COS424
Homework #1
Due Tuesday, February 28, 2012

Turn in your written assignments in class on Tuesday, February 28th and submit your electronic files for R programing questions to CS DropBox at http://dropbox.cs.princeton.edu/COS424_S2012/Homework_1 by the end of the same day.

Question 1. (10 points) An *indicator function* is equal to one when its argument is true, and zero otherwise. Specifically,

$$\mathbf{1}(x) = \begin{cases} 1 & \text{when } x \text{ is true.} \\ 0 & \text{otherwise} \end{cases}$$

Let, X be the outcome of a (fair) six-sided die. Compute

1. The probability that $X > 2$.
2. $\mathrm{E}[\mathbf{1}(X > 2)]$

Question 2. (10 points) Let $C_1, C_2, and C_3$ each be coins such that

$$\begin{aligned} P(C_1 = H) &= 0.5 \\ P(C_2 = H) &= 0.7 \\ P(C_3 = H) &= 0.5 \end{aligned}$$

Suppose we first flip C_1 and record its outcome Z. If it is heads, then we flip C_2 twice and record the outcomes (X, Y). If it is tails, then we flip C_3 twice and record the outcomes (X, Y).

- Compute the joint probability distribution of the outcomes of the flips $P(X, Y)$.
- Compute $P(Z \mid X = H)$.
- Compute $P(X \mid Y = H)$.

Question 3. (10 points) Describe three random binary variables, X, Y, and Z, such that $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$. Give the joint probability distribution of the three variables. Show that your answer is correct using the definitions of marginal independence and conditional independence. i.e. show that $P(X, Y) = P(X)P(Y)$, and $P(X, Y \mid Z) \neq P(X \mid Z)P(Y \mid Z)$.

Also describe a scenario, real or imaginary, in which the random variables have these properties.

Question 4. (10 points) Let $(x_1, \ldots, x_N)$ be continuous observations assumed IID Gaussian with mean μ and variance σ^2. Write down the log likelihood function. Derive the maximum likelihood estimate of μ. Derive the maximum likelihood estimate of σ^2 given an estimate of $\mu = \hat{\mu}$.

Question 5. (10 points) Bayes rule for medical diagnosis (Exercise 2.9.0.5 in MLAPA)

After your yearly checkup, the doctor has bad news and good news. The bad news is that you are tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 99%, as is the probability of testing negative given that you dont have the disease.) The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

The following two questions have an R programming component and a written component. The write-up should be clear, concise, and thoughtful and include relevant output from the programming component, such as plots. Your grade for these questions will be determined by the completeness, correctness, and clarity of your programming and writeup.

Please submit the write-up for these questions along with the written portion of your assignment, and the programming portion of these questions through csdropbox. After each question, we have made clear what the electronic submission entails.

Before beginning, we'd like to point you to a *partial* list of functions that might be useful in these problems. Note that R has excellent documentation. Use the `help` command and `help.search` command liberally.

- `plot`
- `plot.window` and `axis`
- `segments` and `points`
- `rbinom` (use size=1 for the Bernoulli)

Question 6 (R). (25 points) Recall the Gaussian density,

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

Collect 25 continuous univariate data points from the real world (the real world includes the web). Write a function that takes the data as input and plots the following on a single plot:

- the data
- the Gaussian density fit to the data
- a vertical line at the sample mean $\hat{\mu}$
- a horizontal line with length equal to the sample standard deviation $\sqrt{\hat{\sigma}^2}$. It should be centered at the mean and at height $p(\hat{\mu} - \sqrt{\hat{\sigma}^2} \mid \hat{\mu}, \hat{\sigma}^2)$
- 25 additional points drawn from the same distribution and plotted to be distinguished from the original data

Explain where your data came from. What do you notice about the data you collected compared with the fitted Gaussian?

Electronic submissions for this question:

- *A data file named Question6.txt containing your 25 data points.*
- *A file named Question6.R containing a single entry-point function named "make.plot" that will read in your data and generate the plots.*

```
make.plot <- function() {
# your code here
}
```

- *A file named Homework1.pdf containing your write-up and your plots. This file is for both question 6 and 7*

Question 7 (R). (25 points) Write a function that takes a binary vector of data as input and returns the Bernoulli log likelihood *function* $\mathcal{L}(\pi)$ given that data. Specifically, the returned function takes one argument $\pi \in (0, 1)$ and returns the log likelihood of the data given that parameter.

Sample 100 data points $(x_1, \ldots, x_{100})$ from a Bernoulli with $\pi^* = 0.5$. For various values of N, plot the log likelihood function over $\pi \in (0, 1)$ using data $(x_1, \ldots, x_N)$ on a single plot.

How does the log likelihood function change for different values of N? Why?

Electronic submissions for this question:

- *A file named Question7.R containing a single entry-point function named "make.plot" that will generate the plots, and a function named "Bernoulli.log.likelihood.function." This function takes a data set and returns a function. The returned function takes a parameter and returns the Bernoulli log likelihood evaluated for the data and parameter.*

```
Bernoulli.log.likelihood.function <- function(x) {
  # your code here
  }

make.plot <- function() {
 # your code here
}
```

- *A file named Homework1.pdf containing your write-up and your plots. This file is for both question 6 and 7*

Extra credits: Question 8 and 9 are extra credit

Question 8. (10 points) Express mutual information in terms of entropies (Exercise 2.9.0.15 in MLAPA)

Show that

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Question 9. (10 points) Mutual information for naive Bayes classification with binary features (Exercise 5.7.0.11 in MLAPA)

Feature selection is to remove features that are irrelevant to classification. The simplest approach to feature selection is to evaluate the relevance of each feature separately, and then take the top k features. One way to measure relevance is to use mutual information between feature X_j and the class label Y. The mutual information can be thought of as the reduction in entropy on the label distribution once we observe the value of feature j. If a naive Bayesian classifier is used for classification and the features are binary, show that the mutual information can be computed as follows:

$$I_j = \sum_c [\theta_{jc}\pi_c log\frac{\theta_{jc}}{\theta_j} + (1-\theta_{jc})\pi_c log\frac{1-\theta_{jc}}{1-\theta_j}]$$