CIS 490 MACHINE LEARNING HOMEWORK 1

(Full Score: **15** + 5 bonus points)

September 22, 2017 Fall

NAME: Kevin D’Cruz

STUDENT ID: 01656193

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)?

   Answer:

   Machine learning: It can be defined as methods that help detecting data patterns automatically, with data in the future to be predicted using patterns which are yet not covered, or to make decision activities under uncertainty, like collecting more data. The problems addressed by ML are broken down into Supervised, Unsupervised and Reinforcement Learning. Apart from this, we have Semi Supervised learning which falls in between Supervised and Unsupervised Learning. The goal is to build intelligent systems which can learn from their past experiences.

   Supervised Learning: It is a type of problem addressed by machine learning with a task of obtaining a function from labeled training data, where the training data consists of a set of training examples. There is an input object (typically a vector) and a desired output value pair. The supervised learning algorithm goes through/analyzes the training data and produces a result function, which helps in mapping new examples. In other words, to build a function 'f' that maps inputs 'X' to output 'Y', given a training dataset with a known value of Y. Example: Classification, Regression, SVM. It follows a predictive approach.

   Unsupervised Learning: It is a type of problem addressed by machine learning with a task of obtaining a function to describe a hidden structure from unlabeled data. Since the examples given to the learner are unlabeled, the accuracy of the structure cannot be evaluated, which is the relevant algorithms output. Examples include Clustering (k means) and Dimension Reduction(PCA). It follows a descriptive approach. Only X can be visible and there are no measurements for Y, which means, Y is an unknown label.

   Relationship between machine learning and statistical learning: Both statistical models and machine learning have a similarity that both learn from data. This is where data mining comes into picture. Also, both are based on probabilistic or graphical models. There may be same problems but different models to be made. However, there are certain machine learning models which are not based on a probabilistic model. SVM is one such example of a ML model. Some of the differences among the two are: ML focusses more on prediction and analysis of learning algorithms whereas Statistics focusses on understanding the data in terms of models.

   Semi-Supervised Learning: As said earlier, Semi-Supervised Learning falls somewhere between Supervised and Unsupervised Learning. Here, along with unlabeled data, there is some labelled data (supervision information) given to the algorithm for some examples, which are generally the associated targets. Example, the inputs X could be divided into X_1 and X_2 where X_1 has been given labelled data outputs Y_1 whereas X_2 is not provided with any output Y_2. These techniques are usually used when labelled data is very expensive to obtain. An application example of Semi-Supervised Learning could be Speech Analysis or Medical applications.

Reinforcement Learning: It is a Machine Learning problem where a software agent or a machine is allowed to learn its own behavior. This is done based on the environment feedback. The behaviour could be learnt at once or can be kept on learning as time goes by. This is a branch of AI. There can be Reinforcement Learning methods used for military applications. Another example is playing chess with a computer. RL is learning of the best possible actions based on a punishment or a reward.

2. **(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)

Answer:

Random Variable: It is a variable which has an unknown value, or a function having its values assigned to each of the experiment's outcomes. They are usually real numbers since it is needed for them to be measured. The outcomes of a random occurrence in statistics could be quantified using a random variable. Example: Let A represent the event where a coin is tossed 3 times. Hence, there could be 8 possibilities for X from (H, H, H) to (T, T, T).

Random variables are classified as **Discrete** and **Continuous** Variables.

Discrete variables are those variables that have specific values, in other words, a countable number of distinct values. Ex: The number of fish in a fish pool, the number of fruits in a basket.

Probability Distributions:

1. Binomial: The successes number in a sequence of yes or no independent trials, each returning a success with probability

- Multinomial: Generalized from a Binomial Distribution, used when there more categories greater than 2

- Poisson: This type of probability distribution can be used to model how many times an event occurs as per intervals of space or time

Continuous variables are those variables whose values can be anything within a continuous range. In other words, continuous variables take in infinite values. Eg. Measurement quantities. An integral, or, an area under a curve is used to represent a continuous variable

Probability Distributions:

- Uniform: Rectangular Distribution, has a constant probability
- Gaussian: Bell shaped curve, symmetric about the mean
- Pareto

3. **(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

PDF: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

CDF: $\frac{1}{2}[1 + \mathrm{erf}(\frac{x-u}{\sigma\sqrt{2}})]$

Variance: $\sigma^2 > 0$
Mean: $\mu \in R$, R: real number

2) Beta

PDF: $\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ Where, $B(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

CDF: $I_x\,(\alpha, \beta)$

$\alpha, \beta > 0$ and real

3) Gamma

PDF: $\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$ $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

CDF: $\frac{1}{\Gamma(k)} \gamma(k, \frac{x}{\theta})$ $\frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x)$

Here, shape k, rate β, shape α, scale θ are all > 0

4) Uniform

PDF: $\begin{cases} \frac{1}{b-a} & for\ x \in [a, b] \\ 0 & otherwise \end{cases}$

CDF: $\begin{cases} 0 & for\ x < a \\ \frac{x-a}{b-a} & for\ x \in [a, b) \\ 1 & for\ x \geq b \end{cases}$

Here, the Range is: $-\infty < a < b < \infty$

5) Student t

PDF: $\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\boldsymbol{v\pi}}\Gamma\left(\frac{v}{2}\right)}(\mathbf{1}+\frac{\boldsymbol{x}^{\mathbf{2}}}{\boldsymbol{v}})^{-\frac{\boldsymbol{v}+\mathbf{1}}{\mathbf{2}}}$

CDF: $\frac{\mathbf{1}}{\mathbf{2}}+\boldsymbol{x}\Gamma\left(\frac{v+1}{2}\right)\ X\ \frac{2F1\ \left(\frac{1}{2},\frac{v+1}{2},;\frac{3}{2},;-\frac{x^2}{v}\right)}{\sqrt{\pi v}\Gamma(\frac{v}{2})}$

Here v is Degree of freedom

$v>\ 0$, Real

${}_2F_1$ is the hypergeometric function.

6) Chi-square

PDF: $\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$

CDF: $\frac{\mathbf{1}}{\Gamma(\frac{k}{2})}\ \gamma(\frac{k}{2},\frac{x}{2})$

Here, Degree of freedom $k\ \in\ N_{>0}$ and N=Natural

7) Bernoulli

PMF: $\begin{cases} q=(1-p)\ \ for\ k=0 \\ p\qquad\qquad\ \ for\ k=1 \end{cases}$

CDF: $\begin{cases} \quad 0 & for\ k<0 \\ 1-p & for\ 0\ \leq k<1 \\ \quad 1 & for\ k\ \geq 1 \end{cases}$

0 < p < 1

Also, p ∈ $Real\ Number$

8) Binomial

PMF: $\binom{n}{k}p^k(1-p)^{n-k}$

CDF: $I_{1-p}(n-k,1+k)$

$n\ \in\ N_0$ (Trials)

Success probability: $p\ \in(0,1)$

9) Poisson

PMF: $\frac{\lambda^k e^{-\lambda}}{k!}$

CDF: $\frac{\Gamma(\lfloor k+1\rfloor,\lambda)}{\lfloor k\rfloor!}$

$\lambda > 0$, Real

floor function: k
When k ≥ 0, Γ: Upper Incomplete Gamma Function and
Q: Regularized Gamma Function

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)

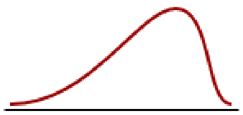
1) pmf and cdf for negative binomial distribution

2) pdf and cdf for
   a. Pareto

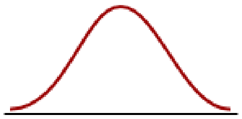
   b. Lognormal

   c. Weibull

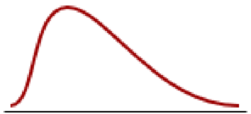
4. **(Total 1 point)** (1) Which one is negatively skewed distribution? (0.5 point)



(a)



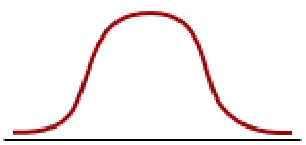
(b)



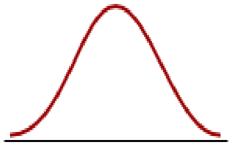
(c)

Answer: A

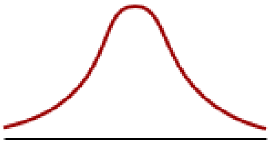
(2) Which distribution has heavier tails? (0.5 point)



(a)



(b)



(c)

Answer: C

5. (**Total 2 points**) What is Central Limit Theorem? Give an example

Answer: Considering that there exists a population with mean μ and a standard deviation of σ, such

that we take out random samples with replacement from the population, then an approximate normal distribution will be obtained, in terms of the distribution of the mean samples. This result which would be obtained would be irrespective of the type of population source, it could be either skewed or normal. However, the sample size should be large. For a normal population, central theorem would work even for smaller samples.
Example: Calculating a probability that the mean is greater than a certain number. Given Standard Deviation sample size and mean, calculating the probability. This can be done using the z-formula.

6. (**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)

Answer: The central tendency/location is the center or middle of a distribution. The measures of central tendency are mean, median and mode. Most commonly occurring value in a distribution is the mode. It describes a whole set of data with one value which is the center of the distribution.

Dispersion of distribution is the describe as to how much spread out a set of data is. The measures of Dispersion are Variance, Standard Deviation and Inter-Quartile Range.

7. (**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

(a) p (X = female, Y= high risk) = ? (0.5 point)
Answer: 1/8

(b) p(X = female) = ? (0.5 point)
Answer: (1/6) + (1/6)+(1/8) = 11/24

(c) p (Y= high risk|X=female) =? (0.5 point)

Answer: p(Y|X) = p(X and Y)/p(X)
= (1/8)/ ((1/6)+(1/6)+(1/8))
= ((1/8)/(11/24))
=3/11

8. **(Total 4 points)**
Use R or Matlab to

(a) 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).

(b) Compute Expected Value, Variance, Standard Deviation, Mode, Median, Skewness and Kurtosis for x1 and x2 generated for Question (a).

(c) 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]

Answer: Code is Attached

(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

(a) 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]

REFERENCES

1. https://www.molgen.mpg.de/3659531/MITPress--SemiSupervised-Learning.pdf
2. Machine Learning: A Probabilistic Perspective. (2012) Kevin P. Murphy. The MIT press. Cambridge, Massachusetts
3. An Introduction to Statistical Learning, (2013) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013