CSC 458 Data Mining and Predictive Analytics I, Fall 2019

Dr. Dale E. Parson, Assignment 5, Comprehensive Final Exam Project.
Due by 11:59 PM on Wednesday December 11 via make turnitin. I will NOT accept solutions to this Assignment 5 after noon on Thursday December 12.

Perform the following steps to set up for this semester's projects and to get my handout. Start out in your login directory on csit (a.k.a. acad).

cd $HOME
mkdir DataMine # This should already be there from assignment 1.
cp ~parson/DataMine/csc458fall2019assn5.problem.zip DataMine/csc458fall2019assn5.problem.zip
cd ./DataMine
unzip csc458fall2019assn5.problem.zip
cd ./csc458fall2019assn5

EDIT THE SUPPLIED README.txt when the following questions starting at Q1 below. Keep with the supplied format, and do not turn in a Word or PDF or other file format. I will deduct 20% for other file formats, because with this many varying assignments being turned in, I need a way to grade these in reasonable time, which for me is a batch edit run on the **vim** editor.

There are three ARFF files in the handout directory.
HawkData20172018Assn5.arff no compression of BWbins=0 instances
HawkData20172018Compressed5.arff compression of contiguous BWbins=0
HawkData20172018Assn5ZDown.arff is HawkData20172018Assn5.arff with 90% of the BWbins==0 instances removed after randomization.

There is also a prep/ subdirectory with the Python scripts used to prepare this assignment. I have included it only for class-time discussion. You will not change or use it directly.
prep/HawkData20172018Assn5PrePython.arff
Input data from a previous assignment, with some editing for assignment 5.
prep/arffio.py
My ARFF I/O library with some enhancements for this assignment.
prep/HawkAssn5GetPrev.py
Maps HawkData20172018Assn5PrePython.arff -> HawkData20172018Assn5.arff
prep/HawkAssn5CompressZeroes.py
Maps HawkData20172018Assn5.arff -> HawkData20172018Compressed5.arff

RULES FOR THE FINAL

This is an exam. Therefore, I will answer questions only in class on December 4 and December 11 (6-8 PM on the 11th per final exam hours), other than to clarify any confusing wording and correct any mistakes in this handout. I will email any replies regarding confusing wording or mistakes to the entire class. Regular assignments are learning experiences, and I am happy to drop hints and otherwise encourage students when asked. However, for a final exam you will have learned how to use Weka and interpret data via previous assignments. Also, note the late restriction above.

STEP1: Load **HawkData20172018Assn5.arff** into Weka. This ARFF file is a modified variant of Hawk Mountain data from previous assignments. It contains the following attributes. Note the attributes tagged with **R** for Remove in **STEP2**. These attributes are also underlined below. We are using only Hawk Mountain North Lookout observation data in Assignment 5. There is no Weather Underground or NOAA Sunrise data.

HawkYear	**R**	2017 or 2018 for this dataset
msnyHstart	**R**	Minutes since observation's previous New Year for hawk watch start.
msmnHstart		Minutes since observation day's previous midnight for hawk watch start.
msnyHend	**R**	Minutes since observation's previous New Year for hawk watch end.
msmnHend	**R**	Minutes since observation day's previous midnight for hawk watch end.
msToYearPeak		Minutes to BW peak count for this year, from Assignment 3.
msDuration		Minutes duration of this instance.
WindSpd		North lookout wind speed as a nominal value, via portable anemometer
		{'0: less than 1km/h (Calm)', '1: 1-5 km/h (1-3 mph)', '2: 6-11 km/h (4-7 mph)', '3: 12-19 km/h (8-12 mph)', '4: 20-28 km/h (13-18 mph)',' '5: 29-38 km/h (19-24 mph)', '6: 39-49 km/h (25-31 mph)', '7: 50-61 km/h (32-38 mph)', '8: 62-74 km/h (39-48 mph)', '9: Greater than 75 km/h'}
WindDir		North lookout wind direction
		{Variable,WNW,NW,SE,E,S,ESE,SW,SSW,N,NNW,NE,ENE,W,WSW,NNE,SSE}
Temp		North lookout Celsius temperature
CloudCover		North lookout cloud cover, units of measure unknown
Visibility		North lookout visibility, units of measure unknown
FlightDIR		Raptor nominal flight direction (SE, etc.), same value set as WindDir
FlightHT		Raptor flight height as a nominal value
		{'0: Below eye level', '1: Eye level to 30m', 2: Unaided eye', '3: limit of unaided vision', '4: Binoculars (to 10X)', '5: At limit of binoculars (10X)', '7: Variable',(none)}
SkyCode		
		{'0: Clear', '1: Partly Cloudy', '2: Mostly Cloudy', '3: Overcast', '4: wind driven sand, snow, dust', '5: Fog or Dense Haze', '6: Drizzle','7: Rain', '8. Snow'}
BW	**R**	Broad-winged Hawk count for that observation interval.
BWbins		BW compressed numeric value per Assignment 3 AddExpression[1].
HTempPrev72		Temp 72 hours before this instance.
HTempDelta72		Temp - HTempPrev72
HTempPrev48		Temp 48 hours before this instance.
HTempDelta48		Temp – HtempPrev48
HTempPrev24		Temp 24 hours before this instance.
HTempDelta24		Temp – HtempPrev24

Attribute List 1 from HawkData20172018Assn5.arff

[1] Assignment 3's **BWbins** AddExpression
ifelse(aBW=0,0,ifelse(aBW=1,1,ifelse(aBW=2,2,ifelse(aBW<30,3,ifelse(aBW<200,4,ifelse(aBW<1000,5,6)))))).

STEP2: **Remove HawkYear**, **msnyHstart**, **msnyHend**, and **msmnHend** (attributes tagged with **R**), since they correlate strongly with either msToYearPeak or msmnHstart. **Remove BW**, also tagged with **R**, because **BWbins** is the class attribute for this assignment, and **BW** correlates strongly with **BWbins**. This removal eliminates trivial prediction of **BWbins** values via **BW** by the models. **Reorder** attributes to place **BWbins** as the final attribute in the Preprocess list, without changing the relative order of the remaining attributes. **SAVE** this 18-attribute working dataset in an ARFF file named **STEP2.arff**. You will work out of this dataset until **STEP3**. Copy your **STEP2.arff** into the acad assignment directory for later **make turnitin**.

Each of Q1 through Q12 is worth 8.33% of this assignment. There are 2 ARFF files to turn in.

Q1: **BWbins** is our class attribute for this assignment. Run **LinearRegression**, **M5P**, and **Rules -> M5Rules** classifiers, and paste the following result values into Q1 in README.txt. All testing in Assignment 5 uses 10 fold cross-validation, i.e., no external test dataset. How do the LinearRegression and M5P results (correlation coefficient & error measures) compare with your results or my results for Q1 of Assignment 4? I will post my Assignment 4 results after I receive all Assignments 4, or 9 AM on Friday December 6, whichever comes first. I'll email the class.

LinearRegression
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

M5P
Number of Rules : N
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

M5Rules
Number of Rules : N (This is the Rule number of the last Rule listed. Examine the Rules.)
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

Q2: **Unsupervised -> attribute -> Discretize** BWbins into 7 bins with useEqualFrequency=False and ignoreClass=True. Be very careful to Discretize ONLY the BWbins attribute. We will UNDO this step later. Make sure the 7 discretized bins have the same instance counts as their pre-Discretize numeric bins in the Preprocessor. See Figure 1 below. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q2 in README.txt. How do the NaiveBayes, BayesNet, and J48 results (% correct and Kappa) compare with your results or my results for Q2 of Assignment 4?

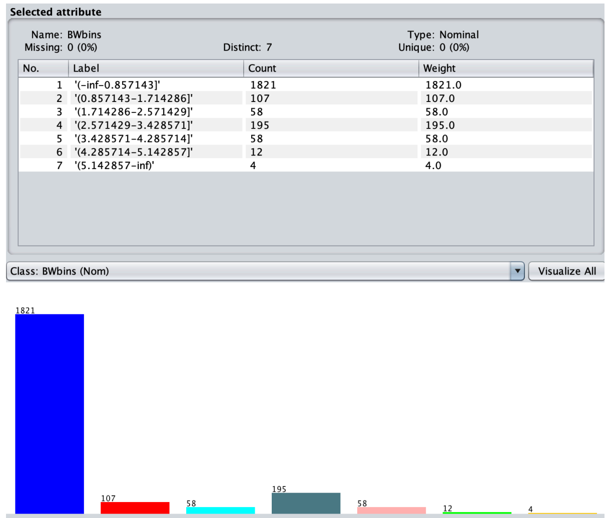

Figure 1: BWbins distribution after Q2 Discretize step.

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

BayesNet
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

J48
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

Q3: What do the changes in Q1 and Q2 in this assignment from the Q1 and Q2 results in Assignment 4 tell you about the importance of collecting weather station data at a separate location (Hamburg) from the raptor data collection site? Note that Q1 and Q2 in Assignment 4 use Hamburg weather station data, while the Assignment 5 dataset does not use weather station data.

Q4: Why do you see the direction of changes (better or worse for each of NaiveBayes, BayesNet, and J48) for Q2 in this assignment from the Q2 results in Assignment 4? In other words, why do these specific modeling algorithms improve or degrade from Assignment 4? (Refer to your Q2 answer above for: How do the NaiveBayes, BayesNet, and J48 results (% correct and Kappa) compare with your results or my results for Q2 of Assignment 4?)

STEP3: Load **HawkData20172018Compressed5.arff** into Weka. This dataset compresses all instances with BW==0 and BWbins==0 (those are the same instances) that are contiguous in time into single instances. The intent is to treat each block of BW==0 instances as a single datum in which *nothing is happening*. The hope is to reduce the effect of the overwhelming number of BW==0 instances on analysis. Assignment 3 compressed the magnitude outliers of BW into BWbins. **HawkData20172018Compressed5.arff** also compresses the BW==0 histogram outlier. Compare Figure 1's BWbins distribution with Figure 2 below. Note that all bins except bin 0 retain their counts from Figure 1.

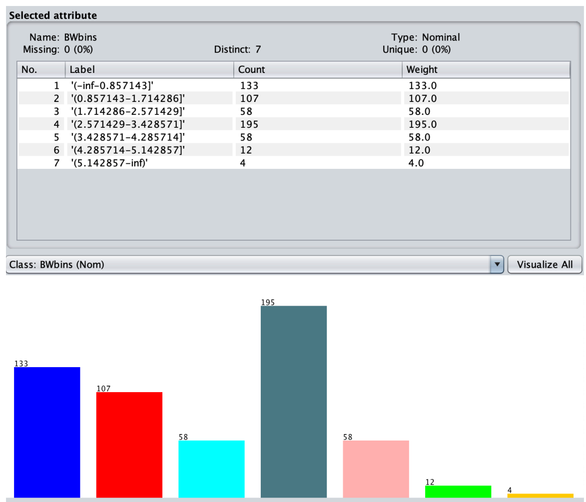

Figure 2: BWbins distribution for HawkData20172018Compressed5.arff.

STEP4: **Remove** the instances removed in **STEP2** (**HawkYear**, **msnyHstart**, **msnyHend**, **msmnHend**, and **BW**) for the same reasons, and **Reorder** attributes to place **BWbins** as the final attribute in the Preprocess list, without changing the relative order of the remaining attributes. In addition to the attributes from **HawkData20172018Assn5.arff**, **HawkData20172018Compressed5.arff** adds the following attributes.

TempMean
CloudCoverMean
VisibilityMean
HTempPrev72Mean
HTempDelta72Mean
HTempPrev48Mean
HTempDelta48Mean
HTempPrev24Mean
HTempDelta24Mean
TempMedian
CloudCoverMedian
VisibilityMedian
WindSpdMedian
FlightHTMedian
SkyCodeMedian
HTempPrev72Median
HTempDelta72Median

HTempPrev48Median
HTempDelta48Median
HTempPrev24Median
HTempDelta24Median
WindSpdMode
WindDirMode
FlightDIRMode
FlightHTMode
SkyCodeMode

Attribute List 2 added by HawkData20172018Compressed5.arff

For BWbins==0 single instances compressed from temporally contiguous BWbins==0 instances in **HawkData20172018Assn5.arff**, the *Mean, *Median, and *Mode attributes give the mean (average), median (center value), and mode (most frequently occurring value) for their named counterparts. For numeric values such as **Temp** this relation holds the Mean and Median. For ordered nominal values such as **WindSpd** it holds the Median and Mode. For cyclic values such as **WindDir** it holds only the Mode; these attributes are intrinsically non-linear, wrapping around at North, so Mode is the only measure that makes sense.

For BWbins>0, non-compressed instances as they are in **HawkData20172018Assn5.arff**, the Mean, Median, and Mode fields are identical to their source values. For example, TempMean == TempMedian == Temp for instances with BWbins>0, since these instances are original, uncompressed instances.

STEP5: Remove the attributes originally in **HawkData20172018Assn5.arff** whose names **PRECEDE** the suffixes **Mean**, **Median**, or **Mode** in Attribute List 2. The first to remove from the above list is **Temp**, and the last is **SkyCode**. **DO NOT REMOVE ANY ATTRIBUTE WITH Mean, Median, or Mode IN ITS NAME**. Also, **Remove msDuration**, since the duration of the instance in minutes correlates strongly with BWbins==0 instances that have been compressed. I used **msDuration** to check output from Python script **HawkAssn5CompressZeroes.py**, but **msDuration** values > 60 basically reflect the compression process. SAVE this 29-attribute working dataset in an ARFF file named **STEP5.arff**. You will complete this assignment using this dataset. Copy your **STEP5.arff** into the acad assignment directory for later **make turnitin**.

STEP6: For any attribute prefix in Attribute List 2 with Mean as its suffix, temporarily Remove attributes with the same attribute prefix from the set of Median and Mode attributes. For example, since TempMean appears, Remove TempMedian. When that removal is complete, for any remaining attribute prefix in Attribute List 2 with Median as its suffix, temporarily Remove attributes with the same attribute prefix from the set of Mode attributes. For example, since WindSpdMedian appears, Remove WindSpdMode. STEP6 eliminates redundant values in the non-target attribute set. For instances with BWbins>0, these attribute values are identical in a given instance. For instances with BWbins==0, these attribute values are strongly correlated in a given instance. We are attempting to estimate the most predictive compressed attributes.

Q5: **BWbins** is our class attribute for this assignment. Run **LinearRegression**, **M5P**, and

M5Rules classifiers, and paste the following result values into Q5 in README.txt. All testing in Assignment 5 uses 10 fold cross-validation, i.e., no external test dataset. <u>How do the LinearRegression and M5P results (correlation coefficient & error measures) compare with your results for Q1 of Assignment 5 above</u>?

LinearRegression
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 567

M5P
Number of Rules : N
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 567

M5Rules
Number of Rules : N (This is the Rule number of the last Rule listed. Examine the Rules.)
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 567

Q6. Look at the tree structure and the Number of Rules for the M5P decision tree in Q5 compared with the M5P Number of Rules in Q1. Do you see any trade-off in minimum description length (tree simplicity) versus prediction accuracy in going from the data of Q1 to the compressed data of Q5?

Q7: **Unsupervised -> attribute -> Discretize** BWbins into 7 bins with useEqualFrequency=False and ignoreClass=True. Be very careful to Discretize ONLY the BWbins attribute. Make sure the 7 discretized bins have the same instance counts as their pre-Discretize numeric bins in the Preprocessor. See Figure 2 above. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q7 in README.txt. <u>How do the NaiveBayes, BayesNet, and J48 results (% correct and Kappa) compare with the Q2 results in this assignment above</u>?

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 567

BayesNet
Correctly Classified Instances N n.n %
Kappa statistic n.n

Total Number of Instances 567

J48
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 567

Q8: How do you judge the effectiveness of collapsing temporally adjacent BWbins==0 instances into single instances in dataset HawkData20172018Compressed5.arff, compared with the uncompressed dataset in HawkData20172018Assn5.arff, in terms of predicting BWbins values?

Q9: In the **Cluster** tab of Weka, run the **SimpleKMeans** clustering algorithm with the default parameters of 2 clusters, and paste the following output table and percentages of instances. Ignoring the Full Data column, what conspicuous differences do you see between the cluster (column) with BWbins==0 compared to the cluster (column) with a non-0 BWbins value?

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(n.n)	(n.n)	(n.n)
msmnHstart			
msToYearPeak			
TempMean			
CloudCoverMean			
VisibilityMean			
HTempPrev72Mean			
HTempDelta72Mean			
HTempPrev48Mean			
HTempDelta48Mean			
HTempPrev24Mean			
HTempDelta24Mean			
WindSpdMedian			
FlightHTMedian			
SkyCodeMedian			
WindDirMode			
FlightDIRMode			
BWbins	'(n.n-n.n]'	'(n.n-n.n]'	'(n.n-n.n]'

Clustered Instances
0 N (n%)
1 N (n%)

PREP for Q10: Load HawkData20172018Assn5ZDown.arff into Weka. This is the dataset of HawkData20172018Assn5.arff (Q1) in which I have: A) used Weka's **instance -> RemoveWithValues** to partition these instances into two ARFF files, one with all BWbins>0 instances, and the other with all BWbins==0 instances; B) used **instance -> Randomize** several times on the BWbins==0 instances, and then used instance -> RemovePercentage to remove 90% of the BWbins==0 instances; C) finally, I used the vim editor to merge the remaining 10% of BWbins==0 with all of the BWbins>0 instances to get the histogram illustrated in Figure 3. You do not have to do anything in this step other than loading HawkData20172018Assn5ZDown.arff.

The reduction in BWbins==0 instances is already saved in that file. There is no compression of temporally contiguous BWbins==0 instances in HawkData20172018Assn5ZDown.arff.

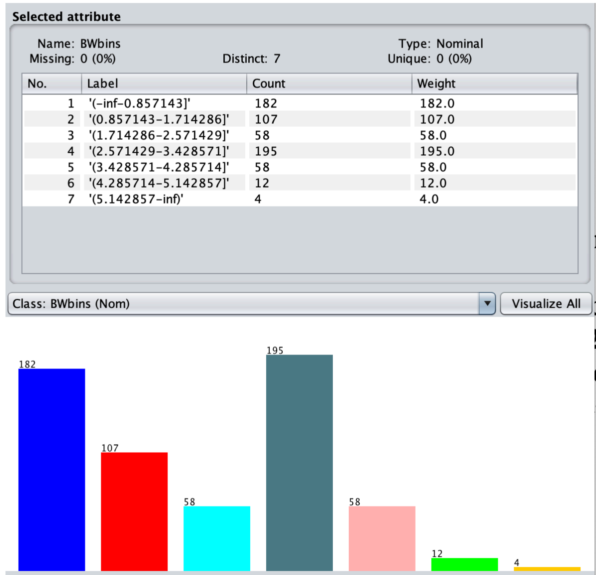

Figure 3: BWbins distribution for HawkData20172018Assn5ZDown.arff

Q10: Run the **M5P** classifier on this dataset and record results below. Next, run **Unsupervised -> attribute -> Discretize** BWbins into 7 bins with useEqualFrequency=False and ignoreClass=True. Then run the **BayesNet** classifier on this dataset. Record your results below, and compare the accuracy (correlation coefficient & kappa respectively) to the results of both Q1&Q2 from this assignment (for M5P and BayesNet), and to the compressed BWbins==0 results of Q5&Q7 (for M5P and BayesNet). **Strictly in terms of correlation coefficient for M5P, and Kappa for BayesNet**, how does this form of BWbins==0 instance count reduction compare with the multiple, temporally contiguous BWbins==0 instance compression of Q5&Q7? Are the correlation coefficient and kappa of Q10 within 10% of their values for Q1 and Q2? Use the formula **(Q10 metric – Q1orQ2 metric) / Q1orQ2 metric** to determine the percentage rise or fall from Q1 or Q2's metric, where **metric** is M5P's correlation coefficient or BayesNet's kappa.

Example: (.4-.5)/.5 would be a 20% drop from .5, not a 10% drop.

M5P from Q10:
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 616

BayesNet from Q10:
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 616

Q11: Review the histograms in Figures 1, 2, and 3, which show BWbins Discretized into 7 bins. Think about for which one ZeroR would achieve the highest Correctly Classified Instances, the original instances of Q2 in Figure 1, the compressed BWbins==0 instances of Q7 in Figure 2, or the sampled BWbins==0 instances of Q10 in Figure 3. ZeroR on a discretized class attribute always has a Kappa of 0. How is it possible that the BayesNet Kappa of Q10 (Figure 3) approaches the BayesNet Kappa of Q2 (Figure 1), despite the substantial reduction in Correctly Classified Instances going from Q2 to Q10?

Q12: These points are for a correctly saved and turned in STEP2.arff and STEP5.arff.

When you have completed all of your work and double-checked the assignment requirements, and your **README.txt** that answers Q1 through Q12, and files **STEP2.arff** and **STEP5.arff** are sitting in your **csc458fall2019assn5**/ directory, then run **make turnitin** by the due date. Late assignments lose 10% per day late. **Due by 11:59 PM on Wednesday December 11 via make turnitin. I will NOT accept solutions to this Assignment 5 after noon on Thursday December 12.**