many time intervals had X number of packets, we expect to see something like the Poisson probability mass function.

The Poisson PMF is defined as,

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

(For the purposes of this problem, this is everything you need to know about Poisson distribution.)

It can be shown that the parameter λ is the mean of the Poisson distribution. In this part, we will estimate this parameter from the number of packets observed per unit time $X_1, \ldots, X_n$, which we assume are drawn I.I.D. from $Poisson(\lambda)$. Show that $\hat{\lambda} = \frac{1}{n}\sum_i X_i$ is the maximum likelihood estimate of λ.

2.2 Maximum A Posteriori (MAP)

Now let's get Bayesian and put a prior distribution over the parameter λ.

Your friend in networking hands you a typical plot showing the counts of computers at a university cluster with different average packet arrival densities. Your extensive experience in statistics tells you that the plot resembles a Gamma distribution PDF, so you believe a good prior distribution for λ may be a Gamma distribution.

Recall that the Gamma distribution has pdf:

$$P(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \qquad \lambda > 0$$

Assuming that λ is distributed according to $\Gamma(\alpha,\beta)$, define an analytic expression for the MAP of λ.

(Hint: it may be helpful to consider that, if $\lambda \sim \Gamma(\alpha,\beta)$, then it has mean α/β and the mode is $(\alpha-1)/\beta$ for $\alpha > 1$. This fact isn't absolutely necessary to solving this problem, but can provide an alternate proof.)

2.3 MATLAB Implementation

You will now use the equations you've just derived to compute the MLE and MAP estimates for some sampled data.

1. In MATLAB, load the file "poisson_data.mat" to restore variables X_1 and X_2 – both of which are sampled from the same Poisson distribution, one with 100 samples and one with 10,000 samples.
2. Plot histograms of each data set, using 50 bins (or whatever looks good to you). Save them as image files and add them to your homework submission.
3. Compute the MLE of λ for both data sets.
4. For each data set, compute the MAP estimate of λ, using $\alpha = 10$ and $\beta = 10$.
5. On two separate plots (one for each data set), plot the Poisson curve using
   (a) the MLE of λ (in red)
   (b) the MAP estimate of λ (in blue – *on the same plot*)

   Save the plots as image files and add them to your homework submission.
6. Repeat steps 3-5 using $\alpha = 1000, \beta = 1000$.

Hint: In order to plot the Poisson curve, you will need to compute it for various values of t.

3 Jensen's Inequality

(PRML Problem 1.40)

By applying Jensen's inequality (equation 1.115) with $f(x) = ln(x)$, show that the arithmetic mean A of a set of real numbers X is never less than their geometric mean G.

Note: Given a set of real numbers $X = \{x_1, \dots, x_n\}$,

- the arithmetic mean A is defined as

$$A \triangleq \sum_{i}^{n} x_i$$

- the geometric mean G is defined as

$$G \triangleq \left(\prod_{i}^{n} x_i \right)^{\frac{1}{n}}$$

4 Decision Trees

Consider the following set of training examples for an unknown target function $< X_1, X_2 > \rightarrow Y$. Both X_1, X_2 are boolean and Y takes values $\{+, -\}$. Each row indicates a permutation of the values, and how many times the given permutation occurs in the training set. For example, $(+, T, T)$ was observed 3 times, while $(-, T, T)$ was never observed.

Y	X_1	X_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

Table 1: Training data

4.1 Entropy

Compute the sample entropy $H(Y)$ for this training data (using $\log_2$). Show your work as much as possible.

4.2 Information Gain

Now, compute the information gains $IG(X_1)$ and $IG(X_2)$. Show your work as much as possible.

4.3 DT Learning

Now, draw the decision tree that would be learned by the ID3 algorithm, given this data set.

(It is strongly suggested that you use some sort of diagramming software, such as Omnigraffle, Microsoft PowerPoint or Microsoft Visio, to generate an image file, though we will accept hand-written, scanned images.