Assignment 3.1 [Hand]

University of San Diego

ADS 502

Dingyi Duan

16. You are asked to evaluate the performance of two classification models, M1 and 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, …, Z)=1−P(+|A, …, 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.

- Sort all instances in descending order: (For M_1)
- Pick the largest probability. 0.73.
- Consider 0.73 as positive class and the rest are all negative.
- Compare 0.73 with its actual class, which is positive. Thus, we have $TP + 1 \Rightarrow 1$.
  $FP \Rightarrow 0$.
- Since there are 5 positive, 5 negative,

$\Rightarrow$

	TP	FP	TN	FN
+ 0.73	1	0	5	4

DP

- $TPR = \frac{TP}{TP+FN} = \frac{1}{1+4} = 0.2$.
- $FPR = \frac{FP}{FP+TN} = \frac{0}{0+5} = 0$.
- Apply this to the rest of M_1 and M_2.

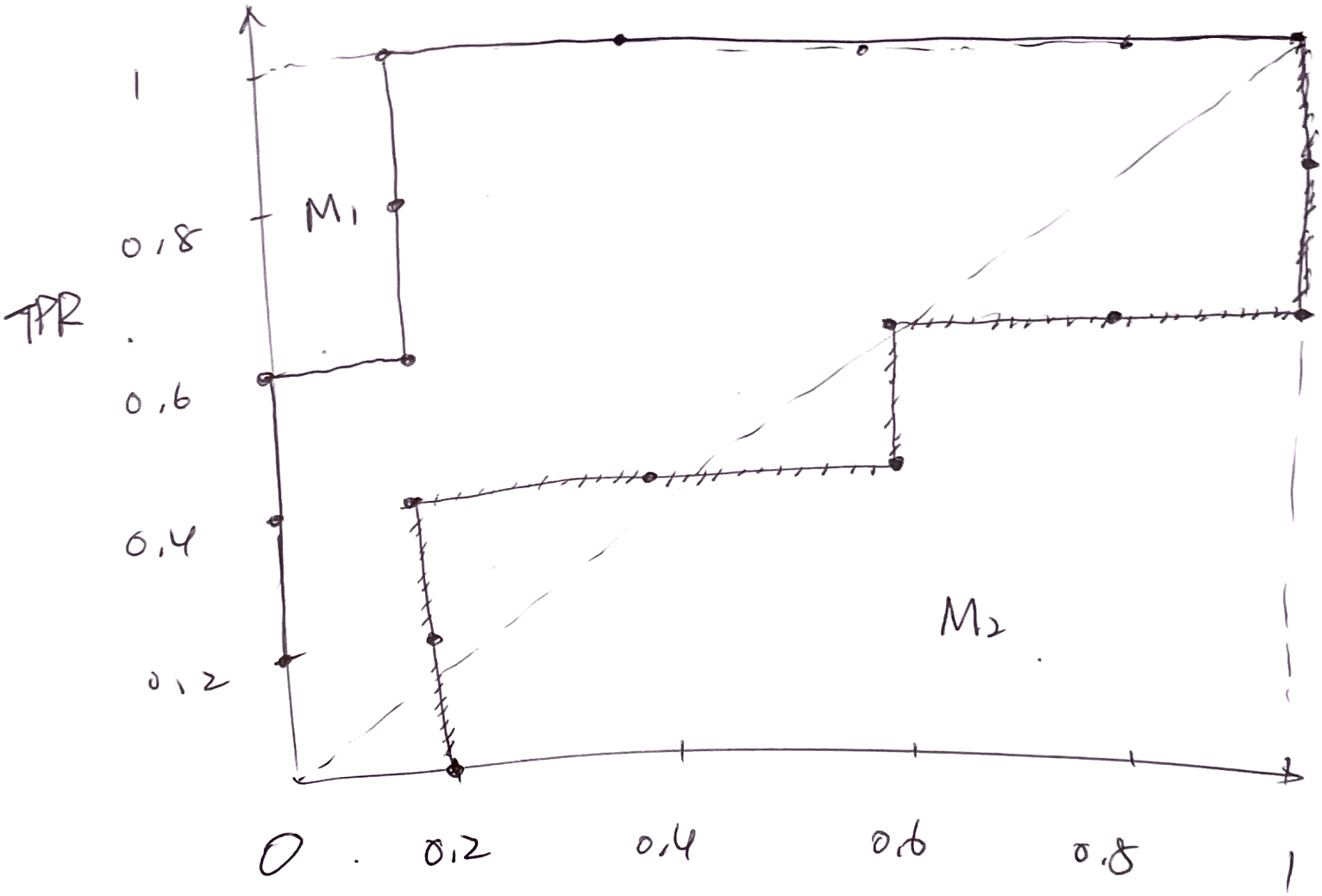
	+	+	+	−	+	+	−	−	−	−
M1:	0.73	0.69	0.67	0.55	0.47	0.45	0.44	0.35	0.15	0.08
TP	1	2	3	3	4	5	5	5	5	5
FP	0	0	0	1	1	1	2	3	4	5
TN	5	5	5	4	4	4	3	2	1	0
FN	4	3	2	2	1	0	5	0	0	0
TPR	0.2	0.4	0.6	0.6	0.8	1	1	1	1	1
FPR	0	0	0	0.2	0.2	0.2	0.4	0.6	0.8	1

	−	+	+	−	−	+	−	−	+	+
M2	0.68	0.61	0.45	0.38	0.31	0.09	0.05	0.04	0.03	0.01
TP	0	1	2	2	2	3	3	3	4	5
FP	1	1	1	2	3	3	4	5	5	5
TN	4	4	4	3	2	2	1	0	0	0
FN	5	4	3	3	3	2	2	2	1	0
TPR	0	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1
FPR	0.2	0.2	0.2	0.4	0.6	0.6	0.8	1	1	1

DD

M_1 is better : the closer the area is to 0.5. the less accurate the model is (M_2).

the closer the area is to 1.0. the more accurat the model is (M_1).



DP

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.

The confusion matrix of M_1.

		Predicted +	Predicted −
Actual	+	3	2
	−	1	4

Predicted + Actual +
0.73. 0.69. 0.67.

Predicted + Actual −
0.55

Predicted − Actual +
0.47 0.45.

Predicted − Actual −
0.44. 0.35, 0.15, 0.08.

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{3}{3+1}$$

$$= \boxed{0.75}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+2}$$

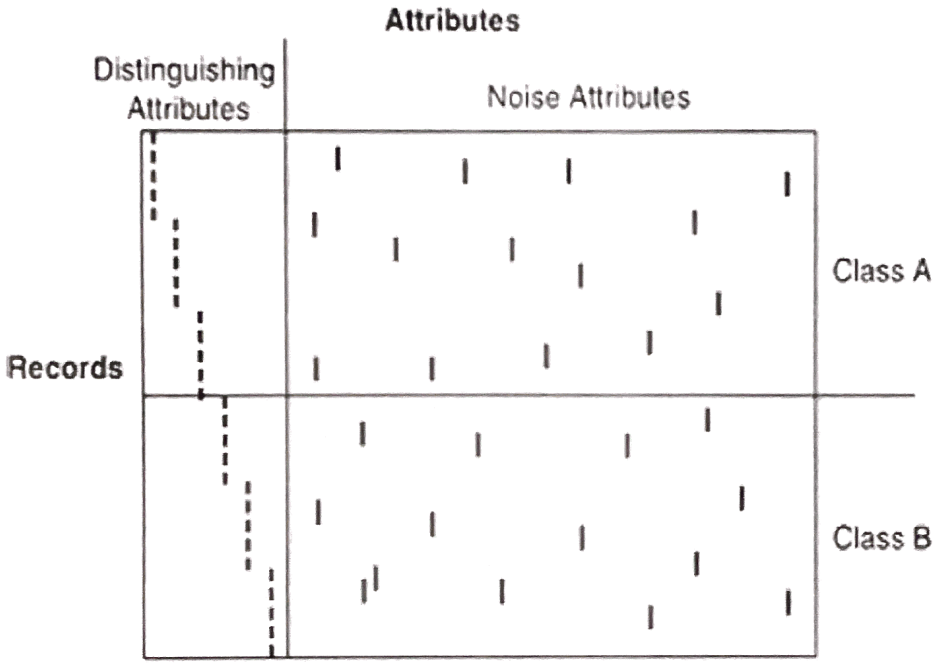
$$= \boxed{0.60}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$= (2 \cdot 0.75 \cdot 0.60) / (0.75 + 0.60)$$

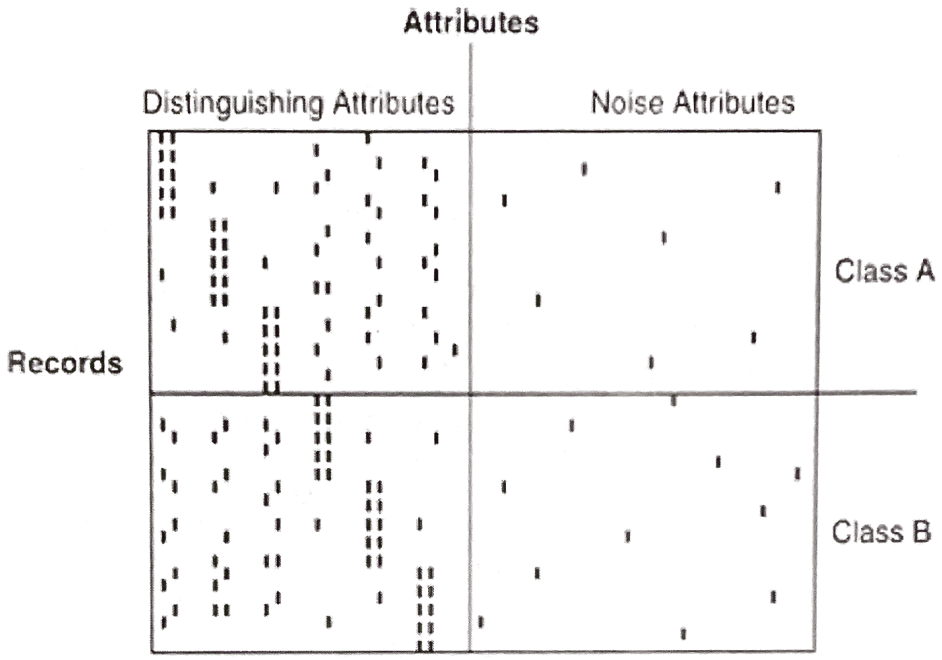
$$\doteq \boxed{0.667}$$

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



(a) Synthetic data set 1.

KNN wouldn't do well because of noise. So DT and Naïve Bayes would perform well due to the distinguishing feature of the samples.

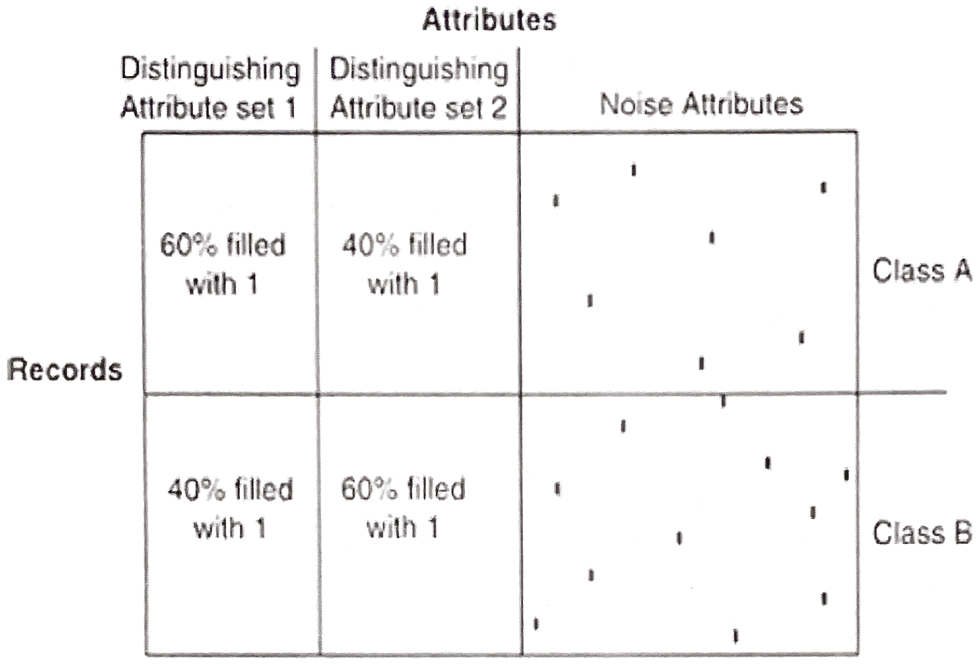


(b) Synthetic data set 2.

KNN wouldn't perform well because of noise.

DT and Naïve Bayes would perform well due to the distinguishness of the samples.

DP

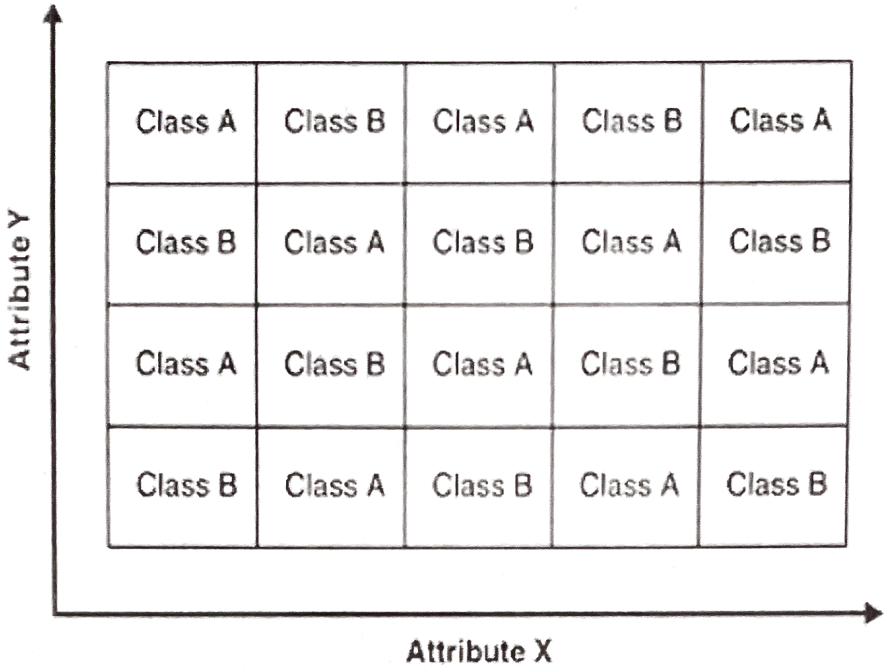


(c) Synthetic data set 3.

KNN wouldn't do well due to the noise.

DT probably wouldn't do well neither due to the mixed "1" in the sample.

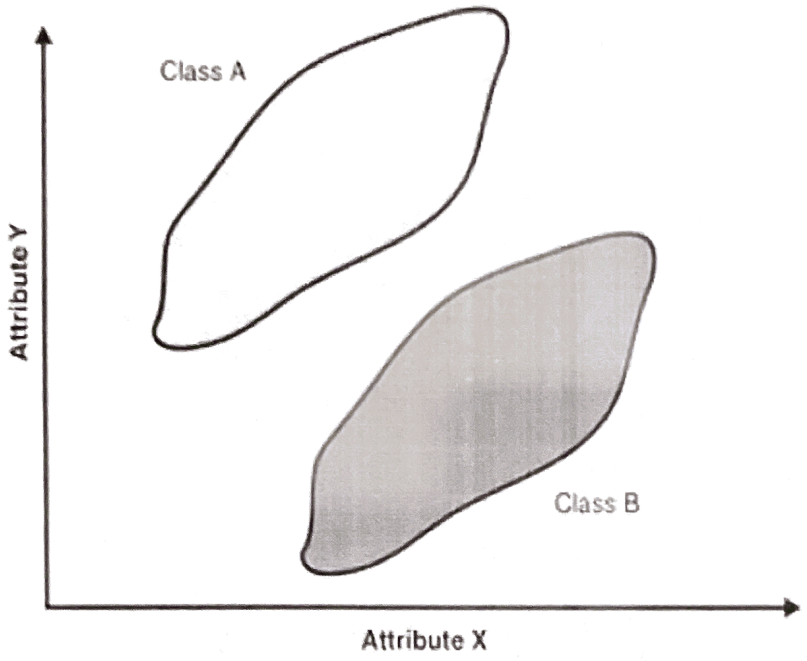
Naïve Bayes would perform well due to the explicit percentages.



(d) Synthetic data set 4.

KNN would perform the best due to the similarities between classes. and no noise.

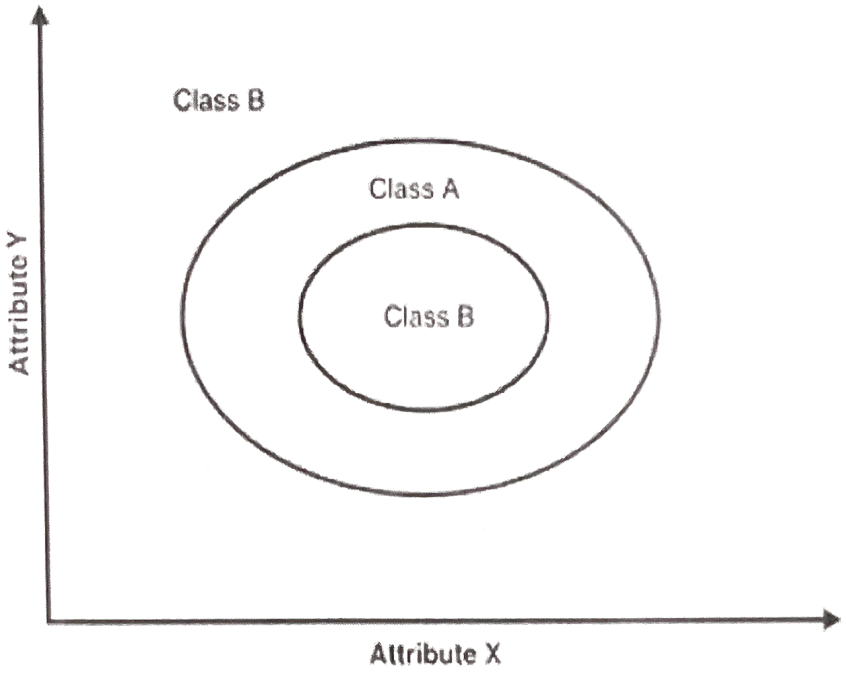
PD



(e) Synthetic data set 5.

KNN would perform the best. due to the similarities of the samples and no noise.

10D



(f) Synthetic data set 6.

KNN would perform the best due to the ~~similarest~~ similarities of the samples and no noise.