CIS 490 MACHINE LEARNING HOMEWORK 1

(Full Score: **20** + 5 bonus points)

September 22, 2017 Fall

NAME:

STUDENT ID:

INSTRUCTOR: Dr. Julia Hua Fang

1. **(Total 2 points)** What is machine learning (0.5 point)? What is supervised learning (0.5 point)? unsupervised learning (0.5 point)? Please describe the relationship among machine learning and statistical learning (0.5 point).

**(1 Bonus point)** What is Semi-supervised learning (0.5 point)? What is Reinforcement learning (0.5 point)?

1. **(Total 2 point)** What is a Random variable? Give an example. (1 point)

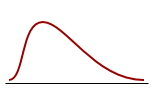
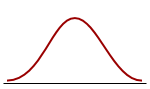
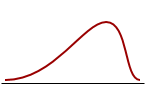
What are the two types of Random Variables? Give three examples of probability distributions for each type. (1 point)

1. **(Total 4.5 points)** Write out the pdf or pmf **and** cdf of following probability distributions (hint: pdf for continuous distributions, pmf for discrete distributions. Both continuous and discrete distributions have cdf) (0.5 point for each distribution, including .25 point for pdf/pmf and .25 point for cdf)

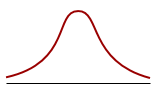
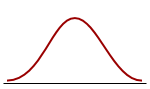
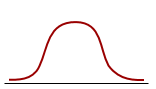
**(Note: Please type in the formulas. Do not copy or paste them from websites)**

* 1. General Gaussian/Normal
  2. Beta
  3. Gamma
  4. Uniform
  5. Student t
  6. Chi-square
  7. Bernoulli
  8. Binomial
  9. Poisson

**2 Bonus points**: Write out pdf/pmf and cdf for the following distributions(Optional): (0.5 point for each distribution, including .25 point for pmf and .25 point for cdf)

1. pmf and cdf for negative binomial distribution
2. pdf and cdf for
   1. Pareto
   2. Lognormal
   3. Weibull
3. **(Total 1 point)** (1) Which one is negatively skewed distribution?(0.5 point)

(a) (b) (c)

(2) Which distribution has heavier tails?(0.5 point)

(a) (b) (c)

1. (**Total 2 points**) What is Central Limit Theorem? Give an example
2. (**Total 3 point**) What are the measures of central tendency/location of a distribution (1.5 point)? What are the measures of variability/dispersion of a distribution? (1.5 point)
3. (**Total 1.5 points)** Consider two random variables X (0 = male; 1 = female) and Y (0= low risk; 1= medium risk; 2 = high risk) with joint pmf given in the Table below.

Table Joint pmf of X and Y

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y=0 | Y=1 | Y=2 |
| X=0 | 1/3 | 1/4 | 1/5 |
| X=1 | 1/6 | 1/6 | 1/8 |

1. p (X = female, Y= high risk) = ? (0.5 point)
2. p( X = female ) = ? (0.5 point)
3. p (Y= high risk|X=female) = ? (0.5 point)

# (Total 4 points)

Use R or Matlab to

1. Randomly generate two random variables, x1 and x2, from a probability distribution of your choice (**use a random seed**, so the grader can replicate your results; comment the distribution you used for random data generation).
2. Compute Expected Value, Variance, Standard Deviation, Mode, Median, Skewness and Kurtosis for x1 and x2 generated for Question (a).
3. Compute the correlation and covariance matrix of x1 and x2 generated for Question (a).

Write out your formula to compute these statistics and your answers on the paper.

Submit your R/Matlab scripts as a whole for generating your data, computing these descriptive statistics, displaying your results and exporting results into two formats, \* .txt and \*.xlsx . The grader will run your **automated** codes and replicate your results.

[name your R script as lastname\_yourstudentid\_random.r or your Matlab script as lastname\_yourstudentid\_random.m]

**(2** Bonus points): download one of the following datasets from University of California, Irvine (UCI) Machine Learning Repository <http://archive.ics.uci.edu/ml/>(optional)

* Iris
* Wine
* Breast Cancer Wisconsin
* Heart Disease
* Waveform Database Generator (Version 2) Data Set

1. Compute Expected Value, Variance, Standard Deviation, Mode, Median, Skewness and Kurtosis of all attributes (hint: X);

(d) Compute the correlation and covariance matrix among two or more attributes of your choice (hint: X);

**Submit** your R/Matlab script as a whole for importing the data set you choose, computing these descriptive statistics, displaying your results and exporting results into two formats, \* .txt and

\*.xlsx . Attach the dataset of your choice. The grader will run your automated codes and replicate your results. [name your R scriot as lastname\_yourstudentid\_real.r or your matlab script lastname\_yourstudentid\_real.m ]