University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning
Spring term 2022

Exercise 6

Exercise 6-1 Conditional Probability (1 point)
Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include an extra battery, and 30% include both a card and a battery. Consider randomly selecting a buyer and let $A = \{\text{memory card purchased}\}$ and $B = \{\text{battery purchased}\}$. Then $\Pr(A) = 0.6$, $\Pr(B) = 0.4$, and $\Pr(A \cap B) = 0.3$.

1. Given that the selected individual purchased an extra battery, what is the probability that an optional card was also purchased?

   Suggested solution:

   $$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{0.3}{0.4} = 0.75$$

2. Given that the selected individual purchased a memory card, what is the probability that an optional extra battery was also purchased?

   Suggested solution:

   $$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{0.3}{0.6} = 0.5$$

Exercise 6-2 Bayes' Theorem (1 point)

Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time.

If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

Suggested solution:

Bayes' Theorem:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

We have the following events:

- A^+ - Actual positive / has disease
- A^- - Actual negative / healthy
- T^+ - Test is positive
- T^- - Test is negative

We are given the following probabilities:

- $\Pr(A^+) = \frac{1}{1000}$
- $\Pr(A^-) = \frac{999}{1000}$
- $\Pr(T^+|A^+) = 0.99$
- $\Pr(T^+|A^-) = 0.02$

Total probability of positive test:

$$\begin{aligned}\Pr(T^+) &= \Pr(T^+|A^+)\Pr(A^+) + \Pr(T^+|A^-)\Pr(A^-) \\ &= 0.99 \cdot 0.001 + 0.02 \cdot 0.999 \\ &= 0.02097\end{aligned}$$

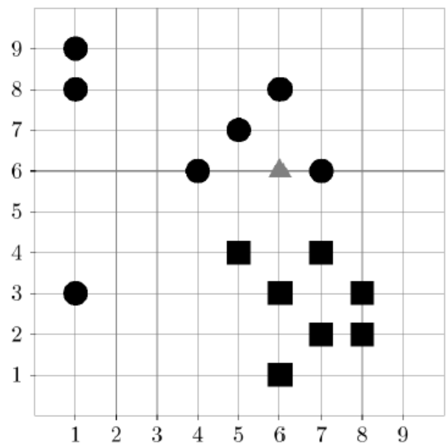
We want to calculate $\Pr(A^+|T^+)$:

$$\begin{aligned}\Pr(A^+|T^+) &= \frac{\Pr(T^+|A^+)\Pr(A^+)}{\Pr(T^+)} \\ &= \frac{0.00099}{0.02097} \\ &= 0.0472\end{aligned}$$

Exercise 6-3 Nearest Neighbour Classification (1 point)

The 2D feature vectors in the figure below belong to two different classes (circles and rectangles). Classify the object at $(6,6)$ - in the image represented using a triangle - using k nearest neighbor classification. Use Manhattan distance (L_1 norm) as distance function, and use the non-weighted class counts in the k-nearest-neighbor set, i.e. the object is assigned to the majority class within the k nearest neighbors. Perform kNN classification and compare the results with your own "intuitive" result for the following k values.

- $k = 4$
- $k = 7$
- $k = 10$

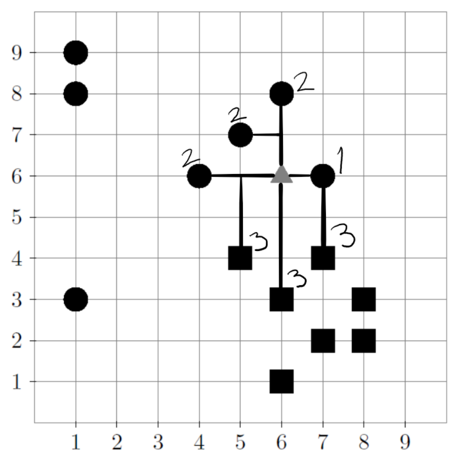


Suggested solution:

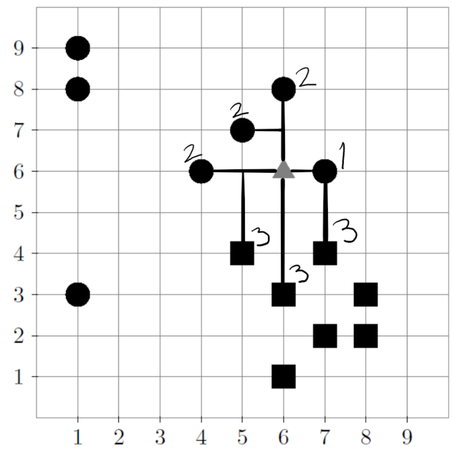
For $k = 4$:

Looking at the closest neighbours, we see that all 4 nearest neighbours are circles.

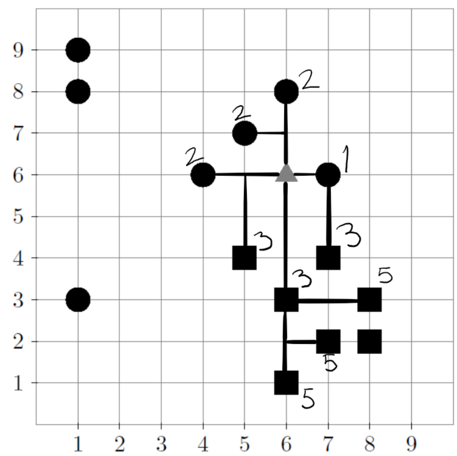
The object at $(6,6)$ would be classified as a circle.



For $k = 7$:
Looking at the closest neighbours, we see that $\frac{4}{7}$ of the nearest neighbours are circles.
The object at $(6, 6)$ would be classified as a circle.

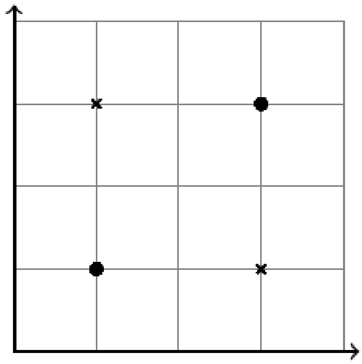


For $k = 10$:
Looking at the closest neighbours, we see that $\frac{4}{10}$ of the nearest neighbours are circles. We also see that $\frac{6}{10}$ of the nearest neighbours are squares.
The object at $(6, 6)$ would be classified as a square.



Exercise 6-4 Nearest Neighbour Classification (1 point)
Find a scenario where we have a set of at least four points in 2 dimensions, such that the Nearest Neighbor classification ($k = 1$) only gives incorrect classification results when using any of these points as query point and the rest as training examples. Use Euclidean distance as distance function.

Suggested solution:
Various solutions are possible, e.g.:



Exercise 6-5 Linearity of Expectation and Variance (1 point)
Suppose that the two variables x and y are statistically independent. Show that the mean and variance of their sum satisfies:

$$E(x + y) = E(x) + E(y)$$
$$\mathrm{var}(x + y) = \mathrm{var}(x) + \mathrm{var}(y)$$

Suggested solution:
Proof for Linearity of Expectation:
Recall definition of expected value over a discrete random variable x, where x_i are the possible values of x.

$$E(x) = \sum_i x_i \cdot \Pr(x_i)$$

Probability of sum of variables:

$$\begin{aligned}
E(x + y) &= \sum_i \sum_j [(x_i + y_j) \cdot \Pr(x = x_i, y = y_j)] \\
&= \sum_i \sum_j \{x_i \cdot \Pr(x = x_i, y = y_j)\} + \sum_i \sum_j \{y_j \cdot \Pr(x = x_i, y = y_j)\} \\
&= \sum_i x_i \sum_j \Pr(x = x_i, y = y_j) + \sum_j y_j \sum_i \Pr(x = x_i, y = y_j) \\
&= \sum_i x_i \Pr(x = x_i) + \sum_j y_j \Pr(y = y_j) \\
&= E(x) + E(y)
\end{aligned}$$

We never assume variables are independent, so this proof also works for dependent variables.

Note: For continuous random variables, the proof is the same, but with integrals rather than sums. The proof can be extended to an arbitrary number of variables by induction.

Variance follows a similar proof. Recall definition of variance over a random variable:

$$\begin{aligned}
\mathrm{Var}(x) &= E(X - E(X))^2 \\
&= E(X^2) - E(X)^2 \\
&= E((x + y)^2) - E(x + y)^2
\end{aligned}$$

We use the linearity of expectation for this proof:

$$\begin{aligned}
\mathrm{Var}(x) &= E((x + y)^2) - E(x + y)^2 \\
&= E(x^2 + y^2 + 2xy) - E(x + y)^2 \\
&= E(x^2) + E(y^2) + E(2xy) - (E(x) + E(y))^2 \\
&= E(x^2) + E(y^2) + E(2xy) - (E(x)^2 + E(y)^2 + 2E(x)E(y)) \\
&= E(x^2) + E(y^2) - E(x)^2 - E(y)^2 \\
&= \mathrm{Var}(x) + \mathrm{Var}(y)
\end{aligned}$$