Gabi Rivera

12Nov2022

ADS502-01

Assignment 3.1: Module 3 Exercise Questions

Introduction to Data Mining: Exercises 4.14

16. You are asked to evaluate the performance of two classification models, M1and M2. The test set you have chosen contains 26 binary attributes, labeled as A through Z. Table 4.13 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, $P(-)=1-P(+)$ and $P(-|A, \ldots, Z)=1-P(+|A, \ldots, Z)$. Assume that we are mostly interested in detecting instances from the positive class.

Table 4.13. Posterior probabilities for Exercise 16.

Instance	True Class	P(+|A, …, Z, M1)	P(+|A, …, Z, M2)
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

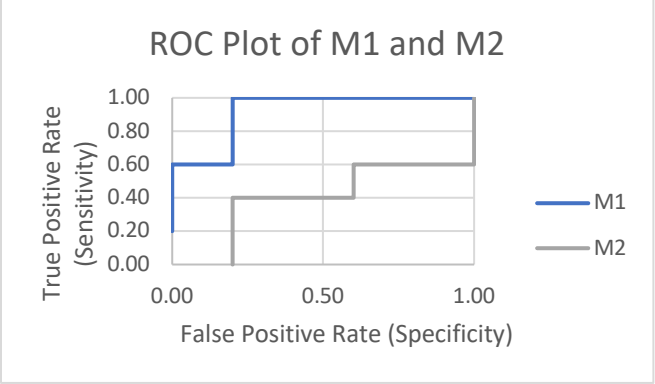
a. Plot the ROC curve for both M1 and M2. (You should plot them on the same graph.)

Which model do you think is better? Explain your reasons.

True Class	M1 Probabilities (+ class)	TP	FP	TPR = TP/P	FPR = FP/N
+	0.73	1	0	0.20	0.00
+	0.69	2	0	0.40	0.00
+	0.67	3	0	0.60	0.00
-	0.55	3	1	0.60	0.20
+	0.47	4	1	0.80	0.20
+	0.45	5	1	1.00	0.20
-	0.44	5	2	1.00	0.40
-	0.35	5	3	1.00	0.60
-	0.15	5	4	1.00	0.80
-	0.08	5	5	1.00	1.00

True Class	M2 Probabilities (+ class)	TP	FP	TPR = TP/P	FPR = FP/N
-	0.68	0	1	0.00	0.20
+	0.61	1	1	0.20	0.20
+	0.45	2	1	0.40	0.20
-	0.38	2	2	0.40	0.40
-	0.31	2	3	0.40	0.60
+	0.09	3	3	0.60	0.60
-	0.05	3	4	0.60	0.80
-	0.04	3	5	0.60	1.00
+	0.03	4	5	0.80	1.00
+	0.01	5	5	1.00	1.00

**0.5 decision threshold*



Answer: M1 is the better model because visually it's area under the curve is larger compared to M2 from looking at the Receiver Operating Characteristic plot.

b. For model M1, suppose you choose the cutoff threshold to be t=0.5. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

Answer: Precision is 75%, Recall is 60%, and F-measure is 67%.

M1 Confusion Matrix, t=0.5 cutoff decision threshold

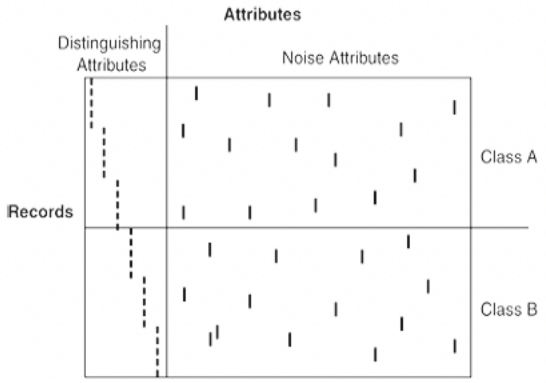
		Predicted	
		Positive	Negative
Actual	Positive	3	2
	Negative	1	4

Precision = True Positive/(False Positive + True Positive) = ¾ = 75%

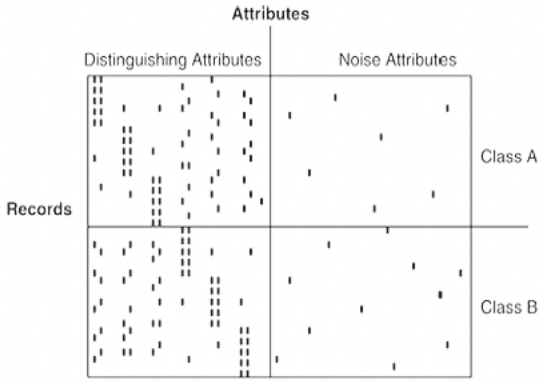
Recall = True Positive Rate (sensitivity) = TP/(TP+FN) = 3/5 = 60%

F-measure = (2*Precision*Recall)/(Recall + Precision) = (2*0.75*0.60)/(0.60*0.75) = 0.67

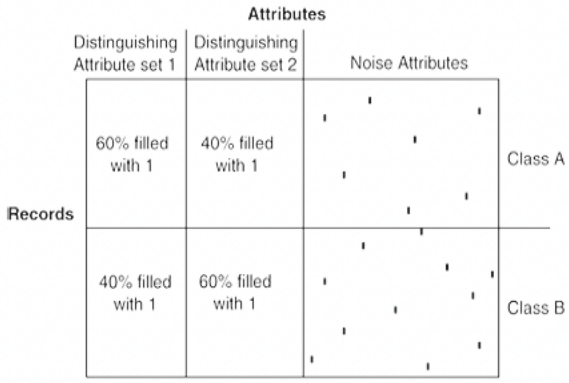
21. Given the data sets shown in Figures 4.59 below, explain how the decision tree, naïve Bayes, and k-nearest neighbor classifiers would perform on these data sets.



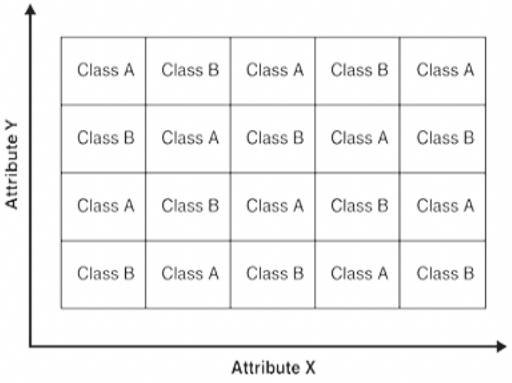
(a) Synthetic data set 1.



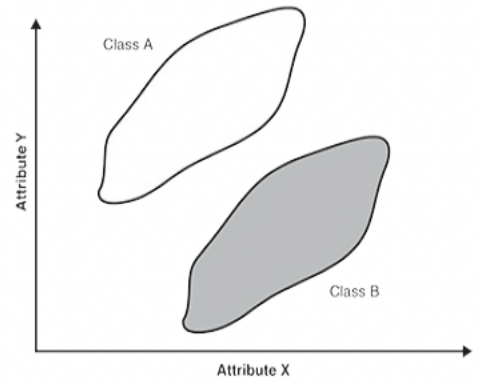
(b) Synthetic data set 2.



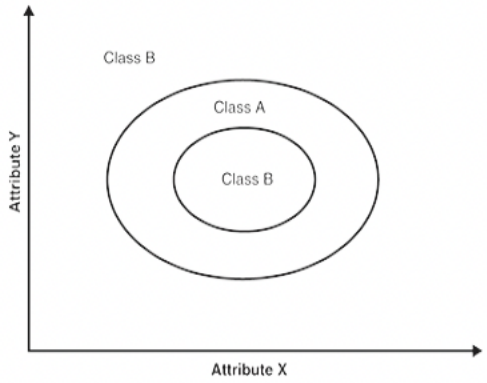
(c) Synthetic data set 3.



(d) Synthetic data set 4.



(e) Synthetic data set 5.



(f) Synthetic data set 6.

Answer:

A. K-NN classifier is sensitive to noise or many interactive attributes so it will not do well with synthetic dataset 1. Naïve Bayes classifiers are robust and can handle noise attributes because they have no impact on the probability estimates. Naïve Bayes will do well on handling synthetic dataset 1. This is the same with decision tree due to entropy gain.
B. Naïve bayes will not do well with synthetic dataset 2 because correlated attributes weaken the performance of this classifier. K-NN and decision tree can handle attributes that are dependent to each other. They will do well with this dataset.
C. Decision tree will not do well because of too many attributes to classify that can cause overfitting. K-NN will do well because it can handle interacting attributes through proximity measures that take account multiple attributes. Naïve bayes will do well too because conditional probability can be used to compare one attribute against the other.
D. Naïve Bayes will not do well because the attributes are dependent on each other. K-NN will do well because it can handle dependent attributes. Decision tree will do well since classes are binary.
E. Same reason with D, decision tree and K-NN will do well but Naïve Bayes will not do well because of dependent attributes.
F. Naïve Bayes will not do well because of dependent attributes. K-NN and decision tree will work well since they can handle attribute dependencies.