Solutions

Exercise 6 : Conditional Probability and Bayes' Theorem, k-Nearest Neighbor Classification

Exercise 6-1 : Conditional Probability

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 battery. Consider randomly selecting a buyer and let A ={memory card purchased} and B = {battery purchased}.
Then $\Pr(A) = 0.6$, $\Pr(B) = 0.4$, and $\Pr(\text{both purchased}) = \Pr(A \cap B) = 0.3$.

(a) Given that the selected individual purchased an extra battery, what is the probability that an optional card was also purchased ?

Suggested solution :
$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\frac{3}{10}}{\frac{4}{10}} = 0.75$

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

Suggested solution :
$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\frac{3}{10}}{\frac{6}{10}} = 0.5$

Exercise 6-2 : Bayes' Theorem

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 :

We have the following events :

- D : has disease
- H : healthy
- P : test is positive
- N : test is negative

We know already about the following probabilities :

- $\Pr(D) = \frac{1}{1000}$
- $\Pr(H) = \frac{999}{1000}$
- $\Pr(P|D) = 0.99$
- $\Pr(P|H) = 0.02$

So we can compute the answer using Bayes' theorem :

$$\begin{aligned}\Pr(D|P) &= \frac{\Pr(P|D)\Pr(D)}{\Pr(P|D)\Pr(D)+\Pr(P|H)\Pr(H)} \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} \\ &= \frac{0.00099}{0.02097} \\ &= 0.0472103\end{aligned}$$

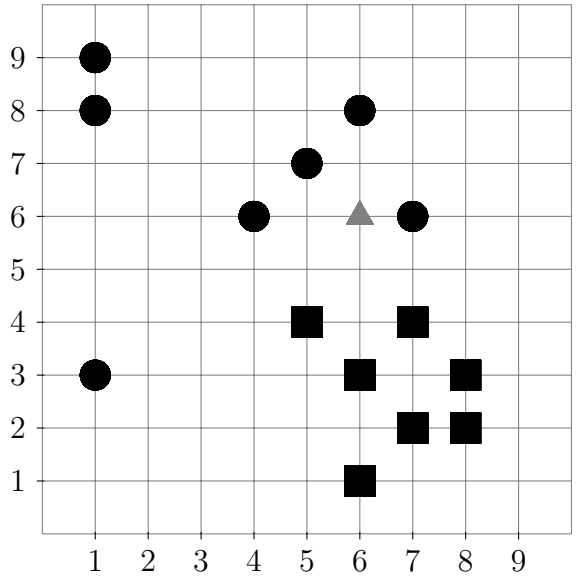
Exercise 6-3 : Nearest neighbor classification

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 for the following values of k and compare the results with your own "intuitive" result.

(a) $k = 4$

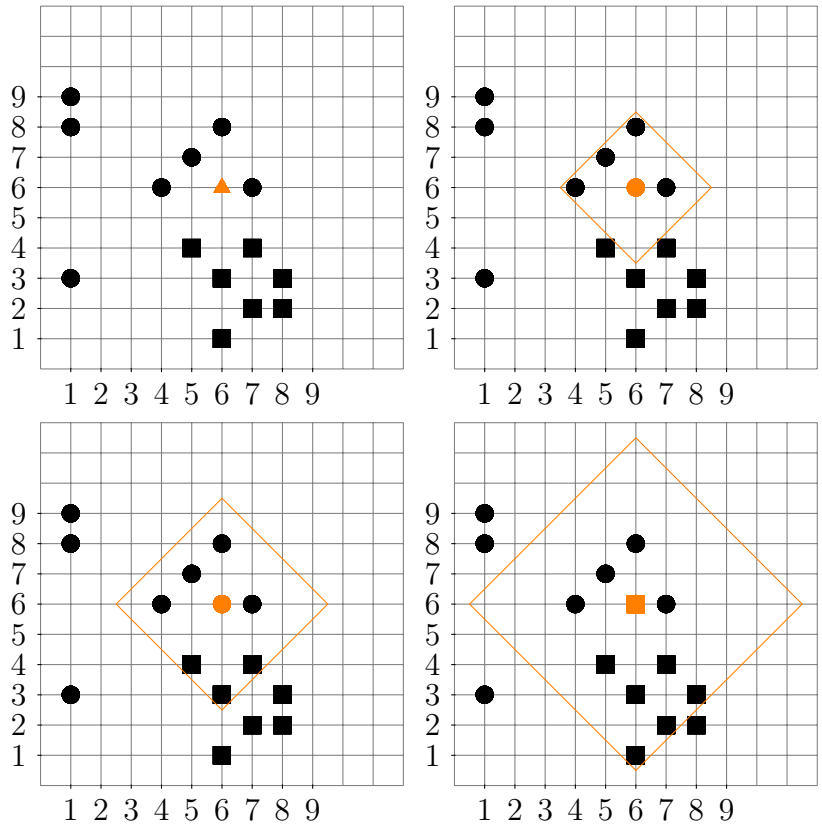
(b) $k = 7$

(c) $k = 10$



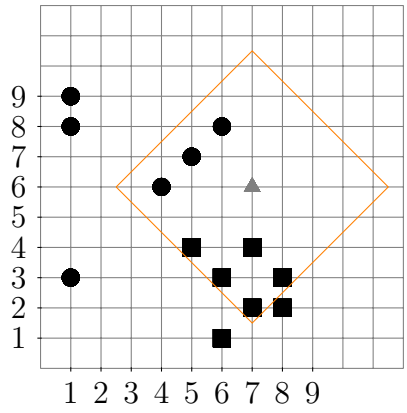
Suggested solution :

Suggestion : draw the equal distance spheres (Manhattan distance !) on the board or on a transparent over the dataset given on the exercise sheet :



Discuss why $k = 7$ might already be too large (e.g., class "circles" contains only 7 objects overall, the object at $(6, 6)$ is close to the class border).

Consider the object at $(7, 6)$, given as example for "circle". How would it be classified with $k = 7$ if it were a test object ?.



Note that we would have 8 neighbors here due to ties in the distance.

Exercise 6-4 : Nearest Neighbor classification

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 points and the rest as training examples. Use Euclidean distance as distance function.

Suggested solution :

Various solutions are possible, e.g. :