Brief Homework 3
Computer Science
Fall 2016
B565

Jivitesh Poojary

October 4, 2016

Declaration

All the work done in this assignment is my and my alone. The *R* Programs have been coded and complied in *RStudio*. I have made use of *knitr* package of *RStudio* to embed the programs in the LaTeX file.

Questions

1. In R load the iris data. Issue the following commands at the prompt:

```
> data(iris)
> pairs(iris[1:4],main="Iris Data", pch=19, col=as.numeric(iris$Species)+1)
> iris.pca <- prcomp(log(iris[,1:4]), scale=TRUE, center=TRUE)
> summary(iris.pca)
> iris.pca
> screeplot(iris.pca, type="lines", col="3")
> iris.pca$sdev^2
```

   (a) What species is red, blue, and green?
   Answer - The 'setosa' species is red in color, followed by the 'versicolor' species in green color and finally the 'virginica' species in blue color. We arrive at this conclusion by seeing the range of values in the scatter plot for each color and cross checking the label in the dataset for those range of values.

   (b) What are the principle components?
   Answer - PC1, PC2, PC3 and PC4 are the principle components. Each principle component gives us a clue about the spacial variance of our data and the direction with the most variation.

   (c) There are three methods to pick the set of principle components: (1) In the plot where the curve *bends*; (2) Add the percentage variance until total 75% is reached (70-90%) (3) Use the components whose variance is at least one. Show the components selected in the iris data if each of these is used.
   Answer -

      i. The curve bends at the PC2 and PC3, so in this case we shall be considering PC1, PC2 and PC3 as it gives us maximum area under the curve.
      ii. Add the percentage variance: we can look at the cummulative proportion of the variance in the output. For PC1 we have the value as 0.7331, however the value increases to 0.9599 when we include PC2 in the calculation. Hence in this case we shall be considering PC1 and PC2.
      iii. Use the components whose variance is at least one: From the output we can see that only for PC1 has the standard deviation greater than 1 (1.7125). As a result the calculated variance will be greater than 1 too. So we shall be only selecting PC1in this case.

   (d) What are the first two unit eigenvectors? What does *unit* mean?
   Answer -

i. The first two unit eigenvectors are:
PC1 (0.5038236, -0.3023682 , 0.5767881, 0.5674952)
PC2 (-0.45499872 ,-0.88914419 , -0.03378802 , -0.03545628).

ii. In the above code we have used the '*scale*' attribute of the '*prcomp*' function to get unit eigenvector values. The value of '*scale*' is set to $TRUE$.

iii. *Unit* eigenvectors are the value of eigenvectors divided by thier magnitude.

(e) How does the PCA differ when no log transform is done of the data?
Answer -

i. We apply the log transformation to reduce the skewness of the data, so that the data is scaled to a more simpler version.

ii. Log transform is necessary as we can condense the range of the values that are present. It has minimum effect on outliers and we can place all the points in a single plot without much deviation.

iii. When we do not perform the log transformation on above data while calculating the components, we do not see a lot of difference in our curve. There is a slight bend in our curve and the values on the axis change proportionaly. We can infer this from the small domain of the values present in the attributes.

2. Assume $X = \{(a, 2), (b, 2), (c, 16)\}, Y = \{(a, 4), (b, 1), (c, 15)\}, Z = \{(a, 10), (b, 4), (c, 4)\}$ are multisets. Calculate the entropy on each. Does is make sense to compare entropy between X, Y, X, Z, and Y, Z? Assume you have another multiset W and its entropy is 0. How many elements does it have? Assume its entropy is the same as X. How many elements does it have? What's the primary difference between multisets and probability distributions?
Answer -

Calculating the entropy of the X: $\mathcal{H}(\Delta_{[X]}) = -\frac{2}{20}log(\frac{2}{20}) - \frac{2}{20}log(\frac{2}{20}) - \frac{16}{20}log(\frac{16}{20})$
$\mathcal{H}(\Delta_{[X]}) = 0.3321928 + 0.3321928 + 0.2575425$
$\mathcal{H}(\Delta_{[X]}) = 0.9219281$

Calculating the entropy of the Y: $\mathcal{H}(\Delta_{[Y]}) = -\frac{4}{20}log(\frac{4}{20}) - \frac{1}{20}log(\frac{1}{20}) - \frac{15}{20}log(\frac{15}{20})$
$\mathcal{H}(\Delta_{[Y]}) = 0.4643856 + 0.2160964 + 0.3112781$
$\mathcal{H}(\Delta_{[Y]}) = 0.9917601$

Calculating the entropy of the Z: $\mathcal{H}(\Delta_{[Z]}) = -\frac{10}{18}log(\frac{10}{18}) - \frac{4}{18}log(\frac{4}{18}) - \frac{4}{18}log(\frac{4}{18})$
$\mathcal{H}(\Delta_{[Z]}) = 0.4711094 + 0.4822056 + 0.4822056$
$\mathcal{H}(\Delta_{[Z]}) = 1.435521$

Yes, we should be comparing the entropies of X, Y. However it does not make sense to compute entropy for X, Z, and Y, Z as these multisets are unequal in cardinality.

If we have another multiset W and its entropy is 0, then it will be having only one member. The number of elements belonging to the member in the multiset should be greater than or equal to 1.

If for multiset W the entropy is same as X, then W will be having the same members . Different multisets can have same entropies for different number of elements. This is because the formula for calculating the entropy only considers the ratio of the occurance of the elements in the multiset. One way of considering the number of elements in W would be to have the multiset be equal to X. Another way is to generalize the above statement as the members be having the same ratio in both the mulisets.

However, this may not be true always as, if the member ratios are interchanged we get the same value of entropy. So a better solution is that each member of W should have elements in a ratio which corresspends to a member in X, but this relationship has to be unique pairwise.

The primary difference between multisets and probability distribution is that we can apply probability distribution on continous data however this is not possible in mulisets directly we have to discretize the data before using it here. Probability distribution gives us an idea of the proportion of the value with respect to the entire

set, but muliset gives us the occurance of the value and we have to calculate the proportion. Nominal and Ordinal data can be represented using either probability distribution or muliset.

3. Consider the following data:

$$\Delta = \{((a,a,a),5),((b,b,a),10),((a,a,b),5),((b,b,a),10)\} \quad (1)$$

Assume we name the attributes A_1, A_2, A_3 and label L. Create a relation instance **r** that describes Δ. Show explicitly which attribute is best to split in using Information Gain. Split on that attribute, then show which attributes are best again. Write the tree as a rule set. What is the label of (a,b,a)?

Answer -

Lets create a relation instance **r** that describes Δ:

A_1	A_2	A_3	L
a	a	a	5
b	b	a	10
a	a	b	5
b	b	a	10

Where $dom(A_1) = dom(A_2) = dom(A_3) = \{a,b\}$ and label $L = \{5,10\}$.
From the table we can see that the attributes A_1 and A_2 are similar.
Calculating the entropy of the Label (L):

$\mathcal{H}(\Delta_{[L]}) = -\frac{2}{4}log(\frac{2}{4}) - \frac{2}{4}log(\frac{2}{4})$
$\mathcal{H}(\Delta_{[L]}) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2})$
$\mathcal{H}(\Delta_{[L]}) = 1$

The information gain after splitting on A_1 is:
Calculating the entropy of the A_1 when the value is 'a' :
$\mathcal{H}(\Delta_{[A_1=a]}) = -\frac{2}{2}log(\frac{2}{2}) - \frac{0}{0}log(\frac{0}{0})$
$\mathcal{H}(\Delta_{[A_1=a]}) = 0$
Calculating the entropy of the A_1 when the value is 'b' :
$\mathcal{H}(\Delta_{[A_1=b]}) = -\frac{2}{2}log(\frac{2}{2}) - \frac{0}{0}log(\frac{0}{0})$
$\mathcal{H}(\Delta_{[A_1=b]}) = 0$

$\mathcal{I}(\Delta_{[A_1]}) = \mathcal{H}(\Delta_{[L]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_1=a]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_1=b]})$
$\mathcal{I}(\Delta_{[A_1]}) = 1$ - 0 - 0 = 1

The information gain after splitting on A_2 is:
Calculating the entropy of the A_2:
Calculating the entropy of the A_2 when the value is 'a' :
$\mathcal{H}(\Delta_{[A_2=a]}) = -\frac{2}{2}log(\frac{2}{2}) - \frac{0}{0}log(\frac{0}{0})$
$\mathcal{H}(\Delta_{[A_2=a]}) = 0$
Calculating the entropy of the A_2 when the value is 'b' :
$\mathcal{H}(\Delta_{[A_2=b]}) = -\frac{2}{2}log(\frac{0}{0}) - \frac{2}{2}log(\frac{0}{0})$
$\mathcal{H}(\Delta_{[A_2=b]}) = 0$

$\mathcal{I}(\Delta_{[A_2]}) = \mathcal{H}(\Delta_{[L]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_2=a]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_2=b]})$
$\mathcal{I}(\Delta_{[A_2]}) = 1$ - 0 - 0 = 1

The information gain after splitting on A_3 is:
Calculating the entropy of the A_3:
Calculating the entropy of the A_3 when the value is 'a' :
$\mathcal{H}(\Delta_{[A_3=a]}) = -\frac{2}{3}log(\frac{2}{3}) - \frac{1}{3}log(\frac{1}{3})$
$\mathcal{H}(\Delta_{[A_3=a]}) = 0.389975 + 0.5283208$
$\mathcal{H}(\Delta_{[A_3]})$ =0.9182958
Calculating the entropy of the A_2 when the value is 'b' :
$\mathcal{H}(\Delta_{[A_3=b]}) = -\frac{1}{1}log(\frac{1}{1}) - \frac{0}{0}log(\frac{0}{0})$

$\mathcal{H}(\Delta_{[A_3=b]}) = 0$

$\mathcal{I}(\Delta_{[A_3]}) = \mathcal{H}(\Delta_{[L]})$ - $\frac{3}{4}\mathcal{H}(\Delta_{[A_3=a]})$ - $\frac{1}{4}\mathcal{H}(\Delta_{[A_3=b]})$
$\mathcal{I}(\Delta_{[A_3]}) = 1$ - 0.6887218 - 0 = 0.3112782

Ruleset - Considering $a' = b$ and $b' = a$,

Ruleset
$A_1 = a \rightarrow L = 5$ $A_1 = b \rightarrow L = 10$ $A_2 = a \rightarrow L = 5$ $A_2 = b \rightarrow L = 10$ $A_3 = b \rightarrow L = 5$
$(A_1 = a) \wedge (A_3 = b) \rightarrow L = 5$ $(A_2 = a) \wedge (A_3 = b) \rightarrow L = 5$ $(A_1 = a) \vee (A_1 = b') \rightarrow L = 5$ $(A_1 = b) \vee (A_1 = a') \rightarrow L = 10$ $(A_2 = a) \vee (A_2 = b') \rightarrow L = 5$ $(A_2 = b) \vee (A_2 = a') \rightarrow L = 10$ $(A_1 = a) \vee (A_2 = a) \rightarrow L = 5$ $(A_1 = b) \vee (A_2 = b) \rightarrow L = 10$
$((A_1 = b) \wedge (A_2 = b)) \wedge (A_3 = a) \rightarrow L = 10$ $((A_1 = a) \wedge (A_2 = a)) \wedge (A_3 = b) \rightarrow L = 5$ $((A_1 = a) \wedge (A_2 = a)) \wedge (A_3 = a) \rightarrow L = 10$

Tree - Attributes A_1 and A_2 have the highest values of Information Gain (1). Hence we can choose either node for building the decision tree. The ruleset present in the final block are overfitting the decision tree. If we get a different value, we may get an error.

Label - The label of (a, b, a) will depend on the ruleset that we choose, here $A_1 = a, A_2 = band A_3 = a$

(a) If we choose, $A_1 = a \rightarrow L = 5$ we get the answer as 5

(b) If we choose, $A_2 = b \rightarrow L = 10$ we get the answer as 10

4. ID3/C4.5 is a greedy algorithm. What is the implication for the kind of trees ID3 produces?
Answer -

(a) ID3/C4.5 is a greedy algorithm, the reason we say this is because the algorithms look for a local minima at each step of iteration. The entropy function is calculated locally with available attributes.

(b) In ID3/C4.5 we do not have an option of backtracking and changing the intermediate nodes once they have been constructed. The intermediate node selection may have been an optimal decision at that the time we were selecting the best attribute for tree building. However, we can get a better decision tree which requires fewer nodes for prediction and is less overfitted.

(c) The implication is that we may not get a best solution with this approach of decision tree construction.

(d) Another point to be considered here is that because of the greedy nature of tha algorithm there is an indusctive bias towards creation of shorter trees and collection of higher information gain nodes near the root of the tree.

5. (Pick One) Error in the training set is usually **more** or **less** than the true error of the real function?
Answer - Less

(a) The error in the training set can be **more** or **less** than the true error of the real function depending on the decision tree that is constructed and the number of nodes that are part of the decision tree.

(b) When the nodes in the decision tree are few, the error rate on training and test data is very high, the reason attributed to this is underfitting of the model. This mean the nodes are so few that they are yet to learn the true structure of the data.

(c) We can reduce the error on the training data set if we were to take in all the combinations of available data and create nodes in the decision tree that satisfy this condition. This approach will reduce the error but will increase the complexity of our model.

(d) However, we are just memorizing the training data in our model. This data may not be a true representation of the data available in the unseen real data present in the test set. Thus resulting in higher error rates when the we test our model on this data. This is called as overfitting. We can resolve this issue by having an acceptable error present in our training data and prunning the decision tree after completely building it.

6. Assume you have the dissimilarity matrix Q of data A,B,C,D.

	A	B	C	D
A	0			
B	1	0		
C	4	8	0	
D	5	2	2	0

Using this ultrametric distance:

$$Q(x,(yz))) = \min\{Q(x,y),Q(x,z)\}$$

Show a tree that results when using hierarchical alggomerative clustering. Show how Q changes at each iteration.

Answer -

From the above table we can see that the distance between A and B is the least, if we were to combine these clusters first we will be getting the following table:

$$\begin{aligned} Q((A,B),C) &= \min\{Q(A,C),Q(B,C)\} \\ &= \min\{4,8\} \\ &= \{4\} \end{aligned}$$

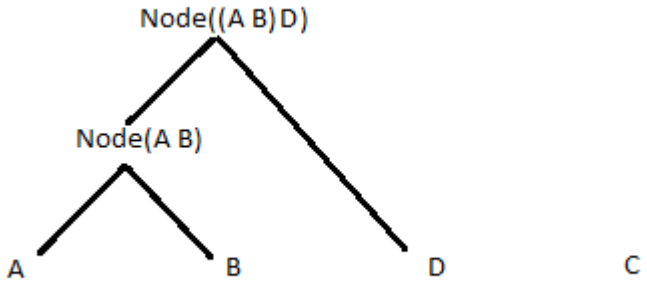
$$\begin{aligned} Q((A,B),D) &= \min\{Q(A,D),Q(B,D)\} \\ &= \min\{5,2\} \\ &= \{2\} \end{aligned}$$

	(A B)	C	D
(A B)	0		
C	4	0	
D	2	2	0

From the above table we can see that the distance between (A B) and D, also between C and D is the least. In our calculation we will be merging the cluster (A B) with D first .If we were to combine these clusters first we will be getting the following table:

$$\begin{aligned} Q(((AB)D),C) &= \min\{Q((AB),C),Q(D,C)\} \\ &= \min\{4,2\} \\ &= \{2\} \end{aligned}$$

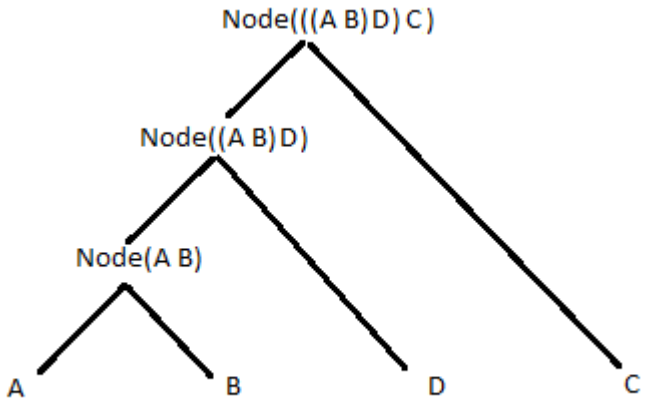
	((A B) D)	C
((A B) D)	0	
C	2	0



Final representation of our dissimilarity table:

	(((A B) D) C)
(((A B) D) C)	0

We can draw a tree as follows:



7. (True or False) For clustering that has less than 1000 data points, enumeration of all possible clusters is feasible.
Answer - False

   (a) For clustering that has than 1000 data points, enumeration of all possible clusters is possible however this is not a feasible process

   (b) In this case the worst case scenario is when we are having a cluster associated with each data point. Thereby increasing the complexity of the process.

   (c) Similarly enumeration of all possible clusters increases the runtime of the process exponentially, thus making it infeasible.

8. (True or False) The run-time of agglomerative clustering is $O(n^2)$.
Answer - False

(a) In agglomerative clustering we perform a linear search for building the proximity matrix.

(b) The run-time of agglomerative clustering is $O(n^2.log(n))$ where n is the number of data points. We get result if the distance between clusters are stored in a sorted list (or heap).

(c) However, if we didn't store the distances in a sorted manner the complexity would increase to $O(n^3)$ as at each instance the algorithm would be searching in the proximity matrix for the best possible point or cluster.

9. Assume a data set Δ for classification has two attributes A_1, A_2 that are binary, $dom(A_1) = dom(A_2) = \{0, 1\}$ and label $L = \{0, 1\}$ with a binary outcome. The information gain for *both* A_1, A_2 are identical and $\mathcal{H}(\Delta_{[L]}) = k$ for $k > 0$. Fill-in the entries below that satisfy these constraints.

A_1	A_2	L	Count
⋮	⋮	⋮	⋮

Answer -
The domain of attributes A_1 , A_2 and L is $\{0, 1\}$. Also the information gain for *both* A_1, A_2 are identical and $\mathcal{H}(\Delta_{[L]}) = k$ for $k > 0$.

The entries that satisfy these constraints are, for $n \geq 1$:

A_1	A_2	L	Count
0	0	0	n
0	1	1	n
1	0	1	n
1	1	0	n

Calculating the entropy of the Label (L):
$\mathcal{H}(\Delta_{[L]}) = -\frac{2}{4}log(\frac{2}{4}) - \frac{2}{4}log(\frac{2}{4})$
$\mathcal{H}(\Delta_{[L]}) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2})$
$\mathcal{H}(\Delta_{[L]}) = 1$

Calculating the entropy of the A_1:
Calculating the entropy of the A_1 when the value is '0' :
$\mathcal{H}(\Delta_{[A_1=0]}) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2})$
$\mathcal{H}(\Delta_{[A_1=0]}) = 1$
Calculating the entropy of the A_1 when the value is '1' :
$\mathcal{H}(\Delta_{[A_1=1]}) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2})$
$\mathcal{H}(\Delta_{[A_1=1]}) = 1$

$\mathcal{I}(\Delta_{[A_1]}) = \mathcal{H}(\Delta_{[L]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_1=0]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_1=1]})$
$\mathcal{I}(\Delta_{[A_1]}) = 1$ - $\frac{1}{2}$ - $\frac{1}{2} = 0$

Calculating the entropy of the A_2:
Calculating the entropy of the A_2 when the value is '0' :
$\mathcal{H}(\Delta_{[A_2=0]}) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2})$
$\mathcal{H}(\Delta_{[A_2=0]}) = 1$
Calculating the entropy of the A_2 when the value is '1' :
$\mathcal{H}(\Delta_{[A_2=1]}) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2})$
$\mathcal{H}(\Delta_{[A_2=1]}) = 1$

$\mathcal{I}(\Delta_{[A_2]}) = \mathcal{H}(\Delta_{[L]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_2=0]})$ - $\frac{2}{4}\mathcal{H}(\Delta_{[A_2=1]})$
$\mathcal{I}(\Delta_{[A_2]}) = 1$ - $\frac{1}{2}$ - $\frac{1}{2} = 0$

From the above calculation we can see that, $\mathcal{H}(\Delta_{[L]}) = 1$ for $1 > 0$ thus satisfying the condition. Similary the information gain for *both* A_1, A_2 are identical and equal to 0.

We also get an alternate solution where we have to invert the label values:

A_1	A_2	L	Count
0	0	1	n
0	1	0	n
1	0	0	n
1	1	1	n

10. Read Quinlan's, "Induction of Decision Trees," Machine Learning 1:81-106 (1986), Kluwer Academic Publishers. In no more than four paragraphs, discuss what he writes about Noise and Unknown Attribute Values for ID3. How will this impact your use of ID3/C4.5?
Answer -
Noise -

    (a) A correct decision tree for this corrupted training set would now have to explain the apparent special cases. Two problems: errors in the training set may cause the attributes to become inadequate, or may lead to decision trees of spurious complexity. Non-systematic errors of this kind in either the values of attributes or class information are usually referred to as noise.

    (b) One solution to this dilemma might be to require that the information gain of any tested attribute exceeds some absolute or percentage threshold. However, this method has some shortfalls and the performance of the tree-building procedure is degraded in the noise-free case. An alternative method based on the chi-square test for stochastic independence has been found to be more useful. For higher noise levels, the performance of the correct decision tree on corrupted data was found to be inferior to that of an imperfect decision tree formed from data corrupted to a similar level.

    Unknown Attribute Values -

    (a) Some of the methods for computing the known attribute values include: placing the most common value, creating a probability distribution whereby we can get the most likely value or divide the objects into fractional objects. Another method is to treat the missing attribute value as a class variable and the class variable as an attribute and classify the value on this basis.

    (b) Another method for handling unknown values was to use token values there by creating multiple branches at the leaf node level, summation of the value and classifying according to the highest occurrance for each class. All the methods discussed have their set of assumptions and do not offer optimal solutions when a relatively large proportion of data is missing.

    Impact -

    (a) We can see that if we were to consider some of the methods discussed in our ID3/C4.5 algorithm we shall get a better decision tree. Along with this there will be increase in accuracy and lower overfitting with better generalization.

11. Let $\Omega = \mathcal{Z}$. We define a random variable X as:

$$p_X(x) \quad = \quad \begin{cases} 1/9, \ -4 \leq x \leq 4 \\ 0, \ o.w. \end{cases} \tag{2}$$

    (a) Calculate E[X]
    Answer -
    E[X] = $\int_{-\infty}^{\infty} x.p_X(x)dx$
    E[X] = $\int_{-\infty}^{-4} x.p_X(x) + \int_{-4}^{4} x.p_X(x)dx + \int_{4}^{\infty} x.p_X(x)dx$
    E[X] = $0 + \int_{-4}^{4} x.p_X(x)dx + 0$ ..... according to the definition of $p_X(x)$
    E[X] = $\int_{-4}^{4} x.p_X(x)dx$
    E[X] = $\frac{1}{9}[\frac{1}{2}x^2]$
    E[X] = $\frac{1}{9}[\frac{1}{2}(4)^2 - \frac{1}{2}(-4)^2]$
    E[X] = 0

(b) Let $Y = |X|$. Calculate E[Y]

Answer -

$\mathrm{E}[X] = \int_{-\infty}^{\infty} |x.p_X(x)|dx$

$\mathrm{E}[X] = \int_{-\infty}^{-4} |x.p_X(x)| + \int_{-4}^{4} |x.p_X(x)|dx + \int_{4}^{\infty} |x.p_X(x)|dx$

$\mathrm{E}[X] = 0 + \int_{-4}^{4} |x.p_X(x)|dx + 0$ ..... according to the definition of $p_X(x)$

$\mathrm{E}[X] = \int_{-4}^{4} |x.p_X(x)|dx$

$\mathrm{E}[X] = 2\int_{0}^{4} |x.p_X(x)|dx$ ..... $p_X(x)$ is a constant funtion in the interval

(We are integrating over a modulus function, it will have equal value for both positive and negative values)

$\mathrm{E}[X] = 2.\frac{1}{9}[\frac{1}{2}x^2]$

$\mathrm{E}[X] = 2.\frac{1}{9}[\frac{1}{2}(4)^2 - \frac{1}{2}(0)^2]$

$\mathrm{E}[X] = \frac{16}{9}$

(c) Calculate Var[X]

Answer -

$\mu = \mathrm{E}[X] = 0$ .....Mean of the distribution

$\mathrm{Var}[X] = \int_{-\infty}^{\infty}(x-\mu)^2.p_X(x)dx$

$\mathrm{Var}[X] = \int_{-\infty}^{-4}(x-\mu)^2.p_X(x) + \int_{-4}^{4}(x-\mu)^2.p_X(x)dx + \int_{4}^{\infty}(x-\mu)^2.p_X(x)dx$

$\mathrm{Var}[X] = 0 + \int_{-4}^{4}(x-\mu)^2.p_X(x)dx + 0$ ..... according to the definition of $p_X(x)$

$\mathrm{Var}[X] = \int_{-4}^{4}(x-\mu)^2.p_X(x)dx$

$\mathrm{Var}[X] = \frac{1}{9}[\frac{1}{3}(x-\mu)^3]$

$\mathrm{Var}[X] = \frac{1}{9}[\frac{1}{3}(4-0)^3 - \frac{1}{3}(-4-0)^3]$

$\mathrm{Var}[X] = \frac{1}{9}\frac{1}{3}[(4)^3 - (-4)^3]$

$\mathrm{Var}[X] = \frac{1}{27}[64+64]$

$\mathrm{Var}[X] = \frac{128}{27}$

12. Let Ω be the throw of a fair die twice. Let A be the event a 3 on either individual throw was observed or the sum is at least 5. Let B be the event that the difference between the two throws is exactly one. Are A, B independent? What is the probability of B given A? What is the probability of A given B?

Answer -

Probability of 3 appearing on a single throw of a die is $\frac{6}{36}$

The possible combinations become - {(3,1), (3,2), (3,4), (3,5), (3,6), (1,3), (2,3), (4,3), (5,3), (6,3), (3,3)}

P(3 on either individual throw of die) = $\frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$

When computing the possibility of the sum being atleast 5, we consider the sum being 5,6,7,8,9,10,11 or 12

We can also look at the problem as the sum not being 2,3 or 4 .The possible combinations become - {(1,1), (1,2), (2,1), (2,2), (3,1),(1,3)}

P(the sum not being at least 5) = $\frac{6}{36}$

P(the sum is at least 5) = 1 - $\frac{6}{36} = \frac{30}{36}$

P(A) = P(3 on either individual throw of die) + P(the sum is at least 5) - P(3 on either individual throw of die AND the sum is at least 5)

P(A) = $\frac{11}{36}$ + $\frac{30}{36}$ - $\frac{9}{36}$

P(A) = $\frac{32}{36}$

For getting the difference as 1, we consider the following possible combinations:

{(1,2), (2,1), (2,3), (3,2), (3,4), (4,3), (4,5), (5,4), (5,6), (6,5)}

P(B) = $\frac{10}{36}$

If A and B are independent events, then $P(A \cap B) = P(A).P(B)$

$P(A \cap B) = \frac{8}{36}$

$P(A).P(B) = \frac{32}{36}.\frac{10}{36} = \frac{320}{1296}$

$P(A \cap B) \neq P(A).P(B)$, A and B are not independent events

Probability of B given A:
$P(B|A) = \frac{P(A \cap B)}{P(A)}$
$P(B|A) = \frac{\frac{8}{36}}{\frac{32}{36}} = \frac{8}{32}$
$P(B|A) = \frac{1}{4}$

Probability of A given B $P(A|B) = \frac{P(A \cap B)}{P(B)}$
$P(A|B) = \frac{\frac{8}{36}}{\frac{10}{36}} = \frac{8}{10}$
$P(A|B) = \frac{4}{5}$

13. Consider the a random variable X with probability distribution:

$$f_X(x) \quad = \quad \begin{cases} \frac{1}{2}x^{-\frac{1}{2}}, \ 0 < x \leq 1 \\ 0, \ o.w. \end{cases} \tag{3}$$

(a) Show f_X is a probability distribution
Answer -
$\text{PD}[X] = \int_{-\infty}^{\infty} f_X(x)dx$
$\text{PD}[X] = \int_{-\infty}^{0} f_X(x) + \int_0^1 f_X(x)dx + \int_1^{\infty} f_X(x)dx$
$\text{PD}[X] = 0 + \int_0^1 f_X(x)dx + 0$ ..... according to the definition of $f_X(x)$
$\text{PD}[X] = \int_0^1 f_X(x)dx$
$\text{PD}[X] = \int_0^1 \frac{1}{2}x^{-\frac{1}{2}}dx$
$\text{PD}[X] = \frac{1}{2}\int_0^1 x^{-\frac{1}{2}}dx$
$\text{PD}[X] = \frac{1}{2}.\frac{2}{1}[x^{\frac{1}{2}}]$
$\text{PD}[X] = [\frac{1}{3}x^{\frac{3}{2}}]$
$\text{PD}[X] = [(1)^{\frac{3}{2}} - (0)^{\frac{3}{2}}]$
$\text{PD}[X] = 1$
Since the area below the curve for this distribution is 1 we can show it is a probability distribution.

(b) Calculate $\text{E}[X]$
Answer -
$\text{E}[X] = \int_{-\infty}^{\infty} x.f_X(x)dx$
$\text{E}[X] = \int_{-\infty}^{0} x.f_X(x) + \int_0^1 x.f_X(x)dx + \int_1^{\infty} x.f_X(x)dx$
$\text{E}[X] = 0 + \int_0^1 x.f_X(x)dx + 0$ ..... according to the definition of $f_X(x)$
$\text{E}[X] = \int_0^1 x.f_X(x)dx$
$\text{E}[X] = \int_0^1 x.\frac{1}{2}x^{-\frac{1}{2}}dx$
$\text{E}[X] = \frac{1}{2}\int_0^1 x^{\frac{1}{2}}dx$
$\text{E}[X] = [\frac{1}{2}\frac{2}{3}x^{\frac{3}{2}}]$
$\text{E}[X] = [\frac{1}{3}x^{\frac{3}{2}}]$
$\text{E}[X] = [\frac{1}{3}(1)^{\frac{3}{2}} - \frac{1}{3}(0)^{\frac{3}{2}}]$
$\text{E}[X] = \frac{1}{3}$

(c) Calculate $\text{Var}[X]$
Answer -
$\text{Var}[X] = \int_{-\infty}^{\infty}(x-\mu)^2.f_X(x)dx$
$\text{Var}[X] = \int_{-\infty}^{0}(x-\mu)^2.f_X(x) + \int_0^1(x-\mu)^2.f_X(x)dx + \int_1^{\infty}(x-\mu)^2.f_X(x)dx$
$\text{Var}[X] = 0 + \int_0^1(x-\mu)^2.f_X(x)dx + 0$ ..... according to the definition of $f_X(x)$
$\text{Var}[X] = \int_0^1(x-\mu)^2.f_X(x)dx$
$\text{Var}[X] = \int_0^1(x-\mu)^2.\frac{1}{2}x^{-\frac{1}{2}}dx$
$\text{Var}[X] = \int_0^1(x-\frac{1}{3})^2.\frac{1}{2}x^{-\frac{1}{2}}dx$ .....substituting μ
$\text{Var}[X] = \frac{1}{2}\int_0^1(x^2 - \frac{2}{3}x + \frac{1}{9}).x^{\frac{-1}{2}}dx$
$\text{Var}[X] = \frac{1}{2}\int_0^1 x^{\frac{3}{2}} - \frac{2}{3}x^{\frac{1}{2}} + \frac{1}{9}.x^{\frac{-1}{2}}dx$

$$\text{Var}[X] = [\tfrac{1}{2}(\tfrac{2}{5}.x^{\frac{5}{2}} - \tfrac{4}{9}x^{\frac{3}{2}} + \tfrac{2}{9}.x^{\frac{1}{2}})]$$
$$\text{Var}[X] = [\tfrac{1}{2}(\tfrac{2}{5}.x^{\frac{5}{2}} - \tfrac{4}{9}x^{\frac{3}{2}} + \tfrac{2}{9}.x^{\frac{1}{2}})]$$
$$\text{Var}[X] = [\tfrac{1}{2}(\tfrac{2}{5}.(1)^{\frac{5}{2}} - \tfrac{4}{9}.(1)^{\frac{3}{2}} + \tfrac{2}{9}.(1)^{\frac{1}{2}}) - 0]$$
$$\text{Var}[X] = \tfrac{1}{2}.\tfrac{8}{45}$$
$$\text{Var}[X] = \tfrac{4}{45}$$

What to Turn-in

The *pdf of the written answers to this document.