Brief Homework 3
Computer Science
Fall 2016
B565

Anusha Ramamurthy
anuramam@iu.edu

October 13, 2016

All work herein is solely mine.

Directions

Please follow the syllabus guidelines in turning in your homework. I will provide the LaTeX of this document too. You may use it or create one of your own. This homework should be started quickly. Sometimes there are natural questions arising from code. Within a week, AIs can contact students to examine code; students must meet within three days. The session will last no longer than 5 minutes. If the code does not work, the grade for the program may be reduced. Lastly, source code cannot be modified post due date.

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 function $col = (as.numeric(iris\$Species) + 1)$ assigns numerical values to the species. Hence Red is the species setosa, Green is the species versicolor and Blue is the species virginica.

   (b) What are the principal components?
   Answer: We have four Principal components, one corresponding to each attribute. The number of principal components is less than or equal to the number of attributes.

```
> iris.pca <- prcomp(log(iris[,1:4]), scale=TRUE, center=TRUE)
> iris.pca

Standard deviations:
[1] 1.7124583 0.9523797 0.3647029 0.1656840

Rotation:
                PC1         PC2        PC3         PC4
Sepal.Length  0.5038236 -0.45499872  0.7088547  0.19147575
```

```
Sepal.Width  -0.3023682 -0.88914419 -0.3311628 -0.09125405
Petal.Length  0.5767881 -0.03378802 -0.2192793 -0.78618732
Petal.Width   0.5674952 -0.03545628 -0.5829003  0.58044745
```

(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:

Method 1: We generate the screeplot to see where the curve bends. We see that the curve bends at Principal component number 2 and 3. Hence we pick PC1, PC2 and PC3 in this method.

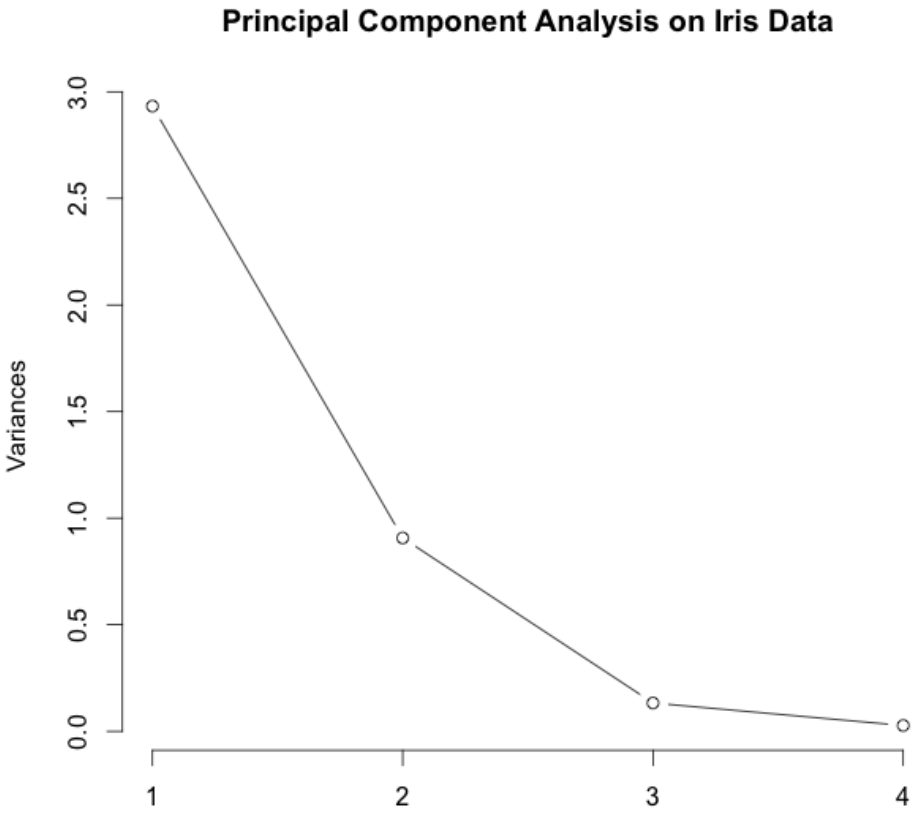

Figure 1: A screeplot plots the variances against the number of the principal component

Method 2: Using the summary function of R, we see that the according to the Proportion of Variance row, PC1 explains to almost 73% variance and PC2 explains 22% of the variance. Hence using this method, we pick PC1 and PC2 which together explain close to 95% of the variance.

```
> summary(iris.pca)
Importance of components:

                          PC1    PC2     PC3     PC4
Standard deviation     1.7125 0.9524 0.36470 0.16568
Proportion of Variance 0.7331 0.2268 0.03325 0.00686
Cumulative Proportion  0.7331 0.9599 0.99314 1.00000
```

Method 3: We see that PC1 explains 0.73 of the variance. It is the only PC that is closest to 1 compared to the others. PC2 also contributes almost 0.22 towards the variance. Hence in this method, we pick PC1 and PC2 and get 0.956 which is ≈ 1

(d) What are the first two unit eigenvectors? What does *unit* mean?
Answer: The PC1 and PC2 are the first two unit eigenvectors. It is a good practice to constrain the eigenvectors to unit norm,in order to use the associated eigenvalue as the variance.

(e) How does the PCA differ when no log transform is done of the data?
Answer: PCA is not invariant or unaffected by the scale of the data. So if each of the features are on different scales, PCA will not give accurate results. Hence it is always a good practice to transform and scale the dataset before using PCA. However since the Iris dataset has features on the same scale, the PCA results for this dataset does not differ much from the results of the running PCA after log transformation on the data.

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: The entropy of X is defined as
$H(X) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i))$

$H(X) = -(\frac{2}{20} log_2 \frac{2}{20} + \frac{2}{20} log_2 \frac{2}{20} + \frac{16}{20} log_2 \frac{16}{20}) = 0.921$

$H(Y) = -(\frac{4}{20} log_2 \frac{4}{20} + \frac{1}{20} log_2 \frac{1}{20} + \frac{15}{20} log_2 \frac{15}{20}) = 0.992$

$H(Z) = -(\frac{10}{18} log_2 \frac{10}{18} + \frac{4}{18} log_2 \frac{4}{18} + \frac{4}{18} log_2 \frac{4}{18}) = 1.430$

Yes, it makes sense to compare entropy between sets because it can help us decide which path to take when we have different nodes/roots with two or more sets. We can pick the node with the largest entropy as that node will have the higher information and pick till we find a node with least information.
If we have another multiset W with entropy zero, it has just one element. If its entropy is same as X, it is not possible to determine the number of elements in it as it could so happen that different sets with same number of elements have different entropy values because of their distributions. Hence multisets can have different probability distributions with same number of elements or can have different number of elements with same probability distributions.

3. Consider the following data:

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

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:

	A1	A2	A3	L
1	a	a	a	5
2	b	b	a	10
3	a	a	b	5
4	b	b	a	10

$H(X) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i))$
The entropy of $H(A1) = 1.0$
The entropy of $H(A2) = 1.0$
The entropy of $H(A3) = 0.811$
The entropy of $H(L) = 1.0$

We calculate the information gain as
Gain $= H(Parent) -$ (weighted average) $* H(Children)$ Hence $Gain(A3) = 1 - \left\{\frac{3}{4}log\frac{3}{4} - \frac{1}{4} * 0\right\} = 0.311$
$Gain(A1) = 1$
$Gain(A2) = 1$
$Gain(L) = 1$
Thus we can split on A1 or A2. So we can pick either. Since the goal is to pick the attribute with the highest Information gain. Spliting on A1 , the label for (a, b, a) is 5 and Spliting on A2 the label for (a, b, a) is 10.

4. ID3/C4.5 is a greedy algorithm. What is the implication for the kind of trees ID3 produces?
Answer: ID3 being a greedy algorithm creates small trees but it also tends to over fit.

5. (Pick One) Error in the training set is usually **more** or **less** than the true error of the real function?
Answer: Error in Training set is generally more than the true error, but depending on the dataset and the classifier used it could change. Hence the answer depends and cannot be universally true.

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

Show a tree that results when using hierarchical agglomerative clustering. Show how Q changes at each iteration.
Answer: Using this ultrametric distance:

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

iteration1 In , we see that the distance from A to B is the least:

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

Iteration2

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

	AB	C	D
AB	0		
C	4	0	
D	2	2	0

Iteration3

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

	ABD	C
ABD	0	
C	2	0

7. (True or False) For clustering that has less than 1000 data points, enumeration of all possible clusters is feasible.
Answer: True. For clustering, large number of data points , enumeration of all possible points is computationally hard and but feasible.

8. (True or False) The run-time of agglomerative clustering is $O(n^2)$.
Answer: False.The run-time of agglomerative clustering is $O(n^3)$.The space complexity is $O(n^2)$. The time required to buid the proximity matrix is $O(n^2)$, where n is the number of data points.

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Answer:

A_1	A_2	L	Count
a	b	0	5
a	b	1	5
b	a	0	5
b	a	1	5

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: The paper discusses ID3 in detail and talks about the concept behind ID3, entropy , information gain and how these are used in ID3 for building the tree.
Noise: Noise in the tree can be if the values in our dataset are mislabelled or if objects of different classes have the same or identical descriptions. Thus the attributes are inadequate to make an informed decision or split.Hence there are two problems with this set. One, these false attribute values may cause the tree to be unnaturally complex and for attributes of the dataset being inadequate to explain the dataset fully. Hence the author suggests two modifications to deal with these problems:One, the algorithm has to be changed to handle these inadequate attributes. Two, the algorithm must be able to stop from building overly complex and spurious trees.The solution he suggests is to give a threshold to the information gain. This threshold can be absolute or a percentile.An alternate solution is to do the chi-square test for stochastic independence.
Unknown Attributes: This happens when we have missing values, or the dataset is incomplete.One work around the author suggests is to fill the values by a probabilistic model. That is we identify the most frequent value and identify the probability of the unknown value being one of the known values.Rather than treating unknown values as separate values, other method suggested is to access the information gain of an attribute A, with the idea that the unknown values are distributed in A in proportion to the relative frequency of these values in the entire dataset. The other question is how to deal with unknown values during classification. The author suggests passing a token with some value T along with the object that is being classified. These token values are summed for each class. Thus the distribution of values over all possible classes can be used to compute a higher degree of confidence levels.

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: We observe that X has a uniform distribution the values for $-4 \leq x \leq 4$. By definition of Uniform distributions $\mathrm{E}[X] = \frac{1}{2} * (a + b) = \frac{1}{2} * (-4 + 4) = 0$

(b) Let $Y = |X|$. Calculate E[Y]
Answer:$\mathrm{E}[Y] = \frac{1}{2} * (0 + 4) = 2$

(c) Calculate Var[X]
Answer: $\mathrm{Var}[X] = \frac{1}{12} * (4 - (-4))^2 = \frac{1}{12} * (8)^2 = 6.666$

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: The set

$$(A \cap B) = \{\{2,3\},\{3,2\},\{3,4\},\{4,3\},\{4,5\},\{5,4\},\{5,6\},\{6,5\}\} \tag{3}$$
$$\tag{4}$$

$P(A \cap B) = \frac{8}{36}$
$P(sum >= 5) = 1 - P(sum <= 4)$
$P(sum <= 4) = \frac{8}{36}$
$P(sum >= 5) = \frac{28}{36}$
$P(3) = \frac{5}{36}$
Hence $P(A) = \frac{33}{36} and P(B) = \frac{10}{36}$
For two events to be independent $P(A \cap B) = P(A) * P(B)$
However $P(A \cap B) = \frac{8}{36} = 0.222$ and $P(A) * P(B) = \frac{33}{36} * \frac{10}{36} = 0.255$ Hence A and B are not independent.
$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.222}{0.278} = 0.8$
$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.222}{0.916} = 0.242$

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

$$f_X(x) \quad = \quad \begin{cases} \frac{1}{2}x^{-\frac{1}{2}}, \ 0 < x \leq 1 \\ 0, \ o.w. \end{cases} \tag{5}$$

(a) Show f_X is a probability distribution

Answer:
For $f(x)$ to be probability distribution

$$\int_{-\infty}^{\infty} f(x).dx = 1$$

i.e.
$\int_{-\infty}^{0} f(x).dx + \int_{0}^{1} f(x).dx + \int_{1}^{\infty} f(x).dx = 0 + \int_{0}^{1} f(x).dx + 0$
$= \int_{0}^{1} f(x).dx = \frac{1}{2}\int_{0}^{1} x^{-\frac{1}{2}}.dx = \frac{1}{2}\frac{\left[x^{-\frac{1}{2}+1}\right]_0^1}{-\frac{1}{2}+1} = \frac{1}{2} * \frac{2}{1} * \left[x^{\frac{1}{2}}\right]_0^1 = \sqrt{1} = 1$
Hence $f_X(x)$ is a probability distribution.

(b) Calculate E[X]

Answer: $\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x).dx = \int_{0}^{1} xf(x).dx = \frac{1}{2}\int_{0}^{1} x.x^{-\frac{1}{2}}.dx = \frac{1}{2}\int_{0}^{1} x^{\frac{1}{2}}.dx = \frac{1}{2} * \frac{\left[x^{(\frac{1}{2}+1)}\right]_0^1}{(\frac{1}{2}+1)} = \frac{1}{2} * \left(\frac{2}{3}\right)\left[x^{\frac{3}{2}}\right]_0^1$
$= \frac{1}{3}\left[\sqrt[3]{x}\right]_0^1$

$\mathrm{E}[X] = \frac{1}{3}$

(c) Calculate $\mathrm{Var}[X]$
Answer: $\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$
$\mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x).dx = \int_0^1 x^2 f(x).dx = \frac{1}{2}\int_0^1 x^2.x^{-\frac{1}{2}}.dx = \frac{1}{2}\int_0^1 x^{\frac{3}{2}}.dx = \frac{1}{2} * \frac{\left[x^{(\frac{3}{2}+1)}\right]_0^1}{(\frac{3}{2}+1)} = \frac{1}{2} * \left(\frac{2}{5}\right)\left[x^{\frac{5}{2}}\right]_0^1$
$= \frac{1}{5}\left[\sqrt[5]{x}\right]_0^1 = \frac{1}{5}$
$\mathrm{Var}[X] = \frac{1}{5} - (\frac{1}{3})^2 = \frac{4}{45} = 0.08$

What to Turn-in

The *pdf of the written answers to this document.

References