Exercises for Chapter 8, Part 2

8.1 Consider the data sets for two classes $X_1 = \{(0,0)\}$ and $X_2 = \{(1,0),(0,1)\}$.

b) Which discriminant line will be found by a support vector machine with linear kernels and infinite penalty for margin violations?

For separating the classes, imagine two parallel lines (or in general hyperplanes). Now we try to position these two lines, so that they separate the classes and at the same time maximize the distance between these lines. If we have only two data points, then the lines will be orthogonal to the difference vector between the two data points, each of the data points will be on one line, and the distance between the lines will be equal to the distance between the two data points. In two dimensions, also for more than two data points, each of the lines will touch at least one data point, one from one class, the other from the other class. For the given data, one line will be through the two points in X_2, so the direction vector is along $(1,-1)$, so the normal vector is along $(1,1)$, so the line is

$$x \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 1 = 0$$

and the other line will be in parallel to this, through the point in X_1

$$x \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0$$

The disriminant line is in the middle of both lines,

$$x \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2} = 0$$

c) Which classification rule will be found by a nearest neighbor classifier?

If we had only the first point in X_2, then the classification rule would be

$$f_1(x) = \begin{cases} 1 & \text{if } x^{(1)} < 0.5 \\ 2 & \text{if } x^{(1)} > 0.5 \\ \text{undefined} & \text{otherwise} \end{cases}$$

If we had only the second point in X_2, then the classification rule would be

$$f_2(x) = \begin{cases} 1 & \text{if } x^{(2)} < 0.5 \\ 2 & \text{if } x^{(2)} > 0.5 \\ \text{undefined} & \text{otherwise} \end{cases}$$

For both points in X_2 we obtain class 1 if both rules vote for 1, so

$$f(x) = \begin{cases} 1 & \text{if } x^{(1)} < 0.5 \text{ and } x^{(2)} < 0.5 \\ 2 & \text{if } x^{(1)} > 0.5 \text{ or } x^{(2)} > 0.5 \\ \text{undefined} & \text{otherwise} \end{cases}$$

d) Which classification rule will be found by a three nearest neighbor classifier?

For any x, all three data points will be the three nearest neighbors, one point from class 1 and two points from class 2, so we get $f(x) = 2$.

e) Draw a receiver operating characteristic for these four classifiers, where class 2 is considered positive.

For this example, the naive Bayes classifier, support vector machine, and nearest neighbor classifier will classify all three data points correctly, so we have TPR=1 and FPR=0, which corresponds to the top left corner of the ROC diagram. The three nearest neighbor classifier will classify both points from class 2 correctly, TP=2, R=2, so TPR=2/2=1, and it will classify the point from class 1 incorrectly, FP=1, I=1, so FPR=1/1=1, which corresponds to the top right corner of the ROC diagram.

8.3 We build a support vector machine classifier for the XOR data set defined by $X_- = \{(-1,-1),(1,1)\}$ and $X_+ = \{(-1,1),(1,-1)\}$.

a) Are the classes of the XOR data set linearly separable?

Obviously not.

b) We use the quadratic kernel $k(x,y) = (xy^T)^2$ and explicitly map the data to $\mathbb{R}^3$, where we use the squares of the original two-dimensional data as the first two dimensions. Determine the formula to compute the third dimension.

In the original two-dimensional space we can write the kernel function as

$$k(x,y) = (xy^T)^2 = (x_1y_1 + x_2y_2)^2 = x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2$$

In the three-dimensional space the first two dimensions are (x_1^2, x_2^2) and (y_1^2, y_2^2), and we want to find the third dimensions x_3 and y_3. The kernel function in the original two-dimensional space is equal to the scalar product in the three-dimensional space

$$(x_1^2, x_2^2, x_3)(y_1^2, y_2^2, y_3)^T = x_1^2y_1^2 + x_2^2y_2^2 + x_3y_3$$

Comparing the two expressions yields

$$x_3 = \sqrt{2}x_1x_2, \quad y_3 = \sqrt{2}y_1y_2$$

c) In the three-dimensional space from (b), determine the normal vector, offset, and margin of the separation plane that the support vector machine will find.

Using the equations from above, in the three-dimensional space we obtain the mappings

$$Y_- = \{(1,1,\sqrt{2}),(1,1,\sqrt{2})\}, \quad Y_+ = \{(1,1,-\sqrt{2}),(1,1,-\sqrt{2})\}$$

so both points in each class fall together. The normal vector is along the difference between the single points from Y_- and Y_+

$$(1, 1, \sqrt{2}) - (1, 1, -\sqrt{2}) = (0, 0, 2\sqrt{2})$$

which we normalize to $w = (0, 0, 1)$. The separation plane is through the mean of these single points

$$((1, 1, \sqrt{2}) + (1, 1, -\sqrt{2}))/2 = (1, 1, 0)$$

so the offset is $b = -(0, 0, 1)^T \cdot (1, 1, 0) = 0$. The margin is the length of the distance between the single points

$$\|(1, 1, \sqrt{2}) - (1, 1, -\sqrt{2})\| = 2\sqrt{2}$$

8.4 A car manufacturer has produced 50.000 cars with 90 kW gasoline engine, 100.000 cars with 120 kW gasoline engine, and 50.000 cars with 90 kW diesel engine. For all calculations use the following assumptions: $-\frac{1}{4}\log_2\frac{1}{4} = \frac{1}{2}$, $-\frac{1}{3}\log_2\frac{1}{3} = \frac{19}{36}$, $-\frac{1}{2}\log_2\frac{1}{2} = \frac{1}{2}$, $-\frac{2}{3}\log_2\frac{2}{3} = \frac{14}{36}$.

a) What is the entropy of these data?

We have a total of $50.000 + 100.000 + 50.000 = 200.000$ objects. For the three classes we have the probabilities $50.000/200.000 = 1/4$, $100.000/200.000 = 1/2$, and $50.000/200.000 = 1/4$. So the entropy is

$$H = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} = 1.5 \text{ bit}$$

b) For a given car, how much information do you gain if you know the kW value?

We have $50.000 + 50.000 = 100.000$ objects with 90kW. For gasoline and diesel we then have the probabilities $50.000/100.000 = 1/2$ each, so

$$H(90\text{kW}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \text{ bit}$$

We have 100.000 objects with 120kW, all gasoline, so we have the probability $100.000/100.000 = 1$, so

$$H(120\text{kW}) = 0 \text{ bit}$$

The probability for 90kW is $(50.000 + 50.000)/200.000 = 1/2$, and the probability for 120kW is $100.000/200.000 = 1/2$, so the expected value is

$$\frac{1}{2} \cdot 1 \text{ bit} + \frac{1}{2} \cdot 0 \text{ bit} = 0.5 \text{ bit}$$

and the expected information gain is

$$1.5 \text{ bit} - 0.5 \text{ bit} = 1 \text{ bit}$$

c) For a given car, how much information do you gain if you know whether it has a gasoline or diesel engine?

We have $50.000 + 100.000 = 150.000$ objects with gasoline. For 90kW we then have the probability $50.000/150.000 = 1/3$, and for 120kW we have $100.000/150.000 = 2/3$, so

$$H(\text{gasoline}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx \frac{19}{36} + \frac{14}{36} = \frac{33}{36} = \frac{11}{12}$$

We have 50.000 objects with diesel, all 90kW, so we have the probability $50.000/50.000 = 1$, so

$$H(\text{diesel}) = 0 \text{ bit}$$

The probability for gasoline is $(50.000 + 100.000)/200.000 = 3/4$, and the probability for diesel is $50.000/200.000 = 1/4$, so the expected value is

$$\frac{3}{4} \cdot \frac{11}{12} \text{ bit} + \frac{1}{4} \cdot 0 \text{ bit} = \frac{11}{16} \text{ bit}$$

and the expected information gain is

$$\frac{24}{16} \text{ bit} - \frac{11}{16} \text{ bit} = \frac{13}{16} \text{ bit} < 1 \text{ bit}$$

d) Sketch the ID3 decision tree.

From the root of the decision tree we first ask for kW. If it is 120kW, then we terminate. And if it is 90 kW, then we next ask for gasoline or diesel.