

BUAN 6341 APPLIED MACHINE LEARNING ASSIGNMENT 2

The purpose of this assignment is to implement three learning algorithms, Support Vector Machines (SVM), Decision Tree, Boosting, on the two datasets, "Seoul Bike-sharing Demand Data" and "IBM HR-Employee-Attrition & Performance." The tasks are to experiment and compare the different algorithms on the response of accuracy.

Solutions

1. Experimental results using the Seoul Bike-Sharing Demand Dataset

(Table 1. Represented Data table of the Seoul Bike-Sharing Demand Dataset)

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes

The machine learning algorithms were used to develop prediction models to predict the hourly bike renting based on the dataset (Table 1). Since the information about hourly bike renting is important for the bike-sharing company to improve its bike renting management. The "Functioning Day" column data were hourly bike renting records and were output data. The bike renting data in a specific hour has two category values: with bike rent=1 or without bike rent=0. Before applying the learning algorithms, the dataset should be processed into the desired format.

Dataset processing steps.

1. Ignored the "Date" column since the prediction model is a 24 hours model with hourly prediction and not considering a specific year and date.
2. Ignored the "Rented Bike Count" column since the "Rented Bike Count" information is highly correlated with the output, the bike renting status. The bike renting status will be "with bike renting" for count value other than zero; otherwise, the renting status will be "no bike renting". No further information is needed if considering the "Rented Bike Count" as an input attribute. Therefore, to develop a sensible hourly prediction model for predicting bike renting status, we ignore the "Rented Bike Count" column data.
3. Encoded "Functioning Day" data into two class labels, which 0 = No (295) and 1 = Yes (8465).
4. As required by the scikit package, the independent categorical variable "Seasons" was encoded into four binary columns: spring, summer, autumn, and winter. Similarly, the "Holiday" column was encoded into two binary columns.
5. Split the dataset into the training set (70% of data) and test set (30% of data). The following is the shape of split data.
[Y train: (6132,), X train:(6132, 15), Y test: (2628,), X test: (2628, 15)]
6. Normalize the dataset to diminish the scale effect of different columns data to avoid numerical difficulties and increase the algorithms' performance.

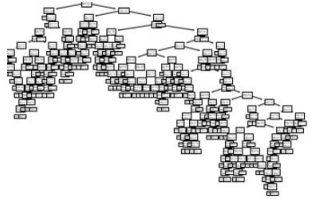

Outline of the principles for applying the machine learning algorithms.

1. Use the SVM models, Decision Trees, and Boosting algorithms supported in the scikit-learn machine learning package as the learning engines.
2. Apply a **10-Fold cross-validation** scheme to build the models.
3. Apply three SVM models with different kernel functions: Radial Basis function, Sigmoid and Polynomial to build the models.
4. Apply Decision Trees to build the model using Entropy to measure the quality of a split. The reasons to choose "Entropy" are most of the variables are not continuous attributes. Despite "Entropy" take more time to compute because of the log, most of the time, the performance of a model won't change whether you use "Gini" or "Entropy."
5. Apply AdaBoost as the Boosting algorithm.

Summary of the machine learning experiments.

Table 2 summarizes the accuracies of the prediction models built by different machine learning algorithms.

(Table 2. Summary of the accuracies for the selected machine learning algorithms)

Model No.	Methods	Model Training Accuracy	10-Fold Cross-Validation Accuracy	Model Testing Accuracy
Model 1	SVM with RBF	96.77 %	96.77 %, SD = 0.06 %	96.31 %
Model 2	SVM with Sigmoid	96.38 %	95.89%, SD = 0.48 %	95.74 %
Model 3	SVM with Polynomial	96.98 %	96.85%, SD = 0.19 %	96.19 %
Model 4	Decision Trees with best Tree	100 %	95.84%, SD = 0.65 %	95.93 %
Model 5	Decision Trees with 6 Depths	96.93 %	96.67%, SD = 0.24 %	96.23 %
Model 6	AdaBoost	96.74 %	96.54%, SD = 0.19 %	96.35 %
Decision Trees		Model 4		Model 5
Plot				

SVM Models

- As indicated in Table 2, the testing accuracy of the RBF kernel is 96.31 % and has the best performance. The reason might be that the RBF kernel can more fit into the expression of data distribution.

Decision Trees

- The training accuracy of Decision Trees with the best tree, Model 4, is higher than the test accuracy and one-standard-error validation accuracy that indicated the model might be overfitting.
- Unlike Model 4, the training accuracy of Model 5 is close to the cross-validation accuracy and without the overfitting problem. Therefore, by limiting the tree depths, we can effectively overcome the overfitting problem and obtain a reasonable tree.
- Comparing Model 4 and Model 5, Model 4 is a high complexity tree model and has no bias (100% training accuracy) but with high variance (95.84% validation accuracy 95.84%). On the other hand, by limiting the tree depth to reduce the model complexity, as Model 5, we have a more balanced model with an increased bias (96.93% training accuracy) and lower variance (96.67% validation accuracy). The result is consistent with the general principle that increasing the model complexity can reduce the model bias but increase the prediction variance.
- The computing time of Model 5 is half of Model 4, and indicates that less tree depth can reduce calculation time. High computing efficiency can help process a large data set.

Boosting (AdaBoost)

- AdaBoost has the highest test accuracy of 96.35% and appropriate training and validation accuracy comparing to other models and might be the best algorithm for this dataset.

Model comparison

- The smaller in-sample error will make a higher out-sample error for all the algorithms because of the overfitting problem or the bias and variance trade-off.
- The training accuracies of all Models except of Model 4 are all within one standard deviation of the 10-Fold cross-validation accuracies and indicated they are good performing models according to the "one-standard-error" rule.

2. Experimental results using the IBM HR Analytics Employee Attrition & Performance

(Table 3. Represented Data table of the IBM Dataset)

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	Mo
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2	Sales Executive	4	Single	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male	61	2	2	Research Scientist	2	Married	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male	92	2	1	Laboratory Technician	3	Single	

Dataset link: <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

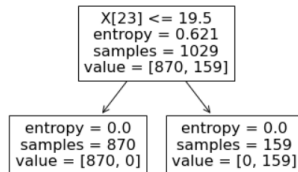
The machine learning algorithms were used to develop prediction models to predict employee performance based on the dataset (Table 3). Since the information about employee performance is important for all organizations to improve their human resource management. The "PerformanceRating" column data were number rating records and were output data. The Performance Rating in a specific number has two category values: with lower rating=3 or higher rating=4. Let's explore the reasons for job performance, and the result might help in my future

career performance. Before applying the learning algorithms, the dataset should be processed into the desired format.

Dataset processing steps.

1. Ignored the "PercentSalaryHike" column because it will direct to classify "Performance Rating" and will lose the meaning of exercise. It leads to 100% on all accuracies.

The resulting plot:



2. Encoded " PerformanceRating " data into two class labels, which 0 = 3 (1224) and 1 = 4 (226).
3. As required by the scikit package, the independent categorical variable "Attrition", "BusinessTravel", "Department", "EducationField", "Gender", "JobRole", "MaritalStatus" and "OverTime"...etc, are encoded into appropriate columns.
4. Split the dataset into the training set (70% of data) and test set (30% of data). The following is the shape of split data.
[Y train: (1029,), X train:(1029, 33), Y test: (441,), X test: (441, 33)]
5. Normalize the dataset to diminish the scale effect of different columns data to avoid numerical difficulties and increase the algorithms' performance


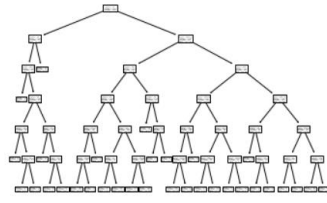
Outline of the principles for applying the machine learning algorithms.

1. Use the SVM models, Decision Trees, and Boosting algorithms supported in the scikit-learn machine learning package as the learning engines.
2. Apply a 10-Fold cross-validation scheme to build the models.
3. Apply three SVM models with different kernel functions: Radial Basis function, Sigmoid and Polynomial to build the models.
4. Apply Decision Trees to build the model using Entropy to measure the quality of a split. The reasons to choose "Entropy" are most of the variables are not continuous attributes. Despite "Entropy" take more time to compute because of the log, most of the time, the performance of a model won't change whether you use "Gini" or "Entropy."
5. Apply AdaBoost as the Boosting algorithm.

Summary of the machine learning experiments.

Table 4 summarizes the accuracies of the prediction models built by different machine learning algorithms.

(Table 4. Summary of the accuracies for the selected machine learning algorithms)

Model No.	Methods	Model Training Accuracy	10-Fold Cross-Validation Accuracy	Model Testing Accuracy
Model 1	SVM with RBF	84.55%	84.55%, SD = 0.25%	84.81%
Model 2	SVM with Sigmoid	83.67%	84.65%, SD = 0.71 %	84.58%
Model 3	SVM with Polynomial	88.92%	84.35%, SD = 0.19 %	84.13%
Model 4	Decision Trees with "best" tree	100.0%	72.89%, SD = 2.97 %	73.24%
Model 5	Decision Trees with 6 Depths	87.46%	80.28%, SD = 2.10 %	79.14%
Model 6	AdaBoost	84.35%	82.80%, SD = 1.22%	82.31%
Decision Trees		Model 4		Model 5
Plot				

SVM Models

- As indicated in Table 4, the RBF kernel's testing accuracy is 84.81 % and has the best performance. However, the testing accuracies with all models are without a big difference. For this dataset, SVM with RBF and Sigmoid kernels are both appropriate prediction models.
- The training accuracy of the Polynomial kernel function, Model 3, is not within one standard error of cross-validation and has lower testing accuracy. Therefore, we assume that Polynomial kernels are facing an overfitting problem.

Decision Trees

- The training accuracy of Decision Trees with the "best" tree, Model 4, is higher than the test accuracy and one-standard-error validation accuracy that indicated the model is overfitting.
- Unlike Model 4, the training accuracy of Model 5 is close to the cross-validation accuracy and has the overfitting problem. By limiting the tree depths, we can effectively reduce the overfitting problem and obtain a good tree.
- We obtain the same result form previous The Seoul Bike-Sharing Demand Dataset, which is consistent with the general principle that increasing the model complexity can reduce the model bias and increase the prediction variance.

Boosting (AdaBoost)

- AdaBoost has a test accuracy of 82.31% and appropriate training and validation accuracy; thus, AdaBoost is also a suitable prediction model.

Model comparison

- According to the experiment result, the SVM with RBF kernel might be the best Model for the IBM dataset because it created the best test accuracy, appropriate training and validation accuracy.
- IBM dataset is not suited for the Decision Trees algorithm because of the high training accuracy and lower testing accuracy. It has an overfitting problem compare to the other two algorithms.
- Some variables have a significant influence on the classification. If we didn't drop the "Percent Salary Hike for the IBM dataset," "Percent Salary Hike" will lead to a directed classification and no need for other variables. Therefore, we need to understand the dataset and the purpose of the analysis to choose the variables and explain the dataset wisely.

3. Conclusion

- There are several kernel functions for SVM. Three testing accuracies within the same datasets are little different. However, we can use training accuracies and cross-validation to find the best kernel.
- Each dataset needs to try multiple times to find the best-fit algorithm. For example, the Boosting for Bike-Sharing Demand Dataset and the SVM of RBF kernel for IBM dataset.
- Pruning the Decision tree might improve your model when it has an overfitting problem.
- High training accuracy usually causes low testing accuracy. If we lower the training accuracy, it will improve the testing accuracy. In other words, the model has bias and variance trade-off issues.
- The results indicated that all cross-validation accuracy is close to the test accuracy calculated by test data. Thus, the results implied the average validation accuracy could be an excellent estimator of the test accuracy.