Data Science

Problem sheet 2

Ex 1

How many logical (Boolean, $f : \{0,1\}^N \to \{0,1\}$)) functions can be generated on N binary attributes? What are the possible functions for $N = 2$?

Solution

N binary attributes can take 2^N different values. The Boolean function takes value 0 or 1 for each of the $n = 2^N$ inputs. The different values can be assigned in $2^n = 2^{2^N}$ different ways.

For $N = 2$ that is $2^{2^2} = 16$. To illustrate, we show a few of these functions (out of the 16):

X_1	X_2	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	...	f_{16}
0	0	0	1	0	0	0	1	1	1	...	1
0	1	0	0	1	0	0	1	0	0	...	1
1	0	0	0	0	1	0	0	1	0	...	1
1	1	0	0	0	0	1	0	0	1	...	1

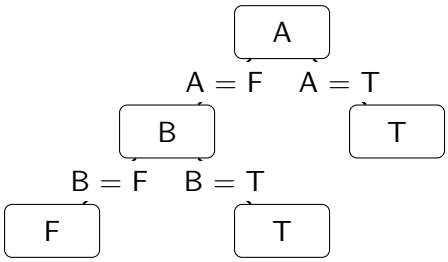
Ex 2

Can decision trees learn logical (Boolean) functions? How to represent the following functions with a decision tree: A **OR** B, A **AND** B, A **XOR** B, where A and B are logical variables.

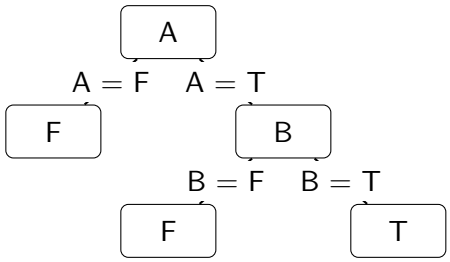
Solution

Yes, they can, see the following examples:

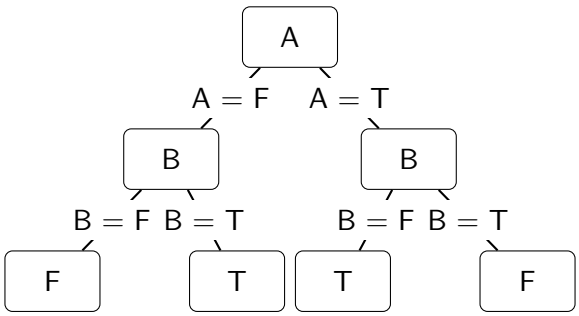
A OR B



A AND B



A XOR B



Ex 3

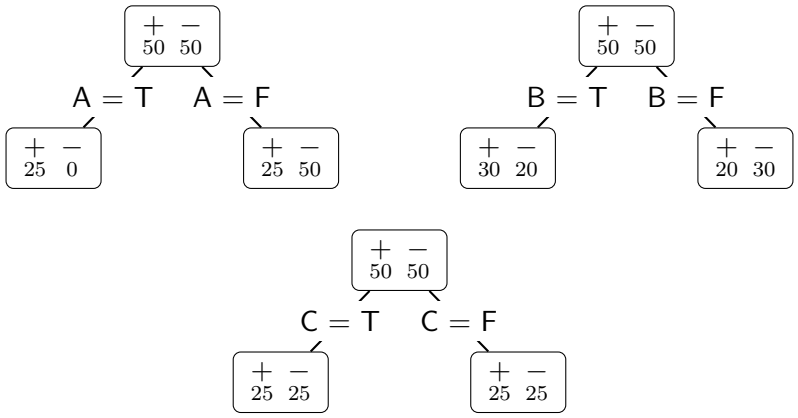
The following table summarizes a data set with three attributes (A, B, C) and two class labels (+, −). Build a two-level decision tree.

1. Use misclassification error as inhomogenity measure. Calculate the gains for each attribute. Which attribute gives the best split?
2. Repeat the previous step for the two children of the root node. Which nodes should be split, and which is the second splitting attribute?
3. Calculate Accuracy, Error rate, Precision, Recall, and F-measure!
4. Choose C as the first splitting attribute and continue building the tree! How a tree of depth 2 would look like in that case?

A	B	C	Number of instances	
			class: +	class: −
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

Solution

1. First let's draw the possible decision trees of depth 1:

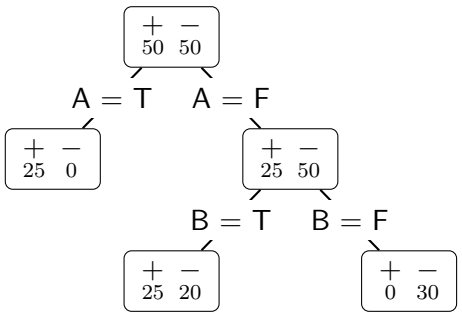


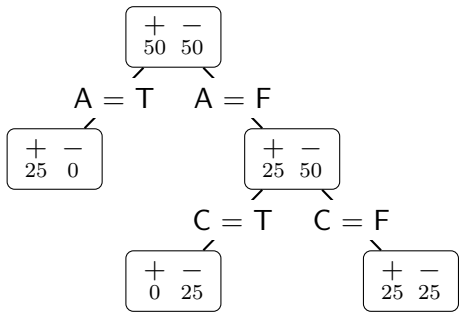
With the help of the figure, we can easily calculate the inhomogenity (misclassification error) of the leaves:

$$\mathrm{I}(\mathit{Root}) = 1 - \max\left(\frac{50}{100}, \frac{50}{100}\right) = \frac{1}{2}$$
$$\mathrm{I}(A = T) = 1 - \max\left(0, \frac{25}{25}\right) = 0$$
$$\mathrm{I}(A = F) = 1 - \max\left(\frac{25}{75}, \frac{50}{75}\right) = \frac{1}{3}$$
$$\mathrm{I}(B = T) = 1 - \max\left(\frac{30}{50}, \frac{20}{50}\right) = \frac{2}{5}$$
$$\mathrm{I}(B = F) = 1 - \max\left(\frac{20}{50}, \frac{30}{50}\right) = \frac{2}{5}$$
$$\mathrm{I}(C = T) = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right) = \frac{1}{2}$$
$$\mathrm{I}(C = F) = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right) = \frac{1}{2}$$
$$\mathrm{Gain}(A) = \mathrm{I}(\mathit{Root}) - \frac{25}{100} \cdot \mathrm{I}(A = T) - \frac{75}{100} \cdot \mathrm{I}(A = F) = \frac{1}{4}$$
$$\mathrm{Gain}(B) = \mathrm{I}(\mathit{Root}) - \frac{50}{100} \cdot \mathrm{I}(B = T) - \frac{50}{100} \cdot \mathrm{I}(B = F) = \frac{1}{10}$$
$$\mathrm{Gain}(C) = \mathrm{I}(\mathit{Root}) - \frac{50}{100} \cdot \mathrm{I}(C = T) - \frac{50}{100} \cdot \mathrm{I}(C = F) = 0$$

From the above calculations, we can conclude that the best first splitting attribute is A.

2. We decided that the first split is on A. Now, it is clear that the node A=T will not be split further since it is totally homogeneous. How to split the other node (A=F), should we split on B or on C?





$$\mathrm{I}(A=F, B=T) = 1 - \max\left(\frac{25}{45}, \frac{20}{45}\right) = \frac{4}{9}$$

$$\mathrm{I}(A=F, B=F) = 1 - \max\left(0, \frac{30}{30}\right) = 0$$

$$\mathrm{I}(A=F, C=T) = 1 - \max\left(0, \frac{25}{25}\right) = 0$$

$$\mathrm{I}(A=F, C=F) = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right) = \frac{1}{2}$$

$$\mathrm{Gain}(B|A=F) = \mathrm{I}(A=F) - \frac{45}{75}\mathrm{I}(A=F, B=T) - \frac{30}{75}\mathrm{I}(A=F, B=F) =$$
$$= \frac{1}{3} - \frac{4}{15} = \frac{1}{15} = 0.06$$

$$\mathrm{Gain}(C|A=F) = \mathrm{I}(A=F) - \frac{25}{75}\mathrm{I}(A=F, C=T) - \frac{50}{75}\mathrm{I}(A=F, C=F) =$$
$$= \frac{1}{3} - \frac{1}{3} = 0$$

The algorithm builds a decision tree of depth 2 that first splits on A, then on B.

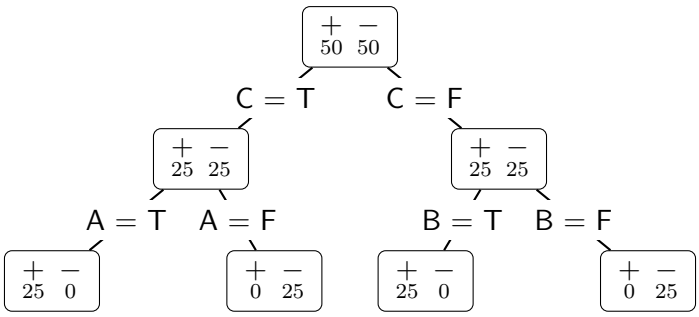
3. We have a tree of depth 2, where the first split is on A and the second split is on B. For this tree we have: TP = 50, FP = 20, TN = 30, and FN = 0.

   Accuracy = $\frac{80}{100}$ Error rate = 1 - Accuracy = $\frac{20}{100}$

   Precision = $\frac{\text{TP}}{\text{TP+FP}} = \frac{50}{70}$ Recall = $\frac{\text{TP}}{\text{TP+FN}} = \frac{50}{50}$

$$F = 2 \cdot \frac{\text{P} \cdot \text{R}}{\text{P} + \text{R}} = 2 \cdot \frac{\frac{5}{7} \cdot 1}{\frac{12}{7}} = 2 \cdot \frac{5}{7} \cdot \frac{7}{12} = \frac{10}{12} = \frac{5}{6} \approx 0.8333$$

4. It is not hard to see that if we fix the first splitting attribute to be C, then the following tree of depth two will be built on the data since splitting the (C=T) node on A and splitting (C=F) node on B results in totally homogeneous leaves:



   We can see that while C is locally the worst first split but globally it turns out to be a pretty good first split if the decision tree algorithm could see a step further. (But it can not, decision tree is a greedy local algorithm and of course - as this example also suggests - not necessarily globally optimal.)

Ex 4

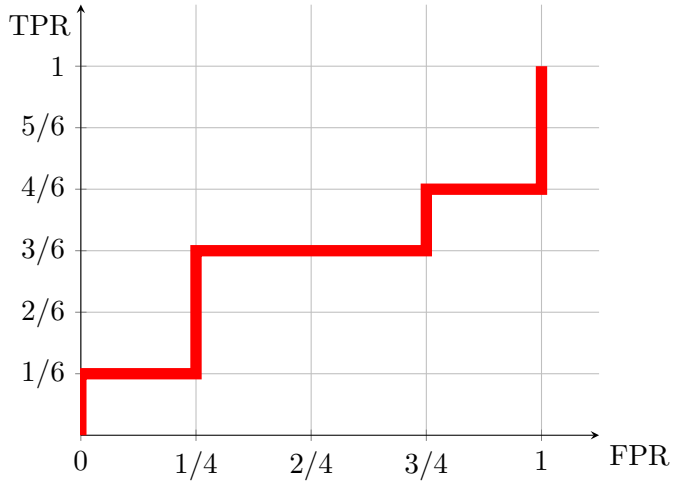
Construct the ROC curve of the following classifier and calculate the AUC. How do you interpret the AUC score? What would you suggest in terms of results? In the table confidence scores increase from left to right.

Label:	+	+	−	+	−	−	+	+	−	+	
TP											
FN											
TN											
FP											
TPR											
FPR											

Solution

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Label:	+	+	−	+	−	−	+	+	−	+	
TP	6	5	4	4	3	3	3	2	1	1	0
FN	0	1	2	2	3	3	3	4	5	5	6
TN	0	0	0	1	1	2	3	3	3	4	4
FP	4	4	4	3	3	2	1	1	1	0	0
TPR	1	$\frac{5}{6}$	$\frac{4}{6}$	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0
FPR	1	1	1	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0



The area of a rectangle is $\frac{1}{24}$ and there are 11 rectangles under the curve, i.e.

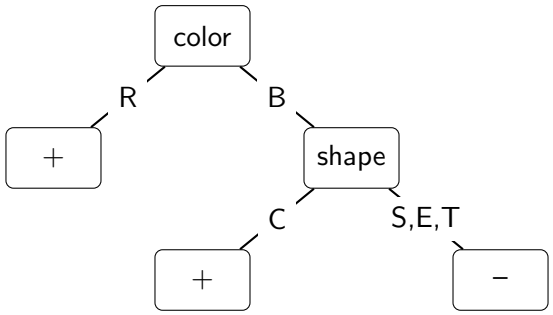
$$\text{AUC} = \frac{11}{24} < 0.5$$

Since the AUC value is < 0.5, the classifier performs worse than a random classifier, so it is worth switching the predicted (positive - negative) labels.

Ex 5

You can see a schematic diagram of a possible decision tree built on training data below.

1. Determine the confidence scores (ratio of positive observations) of the leaves based on the training data (train1, train2, ..., train7).
2. Sort the confidence scores of first three test instances (test1, test2, test3) in ascending order.
3. Construct an ROC curve using the first three instances of the test data (test1, test2, test3) and calculate the AUC.
4. Construct the ROC curve after adding two new test data (test4, test5). If more instances have the same confidence scores ROC curve may change diagonally!

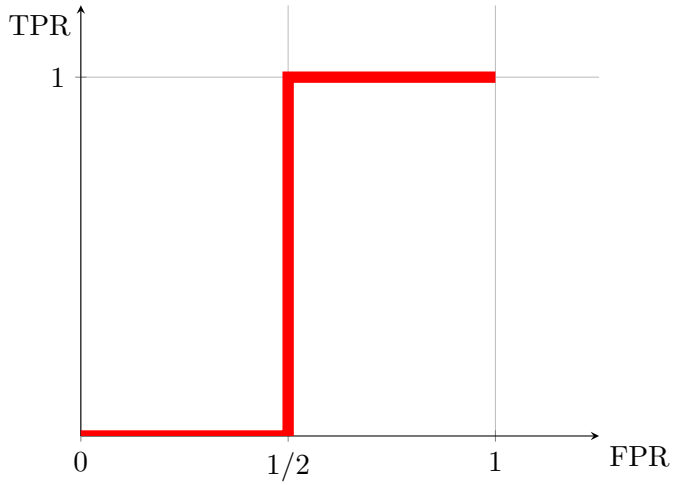


ID	Shape	Color	Size	Class
train1	S	R	L	+
train2	C	R	H	+
train3	C	B	H	+
train4	T	R	L	+
train5	S	B	M	–
train6	E	B	L	–
train7	C	R	M	–

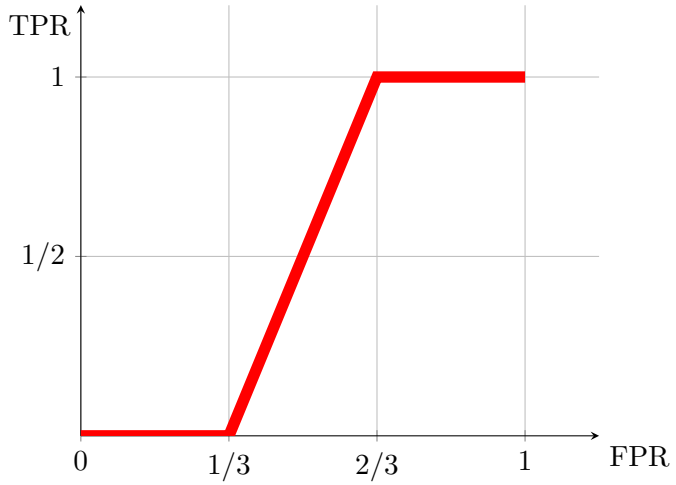
ID	Shape	Color	Size	Class
test1	C	R	H	+
test2	C	B	L	–
test3	E	B	H	–
test4	C	R	L	+
test5	E	R	H	–

Solution

1. The confidence scores corresponding to the leaves (from the left to right in the figure) are respectively: 3/4, 1, 0.
2. Sorting the first three data points with respect to their confidence scores in an ascending order: test3 (0, –), test1 (3/4, +), test2 (1, –).
3. Based on the first three test data points the ROC curve is depicted below. The corresponding AUC value is clearly 1/2.



4. The other two additional test data points (test4, test5) are in the same leaf as test1, so these three data points: test1 (+), test4 (+), and test5 (–) share a common confidence score, that makes the ordering weak, meaning that some data points are tied. So the ROC curve has diagonal segments:



Ex 6

Classify the following record X = (Marital status = Single, Annual income = 90K) using the Naive Bayes classifier based on the training data in the following table, where Default is the class label. Discretize annual income by 20K intervals: $[60K, 80K), [80K, 100K), \ldots$!

1. Use the original estimates!
2. Use Laplace smoothing!

Marital status	**Annual income**	**Default**
Single	125K	No
Married	95K	No
Single	70K	No
Married	120K	No
Divorced	75K	Yes
Married	60K	No
Divorced	220K	No
Single	85K	Yes
Married	75K	No
Single	90K	Yes

Solution

Marital status	Annual income	Discrete income	Default
Married	60K	1	No
Single	70K	1	No
Married	75K	1	No
Divorced	75K	1	Yes
Single	85K	2	Yes
Single	90K	2	Yes
Married	95K	2	No
Married	120K	4	No
Single	125K	4	No
Divorced	220K	9	No

Goal is to maximize the a posteriori estimation:

$$c^* = \arg\max_c \mathbb{P}(X = x|C = c) \cdot \mathbb{P}(C = c)$$

1. X = (Marital status = Single, Income = 90K) = (Single, Discrete income = 2)
   Using the Naive Bayes assumption:

$$\begin{aligned}
\mathbb{P}(X|No) &= \mathbb{P}(Single|No) \cdot \mathbb{P}(2|No) = 2/7 \cdot 1/7 = 2/49 \\
\mathbb{P}(No) &= 7/10 \\
\mathbb{P}(X|No) \cdot \mathbb{P}(No) &= 2/49 \cdot 7/10 = 1/35 \\
\mathbb{P}(X|Yes) &= \mathbb{P}(Single|Yes) \cdot \mathbb{P}(2|Yes) = 2/3 \cdot 2/3 = 4/9 \\
\mathbb{P}(Yes) &= 3/10 \\
\mathbb{P}(X|Yes) \cdot \mathbb{P}(Yes) &= 4/9 \cdot 3/10 = 4/30 \\
\mathbb{P}(X|Yes) \cdot \mathbb{P}(Yes) &> \mathbb{P}(X|No) \cdot \mathbb{P}(No) \Longrightarrow Class = Yes
\end{aligned}$$

2. Using Laplace estimation, the principle is the same, just the estimations are a bit different:

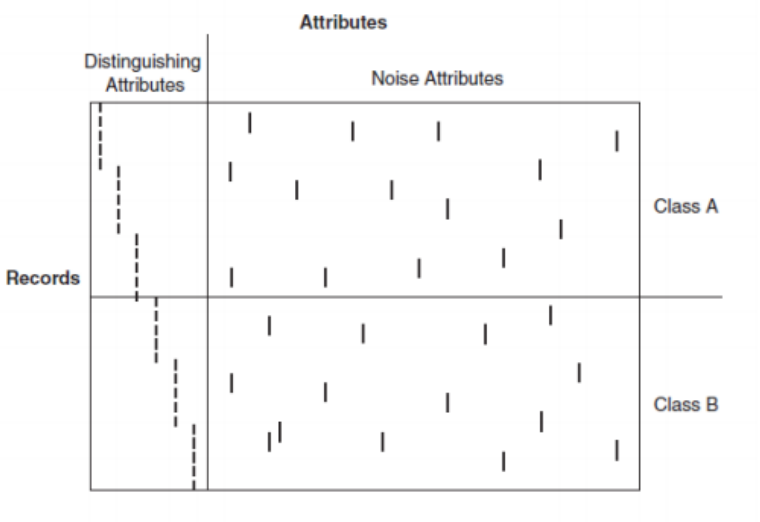
$$\begin{aligned}
\mathbb{P}(X|No) &= 3/9 \cdot 2/9 = 6/81 \\
\mathbb{P}(X|Yes) &= 3/5 \cdot 3/5 = 9/25
\end{aligned}$$

   Again $\mathbb{P}(X|Yes) \cdot \mathbb{P}(Yes) > \mathbb{P}(X|No) \cdot \mathbb{P}(No) \Longrightarrow Class = Yes$

Exercise 7

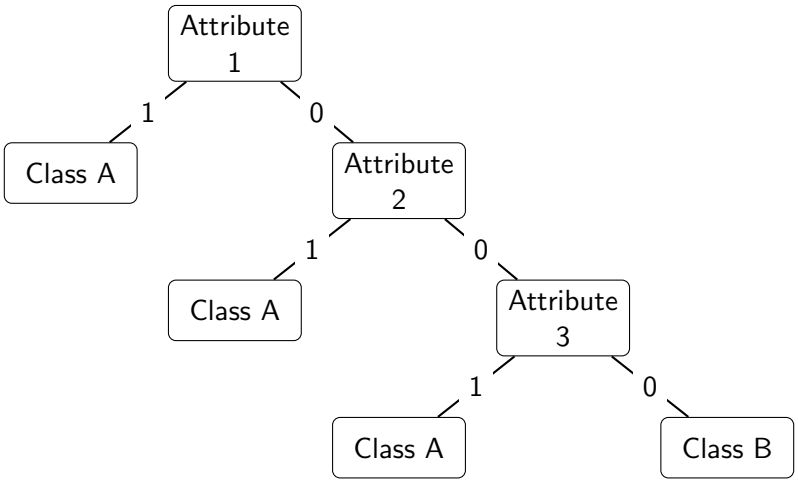
Assume that the following data set contains 1000 records with class label A and 1000 records with label B. There are some binary variables with distinctive power: $X_1, X_2, \ldots,$ in addition, there are many noisy binary attributes, that take value 1 or 0 at random.

1. Sketch a decision tree that learns such data! How would the decision tree classifier perform on this data?
2. Which records are close to the first record? How would the kNN classifier perform on this data?
3. Outline the conditional probabilities! How would the Naive Bayes classifier perform on this data? Consider the first row as an example!



Solution

1. Decision trees are not sensitive to irrelevant attributes. The decision tree will split on the distinguishing attributes, since these are the splits that increase homogeneity. For example a good representation of the data by a decision tree:



2. kNN will not perform well due to relatively large number of noise attributes. kNN is sensitive to irrelevant (noise) attributes. A record from class B may be closer to the first record (first row) because they may agree in some noise attributes due to chance.

3. Naive Bayes will do well on this data set because the distinguishing attributes have better discriminating power than noise attributes in terms of conditional probability. For an example let us consider how the Naive Bayes classifier decides on the first row:

$$\begin{aligned}\mathbb{P}(\text{first row} \,|\, \text{Class A}) &= \mathbb{P}(A_1 = 1 \,|\, \text{Class A}) \cdot \prod_{i=2}^{6} \mathbb{P}(A_i = 0 \,|\, \text{Class A}) \cdot \mathbb{P}(\text{noise} \,|\, \text{Class A}) = \\ &= 1/3 \cdot 2/3 \cdot 2/3 \cdot 1 \cdot 1 \cdot 1 \cdot \mathbb{P}(\text{noise} \,|\, \text{Class A})\end{aligned}$$

$$\text{Moreover } \mathbb{P}(\text{Class A}) = 1/2$$

$$\begin{aligned}\mathbb{P}(\text{first row} \,|\, \text{Class B}) &= \mathbb{P}(A_1 = 1 \,|\, \text{Class B}) \cdot \prod_{i=2}^{6} \mathbb{P}(A_i = 0 \,|\, \text{Class B}) \cdot \mathbb{P}(\text{noise} \,|\, \text{Class B}) = \\ &= \varepsilon \cdot 1 \cdot 1 \cdot 2/3 \cdot 2/3 \cdot 2/3 \cdot \mathbb{P}(\text{noise} \,|\, \text{Class B})\end{aligned}$$

$$\text{Moreover } \mathbb{P}(B) = 1/2$$

In the above the 0 conditional probabilities were replaced by a small ε number (e.g. to mimic Laplace smoothing).

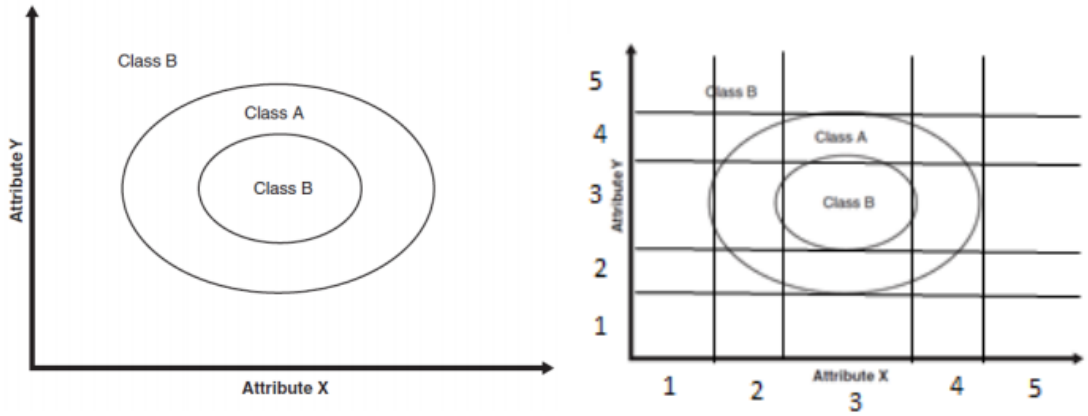
Let us note that $\mathbb{P}(\text{noise} \,|\, \text{Class A}) \approx \mathbb{P}(\text{noise} \,|\, \text{Class B})$, since noise attributes have no distinguishing power.

Therefore: $\mathbb{P}(\text{first row} \mid \text{Class A}) \cdot \mathbb{P}(\text{Class A}) > \mathbb{P}(\text{noise} \mid \text{Class B}) \cdot \mathbb{P}(\text{Class B}) \Longrightarrow$ Class A is assigned (that is the right decision). Using a very similar argument, it can be shown that similarly all the other records are classified correctly.

Ex 8

Consider the following data with two attributes (X and Y) and two possible labels (A and B). The position of class A and class B records in the X-Y space is illustrated below.

1. How would a decision tree work on such data? Indicate decision boundaries!
2. How would the kNN classifier perform on this data? What does its performance depend on?
3. How would the Naive Bayes classifier perform? Outline the conditional probabilities! We assume that the two classes have the same number of records and that the instances are distributed uniformly. Use the possible discretization given below, i.e. both attribute X and Y are discretized into 5 bins!



Solution

1. If the density of the records is high everywhere, then kNN is accurate since the nearest neighbors will belong to the same geometric area.
   If the density is low, kNN will be wrong near the boundaries of the regions.

2. We may have rectilinear boundaries as in the image. We will have error in the small corners around the ellipses. If the density is high, maybe we want to further split the corners. But if the density is low, we are fine, we might not even have records in the corner.

3. We assumed that the two classes have the same number of records so

$$\mathbb{P}(\text{Class A}) \approx \mathbb{P}(\text{Class B})$$

   Let us outline the estimated conditional probabilities

$\mathbb{P}(X = 1 \,\vert\, Class = A) = 0$	$\mathbb{P}(X = 1 \,\vert\, Class = B) \approx 1/4$
$\mathbb{P}(X = 2 \,\vert\, Class = A) \approx 1/3$	$\mathbb{P}(X = 2 \,\vert\, Class = B) \approx 1/8$
$\mathbb{P}(X = 3 \,\vert\, Class = A) \approx 1/3$	$\mathbb{P}(X = 3 \,\vert\, Class = B) \approx 1/4$
$\mathbb{P}(X = 4 \,\vert\, Class = A) \approx 1/3$	$\mathbb{P}(X = 4 \,\vert\, Class = B) \approx 1/8$
$\mathbb{P}(X = 5 \,\vert\, Class = A) = 0$	$\mathbb{P}(X = 5 \,\vert\, Class = B) \approx 1/4$
$\mathbb{P}(Y = 1 \,\vert\, Class = A) = 0$	$\mathbb{P}(Y = 1 \,\vert\, Class = B) \approx 1/4$
$\mathbb{P}(Y = 2 \,\vert\, Class = A) \approx 1/3$	$\mathbb{P}(Y = 2 \,\vert\, Class = B) \approx 1/8$
$\mathbb{P}(Y = 3 \,\vert\, Class = A) \approx 1/3$	$\mathbb{P}(Y = 3 \,\vert\, Class = B) \approx 1/4$
$\mathbb{P}(Y = 4 \,\vert\, Class = A) \approx 1/3$	$\mathbb{P}(Y = 4 \,\vert\, Class = B) \approx 1/8$
$\mathbb{P}(Y = 5 \,\vert\, Class = A) = 0$	$\mathbb{P}(Y = 5 \,\vert\, Class = B) \approx 1/4$

   It is easy to see that if $X = 1, X = 5$ or $Y = 1, Y = 5$, then the record is classified B, since class A has zero probability (see above). So the outside area is classified correctly.
   However, for all other values, A has higher conditional probability than B. Hence the entire inside area is classified A. Hence we misclassify the $(X = 3, Y = 3)$ area. The problem roots in the assumption of conditional independence:

$$0 = \mathbb{P}(X = 3, Y = 3 \,\vert\, Class = A) \neq \mathbb{P}(X = 3 \,\vert\, Class = A)\mathbb{P}(Y = 3 \,\vert\, Class = A) \approx 1/9.$$