CS-470 Machine Learning

Assignment-1

Due Date: Thursday Mar-16-2017

Most of these problems have been taken from the 2nd Edition of Garcia's book. Some problems have been designed by me.

Instructions:

- Please submit only the Assignment problems and NOT the Practice Problems.
- Only submissions through LMS will be considered. Try to submit ahead of time; broken internet connections and the terrible service provided by your ISP are not acceptable excuses.

Axioms of Probability

Question-1 (Practice Problem) A random experiment has sample space $S = \{a, b, c\}$. Suppose that $P[\{a, c\}] = 5/8$ and $P[\{b, c\}] = 7/8$. Use the axioms of probability to find the probabilities of the elementary events. (**Answer:** $P[\{a\}] = \frac{1}{8},\ P[\{b\}] = \frac{3}{8},\ P[\{c\}] = \frac{4}{8}$)

Question-2 (Assignment Problem) Show that

$$P[A \cup B \cup C] = P[A] + P[B] + P[C] - P[A \cap B] - P[A \cap C] - P[B \cap C] + P[A \cap B \cap C]$$

Conditional Probability

Question-3 (Practice Problem) Show that $P[A \cap B \cap C] = P[A|B \cap C]P[B|C]P[C]$.

Question-4 (Assignment Problem) A die is tossed twice and the number of dots facing up is counted and noted in the order of occurrence. Let A be the event "total number of dots is even," and B be the event "both tosses had an even number of dots,". Find $P[A|B]$ and $P[B|A]$. (**Answer:** $P[A|B] = 1,\ P[B|A] = \frac{1}{2}$).

Question-5 (Practice Problem) A data packet, transmitted from router-1, takes between 0 to 60 seconds to arrive at router-2. Assume that the arrival time is uniformly distributed between 0 to 60 seconds. Find the probability that the packet will arrive on (or before) 31 seconds if it has not arrived by 30 seconds.

Question-6 (Assignment Problem) A nonsymmetric binary channel is shown in Figure 1. Assume that the inputs are equiprobable.

a) Find the probability that the output is 0

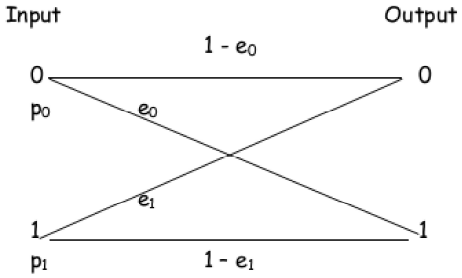

Figure 1

b) Find the probability that the input was 0 given that the output is 1. Find the probability that the input is 1 given that the output is a 1.

Question-7 (Assignment Problem) A 1 year old baby is learning to express her emotions though vocalizations. Currently she has a vocabulary of only three words '*ah*', '*bah*' and '*cha*'. At one time she uses one of these three words to express two emotions: '*Happy*' or '*Sad*'. Based on analysis of her audio recordings we know that most of the time she expresses happiness via and '*ah*'. Similarly, sadness is expressed via a '*cha*' most of the time. For the two emotional states, probabilities of hearing any of the three utterances are shown in Figure 2.

Prior analysis indicates that baby is *Happy* 60% of the time and *Sad* 40% of the time. Calculate the probabilities of the two emotional states when you hear a '*bah*'?

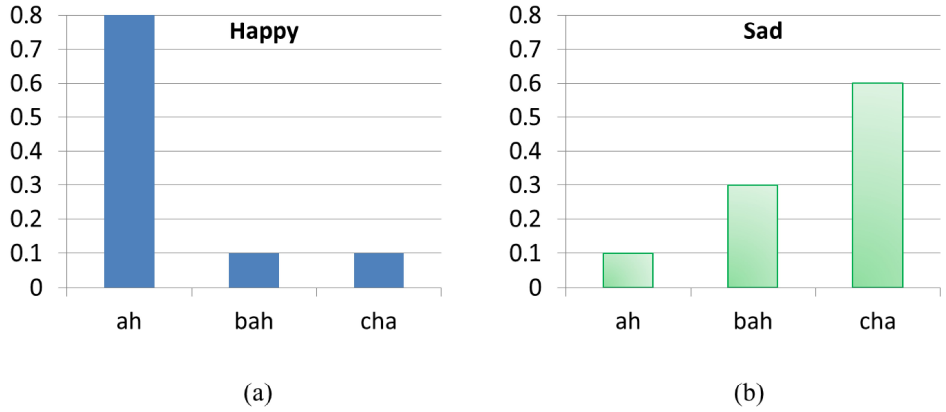

Figure 2 Probabilities of hearing the three distinct vocalizations when baby is (a) *Happy* and (b) *Sad*.

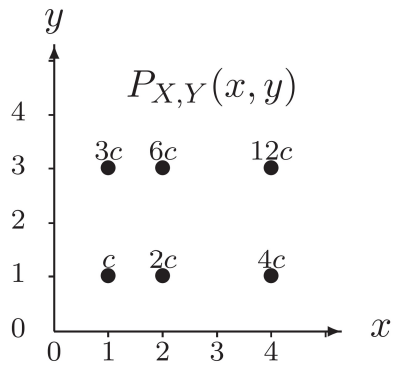

Figure 3 PMF for Question-8. Here, c= 1/28.

Correlation Analysis

Question-8 (Practice Problem) The joint pmf of discrete random variables X & Y is shown in Figure 3. The value of the constant $c = 1/28$, find

a) The marginal pmfs of X&Y, $P_X(x)$& $P_Y(y)$.
b) The expected values $E[X]$& $E[Y]$.
c) The standard deviations σ_X & σ_Y.
d) The expected values of Y/X.
e) The correlation $E[XY]$.
f) The covariance $Co\ \ [X,Y]$.
g) The correlation-coefficient ρ_{XY}.
h) Are the values of covariance and correlation-coefficient what you expect them to be?

Question-9 (Assignment Problem) In this question you will analyze some data samples using correlation analysis; you will need to download the mat file '`CorrelationData.mat`' from LMS. Load this file into matlab; you should see 8 vectors $[X_1, \ldots X_4]$ & $[Y_1, .., Y_4]$.

a) Use the `scatter( )` function to plot a scatterplot of the following 4 pairs of data-vectors $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)$. Comment on what you expect the correlations-coefficients between these 4 data-vectors to be.
b) Now the compute correlation, covariance and correlation coefficients between the 4 pairs of data-vectors. In order to compute these metrics from data you will need to use the ***estimators*** described below. Compare the value of correlation-coefficients with the correlation-coefficient you obtain from matlab using the `corrcoef( )` function; you should get the same values.

A Word on Estimators: One difference between computing metrics using pencil and paper and actual data is that when working with paper you are generally given the relevant pdfs or pmfs that you need to use (as is the case in question-8). For example, if you have the joint-pdf $f_{XY}(x,y)$ of the random variables X & Y then you can use the familiar formula

for correlation which is $E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)dxdy$. However, the problem in real life is that most of the time you do not know the pdf $f_{XY}(x,y)$ and therefore, must estimate the correlation from the data itself. In such scenarios, you will need to use, what is called, the *estimator* of the metric you are interested. For example, estimators for our metrics of interest are given below:

Correlation: $c_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$

Here, n is the number of data samples ($n = 200$ for our data); x_i& y_i represent the i-th sample (or value) of the random variables X & Y respectively.

Covariance: $v_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Here, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the estimator of the mean and is called the *sample-mean*

Correlation-Coefficient: $\rho_{XY} = \frac{v_{XY}}{s_X s_Y}$

Here, $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the estimator of the standard-deviation and is called the sample-standard deviation. The sample standard-deviation s_Y for Y can be calculated similarly.