Project 2 – Decision Trees, Linear Regression, Model Trees, Regression Trees

CS548 Knowledge Discovery and Data Mining - Fall 2016

Prof. Carolina Ruiz

Students: Mu Niu, Rohitpal Singh, Suchithra Balakrishnan

	Classification		**Regression**	
Dataset :				
Dataset Description	/05			
Data Exploration	/10			
Initial Data Preprocessing (if any)	/05			
	Weka	**Python**	**Weka**	**Python**
Code Description:	/10	/10	/10	/10
Experiments:				
Guiding Questions	/10		/10	
Sufficient & coherent set of experiments	/10	/10	/10	/10
Objectives, Parameters, Additional Pre/Post-processing	/10	/10	/10	/10
Presentation of results	/10	/10	/10	/10
Analysis of individual experiments’ results	/10	/10	/10	/10
Summary of Results, Analysis, Discussion, and Visualizations	/20		/20	
Advanced Topic	/30			
Total Written Report	/310 = /100			

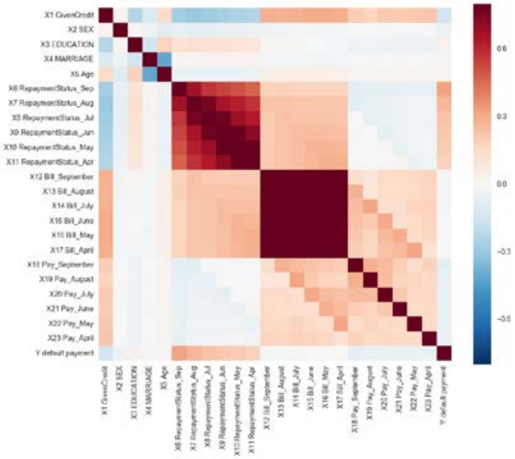
Dataset Description, Exploration, and Initial Preprocessing: (at most 1 page)

[05 points] Dataset Description: (e.g., dataset domain, number of instances, number of attributes, distribution of target attribute, % missing values, …)

This dataset describe the credit card default situation in Taiwan including clients’ basic information like education, age and sex, monthly payment record and bill statement from April to September in 2005. There are 30,000 instances, 24 attributes and no missing values.

The target attribute is binary (1 = Yes, 0 = No), default payment. There are in total 6636 instances to be “1”, contain 22.12%, and the rest 23364 instances are “0”, contain 77.88%.

[10 points] Data Exploration: (e.g., comments on interesting or salient aspects of the dataset, visualizations, correlation, issues with the data, …)



The dataset has some interesting aspects. There’s a problem in attribute X3, according to data description, it should have just 4 values. But instead, it have 7 values in total.

We calculated the correlation for the whole dataset and made a heat map based on correlation matrix. From the correlation heatmap, we found that attributes from X12 to X17, which represent amount of bill statement have really high correlation with each other but very little correlation with the target attribute. So if we selected the best subset, maybe just one of the attribute in these group is needed. Another interesting thing is that it seems like continuous attributes tend to have less correlation with target attribute than nominal attributes.

We think if the dataset can be improved, it would be useful if they make clearer statement of how they collect data and what their data mean in more specific way.

[05 points] Initial data preprocessing, if any, based on data exploration findings: (e.g., removing IDs, strings, necessary dimensionality reduction, …)

We removed the first column represents ID, and the second row that describe attributes’ meaning. We didn’t remove other attributes in the initial preprocessing since they can be useful in later experiences

For attribute X3, we replaced 0, 5 and 6 to be 4 (others). There are 487 replacements in total.

For analyses just with nominal attributes, we replace numbers to text in order to have a better understanding on how the attributes distributed.

Weka Code Description: Inputs, output, and process followed <u>by Weka's code</u> to construct the trees (at most 2/3 page)

[10 points] J4.8 Code Description:

J4.8 is to build decision tree based on entropy. It firstly calculates entropy of target attribute and entropies between attributes and target attributes. They calculate the information gain by algorithm $information\ gain\ =\ entropy(target) -\ entropy(target, attribute)$, then choose the attribute with the highest information gain as the decision node. If the information gain of one branch is zero, it's set as a leaf node. If positive, the attribute is splitted further with newly calculated information gain. After forming the decision tree, nodes with information gain less than 0.25(default value) are pruned to make the tree smaller.

Input: minNoObj, Dataset and Target Attribute, useMDLcorrection and doNotMakeSplitPointActualValue

Output: Tree, No. of Leaves, Size of Tree, Time Taken to build Model, Summary, Performance Metrics (detailed accuracy by class) & confusion Matrix.

[10 points] M5P Code Description:]

M5P is used to form regression tree and model tree based on standard deviation reduction (SDR), which is expressed as $SDR\quad_i = std(target) - \sum_i \quad \frac{|T_i|}{|T|} \times std(T_i)$. The algorithm replaces information gain in J4.8 with SDR. If the SDR is non-positive or number of instances per leaf is less than 4(default), the splitting stopped. Then the leaf is represented as the mean of instances for regression tree and a regression linear function in model tree.

Input: credit.csv and target attribute

Output: Rules, Number of Rules, Time taken to build model, summary and Tree.

[20 points] Python Packages and Functions used (decision trees, linear regression, model/regression trees). Describe inputs & outputs (at most 1/3 page)

	package	Function	Input	Output
Decision tree	Scikit-learn(tree) time	DecisionTreeClassifier time	Dataset starting time	tree, performance score time to processing
Linear regression	sklearn,numpy,pandas,time	KFold(), LinearRegression(), fit(), predict(), score(), mean_squared_error(),corr()	credit_all.csv	R^2, Root Mean Sqr Error, corr matrix, time to create model
Regression tree	sklear,graphviz,numpy,pandas,time	KFold(), DecisionTreeRegressor(), fit(), predict(), score(), mean_squared_error(), f_regression(), export_graphviz()	credit_all.csv	R^2, Root Mean Sqr Error, Tree, coefficient values, time to create model

[10 points] Three Guiding Questions for the Classification Experiments: (at most 1/3 page)

1. **How education influence the default action? Which education level have most defaulters by percentage?**
2. **Which method will give the least processing time?**
3. **Is credit limit determine or at least, highly related with default activity?**

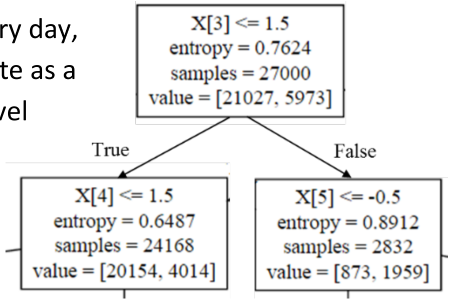
[40 points] Summary of Classification Experiments in Weka. Use 10-fold cross-validation *At most 2/3 page.*												
Tech.	**Guiding questions**	**Pre-process**	**Parameters**	**Post-process & Pruning**	**Accuracy, Precision, Recall, ROC Area**				**Time to build model**	**Size of model**	**Interesting patterns in the model**	**Analysis & observations about experiment**
					Accu (%)	Pres	Rec	ROC	Sec (s)			
ZeroR	2	NA	ALL	NA	77.88	0.607	0.779	0.5	0.12		Only the target attribute is considered	The accuracy is the same as percentage of target attribute with value 0. It takes the shortest time.
OneR	1, 2	NA	X3, Y	NA	77.88	0.607	0.779	0.5	0.05		Accuracy is the same as in OneR method	This shows that X3 is not highly correlated to Y, to produce high accuracy
OneR	2, 3	NA	X1, Y	NA	77.88	0.607	0.779	0.5	0.03		Accuracy is the same as in OneR method	This shows that X1 is not highly correlated to Y, to produce high accuracy
OneR	2, 3	Attr Sel	X6, Y	NA	81.943	0.803	0.819	0.643	0.04		X6 is highly correlated to Y	The accuracy is greatly increased with X6
J4.8	1, 2	NA	Nom Attr	NA	82.016	0.803	0.820	0.708	1.03	224	High accuracy	Using all nominal values gives high accuracy and less time to build
J4.8	1, 2	NA Bin Splits	Nom Attr	NA	81.87	0.802	0.819	0.689	2.14	193	Decrease in size of model due to binary split	Creating binary splits decreases the tree size but accuracy is similar
J4.8	1, 2, 3	NA	All	NA	76.87	0.755	0.769	0.645	18.16	5123	Long time to build model	Using all attributes does not increase the accuracy and large tree is produced.
J4.8	1, 2, 3	Discretize	All	NA	74.87	0.749	0.749	0.648	0.73	18563	Less time to build. Larger tree	Discretization has decreased the time to build the tree but has increased the size of the tree with less accuracy
J4.8	1, 2, 3	Discretize	All	Pruned	81.93	0.802	0.819	0.704	1.75	208	High accuracy	Pruning discretized attributes tree produces higher accuracy, lesser time to build and smaller tree size.

[40 points] Summary of Classification Experiments in Python. Use 10-fold cross-validation *At most 2/3 page.*

Tech.	**Guiding questions**	**Pre-process**	**Parameters**	**Post-process & Pruning**	**Accuracy, Precision, Recall, ROC Area**	**Time to build model**	**Size of model**	**Interesting patterns in the model**	**Analysis & observations about experiment**
ZeroR	2	NA	Y	NA	77.88%	0.0 s	2	We don't need to think about attribute other than the target attribute.	the accuracy is the same as percentage of target attribute with value 0. It takes the shortest time.
OneR	1,2	NA	X6, Y	NA	81.99%	0.09 s	22	X6 is the best attribute in both nominal and continuous attributes	Adding one highly correlated attribute improve accuracy significantly.
OneR	1,2	NA	X3, Y	NA	77.88%	0.02 s	162	The result is no better than ZeroR method.	Education is not deterministic attributes for prediction
OneR	3	NA	X1, Y	NA	77.88%	0.02 s	8	The result is no better than ZeroR method.	Credit limit is not a good representative on target attribute.
Decision tree	1,2,3	NA	All	NA	81.19%	13.4 s	588	It takes a long time to form a huge tree.	The tree is too big to analyze, we need to find a way to shorten it
Decision tree	1,2,3	Discretize	All	NA	81.65%	2.45 s	625	Node number decreased after discretize, but accuracy increases	The tree size and accuracy haven't changed enough after discretization.
Decision tree	1,2	Dice	nominal	NA	82.09%	1.13 s	59	It actually more accurate if we use nominal attributes only	The first splitting point is X3<=1.5 which means graduate level of education significantly impact default action.

[20 points] Summary of Weka and Python Classification Results, Analysis, Discussion, and Visualizations (at most 1/3 page) 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best classification model obtained.

ZeroR method gives the least processing time with acceptable accuracy. If dataset domain need to deal with massive data every day, ZeroR would be a reasonable method for classification. Although X3(education) is not a very good estimator for target attribute as a whole, but according to the decision tree we made, we can see that split between graduate education and other education level did give us the most information gain among dataset. The credit limit does not have enough affection on target attribute, we cannot make prediction based on that attribute. The best classification we have is the decision tree formed by all nominal data, which have 82.09% accuracy with just 59 nodes. In these classification experiences, we learned that containing all data may cause overfitting problem and reduced accuracy instead of increasing. We can use ZeroR as a benchmark to evaluate our model. Plot is the first splitting point.



[10 points] Three Guiding Questions for the Regression Experiments: (at most1/3 page)

1. **What is the underlying behaviour of payment by defaulters and non-defaulters for PAY_AMT6?**
2. **What are the least attributes that can be used to predict the target PAY_AMT6 with high accuracy?**
3. **Number of people who paid highest on an average in April (PAY_AMT6) based on total pay over six months ?**

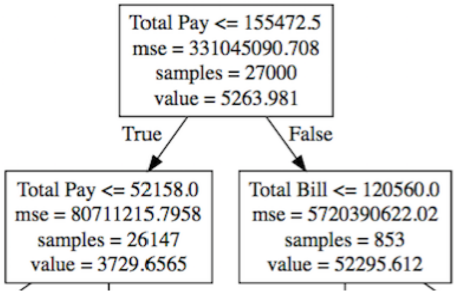
[40 points] Summary of Regression Experiments in Weka. Use 10-fold cross-validation. *At most 2/3 page.*

Tech.	**Guiding questions**	**Pre-process**	**Parameters**	**Post-process & Pruning**	**Correlation Coefficient and Error Metric(s)**	**Time to build model**	**Size of model**	**Interesting patterns in the model**	**Salient observations about experiment**
Linear Regression	1	NA	ALL	NA	CC-0.3691 RMSE-16524 R^2-0.1452	0.32s	17	PAY_AMT6 is highly dependent on LIMIT_BAL & BILL_AMT5	'default payment next month' attr is not considered for predicting PAY_AMT6
Linear Regression	1	Cfs	LIMIT_BAL,PAY_AMT[1..5], 'default pay next month'	NA	CC-0.2951 RMSE-16988 R^2-0.0935	0.06s	7	Defaulters are likely to pay 365 less against non-defaulters	t-stat is negative that indicates 'default payment next month' is not having any linear relationship with PAY_AMT6 .
Regression Tree	2	Cfs,feature creation	LIMIT_BAL,Tot Pay, Tot Balance,Tot time missed	NA	CC-0.6297 RMSE-13827	1.12s	77 Rules	Large tree, not good to interpret. This gives only high accuracy.	Total Pay over last six month is chosen to split the tree at the root node.
Model Tree	3	Cfs,minNumInstances, Feature Creation	LIMIT_BAL,Tot Pay, Tot Balance,Tot time missed	NA	CC-0.5174 RMSE-15223	0.45s	6 Rules	With only few parameters accuracy of predicting **PAY_AMT6** increased.	6087 customers who used to pay '> 40239.5' over six months they also paid highest in april (**PAY_AMT6)**

[40 points] Summary of Regression Experiments in Python. Use 10-fold cross-validation. *At most 2/3 page.*									
Tech.	**Guiding questions**	**Pre-process**	**Parameters**	**Post-process & Pruning**	**Correlation Coefficient and Error Metric(s)**	**Time to build model**	**Size of model**	**Interesting patterns in the model**	**Salient observations about experiment**
Linear Regression	1	Feature Selection (Correlation)	LIMIT_BAL, PAY_AMT[1..5], 'default pay next month'	NA	R^2=0.08 RMSE=16721	0.18s	7	Non Defaulters are likely to pay 95 less than Defaulters.	$R^{2=}$ 0.002 for regression of 'default pay next month' on **PAY_AMT6,** indicates no linear relationship.
Regression Tree	2	Attr. Selection, Feature Creation	**LIMIT_BAL,Total Time Missed,Total Bill,Total Pay**	NA	R^2=0.37 RMSE=13728	0.38s	29	Total Pay & LIMIT_BAL are having best (P-val) but 'Total pay' is imp attribute that increased the diff in variance with root node.	This model explained 37% of variance in the dataset which is very high as compared to baseline model.
Model Tree	3	Attr. Selection, Feature Creation	**LIMIT_BAL,Total Time Missed,Total Bill,Total Pay**	NA	R^2=0.357 RMSE=13971	0.42s	19	One of the leaf node does not belong to any particular category of 'Total Pay' at granular level.	

[20 points] Summary of Weka and Python Regression Results, Analysis, Discussion, and Visualizations (at most 1/3 page) 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best regression model obtained.

1. Original attributes in the dataset are very less correlated with the target and linear regression explained only 14% variance of the data as the baseline model. Creating new features like 'Total Bill' by taking sum of 'bill amt' over six months reduces the dimension as well as increases the R^2 to 37%.
2. Linear regression was useful to testify whether association exists between attributes and the target and attr. Individual's effect in making prediction. Regression tree predicted **PAY_AMT6** at a very high accuracy even with few parameters but it is not good for interpretation at granular level. Model Tree is also useful to see the attribute's effect on target but as size of tree increases it is difficult to interpret.
3. Image on right side for Regression tree i.e. split at root node using 'Total Pay'.



Advanced Topic: <Random Forest>

[7 points] List of sources/books/papers used for this topic (include URLs if available):

- Ho, Tin Kam. "The Random Subspace Method for Constructing Decision Forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence*20.8 (1998): 832-44. Web. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=709601
- Moorthy, Kohbalan, and Mohd Saberi Mohamad. "Random Forest for Gene Selection and Microarray Data Classification." *Communications in Computer and Information Science Knowledge Technology* (2012): 174-83. Web.
- Simon Bernard, Laurent Heutte, and Sebastien Adam, "A study of strength and correlation in RandomForests". Internation Conference on Intelligent Computing, Aug 2010, Changsha, China.
- https://en.wikipedia.org/wiki/Random_forest
- https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[20 points] In your own words, provide an in-depth, yet concise, description of your chosen topic. Make sure to cover all relevant data mining aspects of your topic.

Random forest method can be used in both decision tree and regression tree. It basically involves two methods, Decision tree algorithm and Bagging principle. An ensemble of decision trees are constructed in the training step using the Decision tree algorithm. Then the Bagging principle is applied which uses another Randomization technique called the Random Feature Selection.

In this method, they used bootstrap sample method to randomly choose S samples with equal sizes (usually 2/3 of the whole dataset) as their training data, then they used classical tree methods to form S unpruned trees using the training datasets they had selected. That would be the random forest model they had. Instead of always calculating the best split point for the whole dataset, this method gives prediction result by taking the mode of results from each tree's prediction. That is, for example, in decision tree, if the majority of the predict results for one particular sample is one class, the result from random forest would be this majority result.

To evaluate the performance of this method, they used an approach called *Out-of-bag (OOB) error*. To calculate this value, they used all the data that is not chosen in a particular training subset as their test subset (the rest 1/3 data). So they can calculate the performance for each of the S trees and average the performance to be method's performance. The Out-of-bag data is used to get a running unbiased estimate of the classification error of the trees.

In this method, after the construction of all the trees, all the data are validated with the trees and proximity is calculated for pairs of cases. Based on the proximity, they are normalized with the number of trees. These proximities can then be used to replace missing data, find outliers. The correlation is calculated between the predictions made by the different trees.

[3 points] How does this topic relate to trees and the material covered in this course?

This method can be applied in both classification and regression. Random forest largely uses the Decision tree mechanism to produce several decision trees using randomized data and features and then making a prediction. It has the advantage of avoiding the overfitting problem in typical decision trees.

Authorship: Although each student on the team is expected to be involved in every aspect of the project, describe in detail here the main contributions that each of the team members made to this project. This authorship description must accurately reflect the work done by each team member, and must be approved by all of the members of the team (at most 1/3 page)

	Initial preprocessing	J4.8 code	M5P code	Python package	Classification			Regression			advanced topic
					Weka	Python	summary	Weka	Python	summary	
Rohitpal Singh	√	√	√	√				√	√	√	√
Suchithra Balakrishnan	√	√	√		√		√	√		√	√
Mu Niu	√	√	√	√		√	√				√