CSC591/791: Spatial and Temporal Data Mining
Homework 3/4. Due: 4/9/15 @ 11.55pm EST.

Please **note** that part of h/w 4 is folded into this one and part will be folded into mid-term. Therefore, this h/w counts for **7**% of grading.

Student Name: Abhishek Kumar Agrawal
Student ID: 200061445
Unity ID: akagrawa

Notes:

- Most questions are open-ended, so please chose the best solution based on your reading/understanding.
- Some questions require reading papers and web resources. Likewise you can discuss with fellow students. Whatever the case, answers should always be your own.
- You should always **cite** the source of all materials if they are used in your answer/solution/program.
- Answers should be concise and to the point (and within the space provided).
- Word/PDF are the only accepted formats.

#Q	Max Points	Your Score
1	25	
2	25	
3	25	
4	25	

Q1. For given satellite image (ilk-3b-1024.tif; 3-dimensional, 1024x1024), and ground-truth (training/test) data. Ground-truth format is: <id,x, y, label>

1. **Do classification with at least one classifier from each group: Bayesian, Trees, Neural Networks, and SVMs; and compare (accuracy) and contrast (describe major differences, advantages/disadvantages). Be concise (use tabular format).**
2. **Do one spatial classification (any method and freely available s/w is fine). Compare the output with non-spatial (any classifier used to answer first part of the question). Submit classified images (tiff format) as separate zip file.**

For accuracy measure, submit full error matrix (aka, contingency table).

Solution :

Software : R (v 3.10)

Data Extraction : For the image classification, below steps are being followed:

a. The first step is to convert the image to a data.frame using rgdal package, and readGDAL function returns image features as pixel features (RGB bands and x, y coordinates).
b. Project the x,y coordinates in a frame of 1024 x 1024
c. Enhance the training set by taking 3 x 3 surrounding pixels of training dataset and labeling it with the same corresponding training pixel label.
d. Train model on RGB values of the pixel locations and evaluate and validate the model using test dataset.
e. Predict the label of all the image pixels and plot the output image

Below are models used for classification along with the contingency table.

1) **Using Decision Tree** : Package/Lib : rpart, function : rpart (Image : DecisionTree.tiff)

Contingency Matrix

	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	6	0	1	0	0
2	1	6	0	0	0
3	1	0	6	0	0
4	0	0	0	4	0
5	0	0	0	0	3

2) **Using Naive Bayes**: Package/Lib: e1071, function:naiveBayes (Image : NaiveBayes.tiff)

Contingency Matrix

	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	6	0	1	0	0
2	1	6	0	0	0
3	0	0	6	0	0
4	1	0	0	4	2
5	0	0	0	0	1

3) **Using Neural Network**: Package/Lib : nnet, function : nnet (Image : NeuralNetwork.tiff)

Contingency Matrix

	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	5	0	0	0	0
2	3	5	0	0	0
3	0	1	6	0	0
4	0	0	1	4	2
5	0	0	0	0	1

4) **Using Support Vector Machine** : Package/Lib : e1071, function : svm(gamma=0.5, cost=4) , (Image : SVM.tiff)

Contingency Matrix

	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	7	0	0	0	0
2	1	6	0	0	0
3	0	0	7	0	0
4	0	0	0	4	3
5	0	0	0	0	0

5) **K- Nearest Neighbour :** Package/Lib : class, function : knn (Image : KNN.tiff)

Contingency Matrix

	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	6	0	0	0	0
2	1	6	0	0	0
3	1	0	7	0	0
4	0	0	0	4	2
5	0	0	0	0	1

In the table below, the comparison of different non-spatial classifiers has been made with reference to accuracy measures and visualization of output image.

Classifier	Accuracy on Test Set	Avantage	Disadvantage
Decision Tree	89.28%	Among the other classifiers, it has the highest accuracy. And it has nicely classified some of the water class objects while other classifiers couldn't.	Model Overfitting. Since we are doing classification of the pixel on entire RGB band, the decision boundaries are complex and overfitting.
Naive Bayes	82.14%	By visualization, some class labels as road, house are accurately classified. The weight biasing can be added for certain low probability classes for more accuracy.	The output image has lots of misclassification among grass, trees and water labels as these labels tend to share overlapping RGB bands.
Neural Network	78.57%	The accuracy of the model can be increased by having more training samples along with more hidden layer configurations.	It is performing worst among the given classifiers. The default configuration of neural network is not able to assign appropriate weights for more accurate classification.
Support Vector Machine	85.71%	It has one of the highest accuracy in classification. And by visualization, it has correctly classified most of the road and building pixels. The model is also not much complex as number of Support vectors are less than 50% of the training samples.	The output image is highly accurate but a single class output of water is comparatively poor than other classifiers.
KNN	85.714%(k=17)	The classification results are comparable and it is able to detect most of the labeled objects.	Its a lazy learner, so the classification of heavy image with higher nearest neighbour (k) consideration are computationally extensive.

Q2. For the given dataset:

ID	Temperature	Outlook	Humidity	Y=Play{yes,no}
1	hot	sunny	high	no
2	cool	overcast	normal	no
3	mild	sunny	high	no
4	mild	overcast	high	no
5	hot	sunny	normal	yes
6	hot	rainy	high	yes
7	mild	rainy	high	yes
8	cool	rainy	normal	yes
9	cool	overcast	normal	no
10	cool	sunny	normal	yes
11	mild	rainy	normal	yes
12	mild	sunny	normal	yes
13	mild	rainy	high	yes
14	hot	rainy	normal	yes
15	hot	overcast	normal	yes

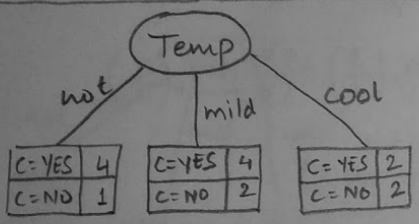
1. **Construct the decision tree using Entropy measure. Show complete work, and draw the final tree.**

   Entropy(t) = $- \sum_{i=0}^{c-1} p(i|t)\ log_2 p(i|t)$

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t)$$

C = Yes	10
C = No	5

$$\text{Entropy}(E) = -\frac{10}{15}\log_2\left(\frac{10}{15}\right) - \left(\frac{5}{15}\right)\log_2\left(\frac{5}{15}\right)$$
$$= 0.918725$$

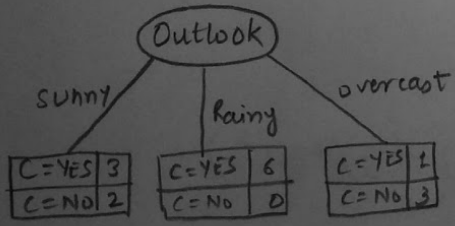


E(temp = hot)
$$= -\left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right)$$
$$= 0.7219$$

E(temp = mild)
$$= -\left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) - \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right)$$
$$= 0.91829$$

E(temp = cold)
$$= -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)$$
$$= 1$$

for temp Δinfo gain = E(Parent) − E(temp)
$$= 0.918725 - \left[\frac{5}{15}\times 0.7219 + \frac{6}{15}\times 0.91829 + \frac{4}{15}\times 1\right]$$
$$= 0.918725 - 0.874616$$
$$= 0.0441$$



E(outlook = sunny)
$$= -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)$$
$$= 0.97095$$

E(Outlook = Rainy)
$$= -\left(\frac{6}{6}\right)\log_2\left(\frac{6}{6}\right) - 0$$
$$= 0$$

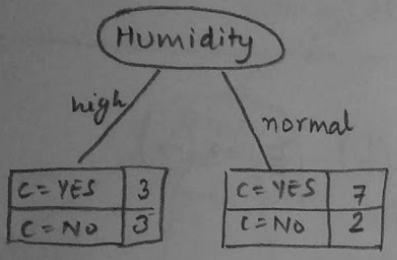
E(Outlook = overcast)
$$= -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right)$$
$$= 0.81128$$

Outlook Δinfo gain = E(Parent) − E(Outlook)

$= 0.918725 - \left[\left(\frac{5}{15}\right) \times 0.97095 + \left(\frac{1}{15}\right) \times 0 + \left(\frac{4}{15}\right) \times 0.811\right]$

$= 0.918725 - 0.53999$

$= 0.37873$



E(Humidity = high)

$= -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right)$

$= 1$

E(Humidity = normal)

$= -\left(\frac{7}{9}\right)\log_2\left(\frac{7}{9}\right) - \left(\frac{2}{9}\right)\log_2\left(\frac{2}{9}\right)$

$= 0.7642$

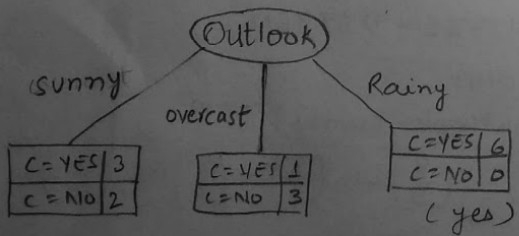
Humidity Δinfogain = E(Parent) − E(humidity)

$= 0.918725 - \left[\frac{6}{15} \times 1 + \frac{9}{15} \times 0.7642\right]$

$= 0.918725 - 0.85852$

$= 0.060205$

Δmfogain (Outlook) > Δinfo gain (Humidity) > Δmfogain (Temp)

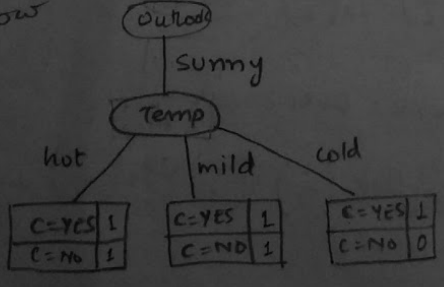


(yes)

Now, E(Outlook = sunny) $= -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)$

$= 0.97095$

Now



E(hot) $= -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)$

$= 1$

E(mild) $= -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)$

$= 1$

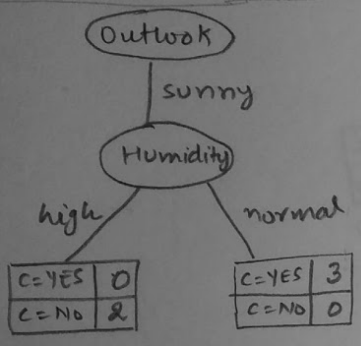
E(cold) $= -\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) - 0$

$= 0$

temp Δinfo gain = E(Parent) − E(temp)

$= 0.97095 - \left[\frac{2}{5} \times 1 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0\right]$

$= 0.17095$



E(Humidity = high)

$= -(0) - \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right)$

$= 0$

E(Humidity = normal) = 0

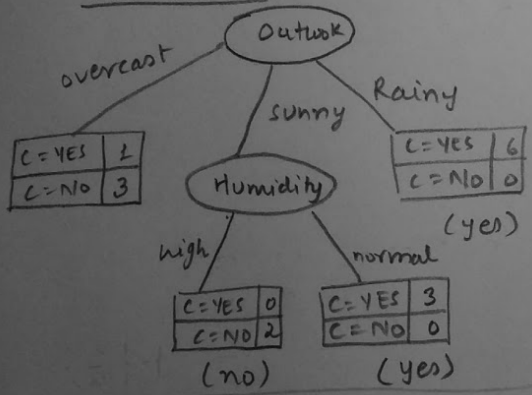
Δinfo gain = E(Parent) − E(humidity)

$= 0.97095 - \left[\frac{3}{5} \times 0 + \frac{2}{5} \times 0\right]$

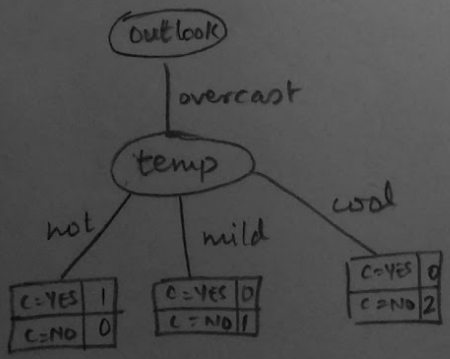
$= 0.97095$

Δinfo gain (humidity) > Δinfo gain (temp)

Decision Tree



For



E(Outlook = Overcast)

$= -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right)$

$= 0.81128$

E(temp = hot) = 0

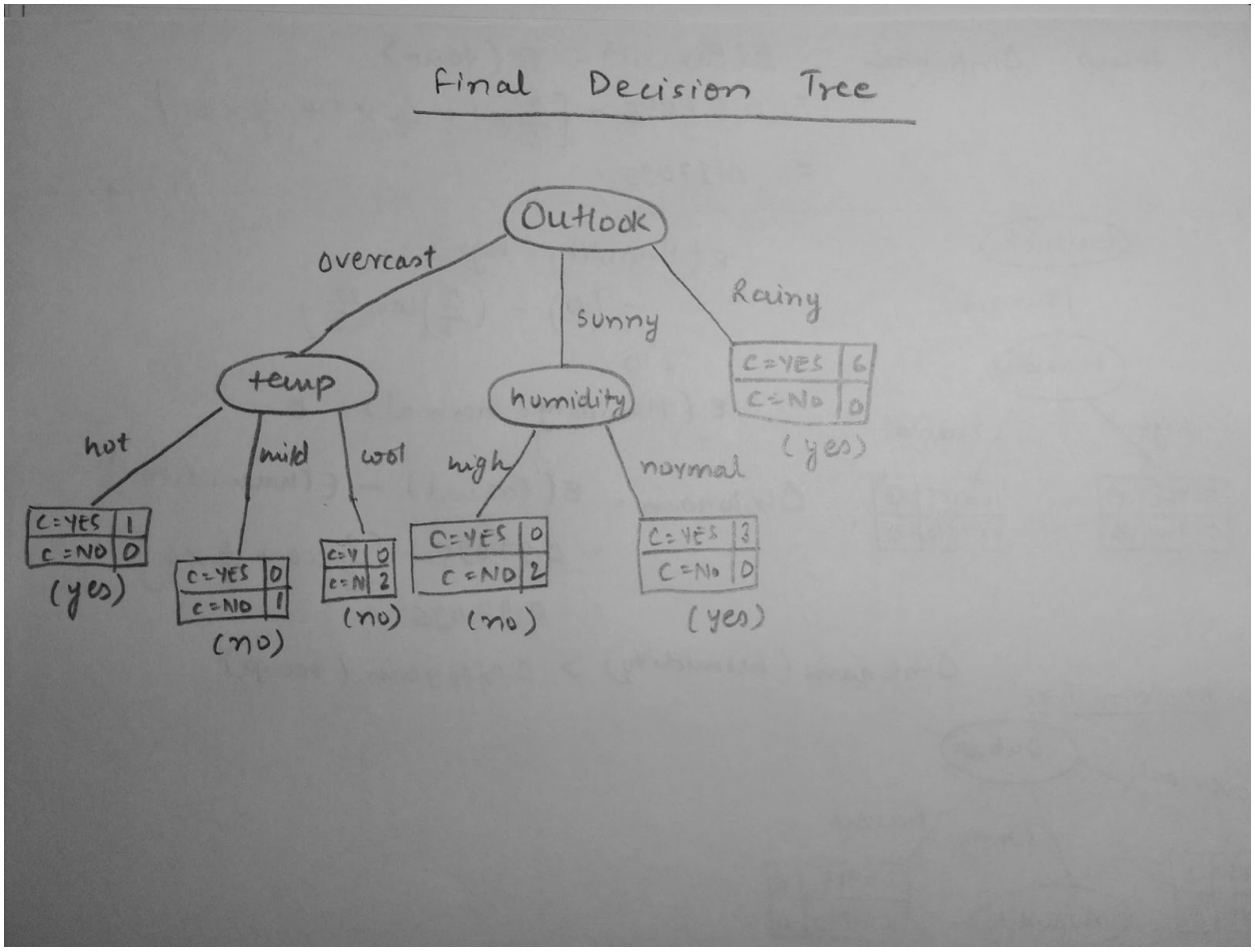
E(temp = mild) = 0

E(temp = cool) = 0

Δinfo gain = E(Parent) − E(temp)

$= 0.81128 - 0$

$= 0.81128$



2.For the following data, predict class label for each instance using the tree constructed in

ID, Temperature, Outlook, Humidity, Y=play (yes or no)
1,hot,overcast,high,? YES
2,mild,overcast,high,? NO
3,cool,sunny,normal,? YES
4,mild,rainy,normal,? YES
5,mild,sunny,normal,? YES
6,cool,rainy,high,? YES

Q3. For given satellite image (ilk-3b-1024.tif; 3-dimensional, 1024x1024), and ground-truth (training/test) data.

1. **Do object-based classification, using the following outline.**
    a. **Segment the image (use any method): See: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/**
    b. **Using the ground-truth, do classification (on segmented image).**
    c. **Report accuracy (full contingency table).**
    d. **Submit image (tiff) as separate file)**

Solution :

Software Package : Monteverdi - 1.22.0[1]

a. Image Segmentation : For the image segmentation, Mean shift clustering segmentation has been used. The steps are as below:

(i) Open image into Monterverdi package and extract ROI from the dataset.

(ii) Now apply Mean Shift clustering as segmentation on the ROI extracted dataset. We can tune the parameters of mean shift clustering. (Spatial Radius =5, Spectral values =0.5, Min region size = 57)

(iii) After this run the setting and save the segmented image as tiff file. (Image: segment_mean_shift.tiff)

b. Image Classification using Segmented Image: In this step, SVM classification is used on the segmented image. Below are the steps:

(i) Used object labeling for attaching the ground truth data

(ii) Selected ground truth label for 5 classes(road as label 0, house as label 1, grass as label 2, trees as label 3,and trees as label 4), 10-15 samples each.

(iii) Selected the SVM classification parameters as Kernel = Radial Basis function, cross validation = 5, (image = svm_segment_classification.tif)

c. Accuracy : From the ground truth, that have been labeled ,

accuracy (On training Set) = **97.5%**

Now we can save the model as a vector output as segment_classification.shp file.

We can load this file as vector output for training set, and input image as original image.

The confusion matrix obtained :

Confusion matrix:

	0	1	2	3	4
0	38	25	25	0	18
1	10	95	4	0	1
2	17	3	77	3	17
3	0	0	0	91	8
4	2	0	7	9	78

Overall accuracy: 0.717803

Kappa: 0.647246

d. Output Image : svm_segment_classification.tif

2. **Discretize the image first (10-bins; choose bin width such that you get roughly a uniform distribution). Then do decision tree classification on the discretized image, and compare (accuracy) it with actual image. Comment on advantages/disadvantages of decision trees with conterminous random**

variables (e.g., actual images) and discrete random variables (discretized image).

Solution :
The following steps have been performed to discretize the image.

a. Load the image using rgdal package in R
b. Create a vector by combining all the RGB bands of 1024 x 1024 pixels.
c. Now we divide the RGB bands into 10 equal frequency bins and project the RGB bands into bin number 1 to 10 by comparing its RGB band value to bin boundaries.
d. Now after discretizing the entire image along with training and test set, we can use the same process as described in question 1.
e. Decision tree model is built and tested and then the same model is applied to predict labels for entire image.

For test dataset: After image discretization (Image : Discretized-10Bin-DT.tiff)
Accuracy : 82.14%

	Reference				
Prediction	**1**	**2**	**3**	**4**	**5**
1	7	0	1	2	0
2	1	6	0	0	0
3	0	0	6	0	0
4	0	0	0	2	1
5	0	0	0	0	2

For decision tree, with conterminous random variables tend to generate complex model i.e. the rectilinear boundaries are complex and overfitting. Since in the image classification, we classify of RGB band values ranging from 0-255, for which certain band range has similar texture and tone. Without discretization, the model try to capture as many nuances of the band variability which leads to **model overfitting.** Discretization smoothes the RGB band range and allow the model to fit more general decision boundary. Thus discretize image decision tree model are less complex and provide more accuracy.

Q4. Using SaTScan (http://www.satscan.org/datasets/nyscancer/index.html), analyze the NY State Cancer data (http://www.satscan.org/datasets/nyscancer/index.html), produce cancer cluster map (similar to Figure 2 on Page 25 of the pre-print paper: http://www.satscan.org/datasets/nyscancer/nyscancerdata.pdf). Go through the online tutorials to familiarize yourself with the s/w first. Briefly write your findings (no more than half-page) and embed the map.

The objective here is to use SaTScan package[2] to find the significant regions in NY State region where the breast cancer observations are higher than expected. Here the data and population files are loaded and configured as input. Spatial coordinates are taken as longitude and latitude and observation period is set for year 2009. Then we run the statistical test for

poisson distribution as probability distribution model . After this we also tune the parameter monte-carlo replications as 999. Running the configured model of SaTScan, following results are obtained :

a. 7 different clusters are observed, and clusters are projected over the significant regions using google earth files.
b. We then project the clusters over NYS region and the output image is embedded.
c. Clusters obtained have significant REL_RISK factor for which spatial regions have been identified. For output file : NYC_BreastCancer.col
d. Cluster 2 and Cluster 4, has more breast cancer cases observed than expected.

References:

[1] Object-based classification (Tutorial), using Monteverdi package, http://wiki.awf.forst.uni-goettingen.de/wiki/index.php/Object-based_classification_(Tutorial)

[2] SatScan tutorial
http://www.satscan.org/tutorials/nyscancer/SaTScanTutorialNYSCancer.pdf