Industrial Big Data Analytics and Machine Learning

Assignment #2

Note: *For the first three problems of this homework assignment, you are expected to build classifiers using the training data based on the algorithms covered in the lectures. Your calculation process should be presented in sufficient details, ideally step-by-step. You could use any basic calculation/computing tools (e.g. NumPy, Julia, Excel, etc.) to facilitate your calculation process. If you only present the final classification model and/or results without necessary intermediate steps and results, you will* ***not*** *get most of the credit.*

Problem 1

Given the following data set on whether to approve credit card applications based on the information of four features (student status, credit rating, available credit, and age group). Please build a Decision Tree Classifier using Information Gain to choose the best attribute, and then use the decision tree to predict the approval outcome of a new application from an unemployed, senior applicant with excellent credit rating and high available credit. Please provide sufficient details on your calculation process of building the classifier, **step-by-step**.

Employment Status	**Credit Rating**	**Available Credit**	**Age**	***Approve Application ?***
Unemployed	Excellent	High	Young	**No**
Unemployed	Fair	High	Young	**No**
Unemployed	Excellent	High	Middle Age	**Yes**
Unemployed	Excellent	Medium	Senior	**Yes**
Employed	Excellent	Low	Senior	**Yes**
Employed	Fair	Low	Senior	**No**
Employed	Fair	Low	Middle Age	**Yes**
Unemployed	Excellent	Medium	Young	**No**
Employed	Excellent	Low	Young	**Yes**
Employed	Fair	Medium	Young	**Yes**
Unemployed	Fair	Medium	Middle Age	**Yes**
Employed	Excellent	High	Middle Age	**Yes**
Unemployed	Fair	Medium	Senior	**No**
Unemployed	Fair	Low	Young	**No**
Unemployed	Excellent	Medium	Middle Age	**Yes**
Employed	Fair	Medium	Senior	**No**
Employed	Excellent	High	Senior	**Yes**
Employed	Fair	Medium	Senior	**No**
Employed	Fair	Medium	Middle Age	**Yes**
Unemployed	*Excellent*	*High*	*Senior*	***???***

Problem 2

Build a Naïve Bayes classifier using the training data in Problem 1, and make prediction for the approval outcome of the same new application from an unemployed, senior applicant with excellent credit rating and high available credit. Please provide sufficient details on your calculation process of building the classifier, **step-by-step**.

Problem 3

Build *k*-nearest neighbor classifiers for the cases of k=1, k=3, and k=5, using the training data in Problem 1, by converting all features of the data instances into numerical values. You can use any distance function that you prefer, but please define it explicitly. After building these three classifiers, please use them to make prediction for the approval outcome of the same new application from an unemployed, senior applicant with excellent credit rating and high available credit. Please provide sufficient details of your calculation of building the k-nearest neighbors classifiers, **step-by-step**. Please use the following table to convert all features of the instances into numerical values.

Feature	Status	Value
Employment Status	Unemployed	0
	Employed	1
Credit Rating	Excellent	1
	Fair	0
Available Credit	High	2
	Medium	1
	Low	0
Age	Senior	2
	Middle Age	1
	Young	0

Problem 4 (*Please feel free to use any machine learning software package for this problem*)

Use a software package to re-solve the classification problems above. Please use the original data set, and build the following classifiers to predict the approval outcome of a new application from an unemployed, senior applicant with excellent credit rating and high available credit.

(a) Decision tree.
(b) Naïve Bayes.
(c) *k*-nearest neighbor for the cases of k=1, k=3, and k=5.
(d) Support vector machine.

Problem 5

Develop two Support Vector Machine classifiers for the following training data set (16 data examples) using two approaches:

(a) Choose a suitable polynomial kernel to build the SVM classifier;
(b) Use Gaussian kernel (set standard deviation δ to 1) to develop the SVM classifier.

Please provide the functional form of the decision boundary for subproblem (a), and present the predictions of both classifiers to predict the outcome for [1.5 1.5 1.5]. Please explicitly present the optimization problem formulation and optimal coefficients of the classifiers, as well as other technical details.

Note: *In this problem, you are expected to build classifiers based on the algorithms covered in the lectures. Your calculation process should be presented in sufficient details, ideally step-by-step. You could use any basic calculation/computing tools (e.g. NumPy, MATLAB without toolboxes, etc.) to facilitate your calculation process. If only the final classification model and/or results are presented without necessary intermediate steps and results, you will* ***not*** *get most of the credit.*

X_1	X_2	X_3	*Y*
1	1	1	**Yes**
1	1	-1	**Yes**
1	-1	1	**Yes**
-1	1	1	**Yes**
-1	-1	1	**Yes**
-1	1	-1	**Yes**
1	-1	-1	**Yes**
-1	-1	-1	**Yes**
2	2	-2	**No**
2	-2	2	**No**
-2	2	2	**No**
-2	-2	-2	**No**
-2	-2	2	**No**
-2	2	-2	**No**
2	-2	-2	**No**
2	2	2	**No**
1.5	*1.5*	*1.5*	***???***

Problem 6

Use a software package to re-solve Problem 5 (classification via SVM). Please use the same data set and build the following classifiers to predict the outcome for [1.5 1.5 1.5].

(a) Support vector machine with a polynomial kernel.
(b) Support vector machine with a Gaussian kernel.