

7.1 Suppose you expect the Bayes boundary to be highly irregular. What kind of values for k will accommodate highly irregular Bayes boundaries? Why?

7.1 Highly irregular Bayes boundary mean that the data is complex and likely nonlinear.

We should use small k as small k 's make the boundary wiggly and twisted

7.2 Which of the following statements are true?

- (a) kNN with the Euclidean distance treats all variables as equally important.
- (b) kNN works best when all variables are on the same scale.
- (c) If k is too large, kNN is very sensitive to noise.
- (d) kNN deteriorates when there are too many noise variables.

7.2 ① True ② True ③ False ④ True

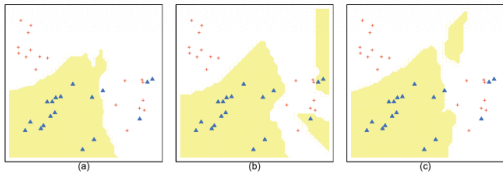


Fig. 7.7 Three kNN fits for different values for k

7.3 Lowest $\rightarrow$ Highest
b $\rightarrow$ a $\rightarrow$ c (smoothest)

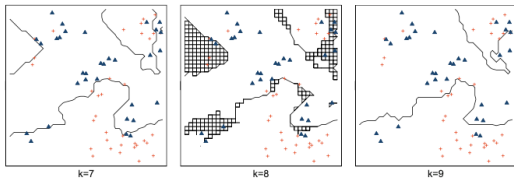


Fig. 7.8 Three kNN fits for different values for k

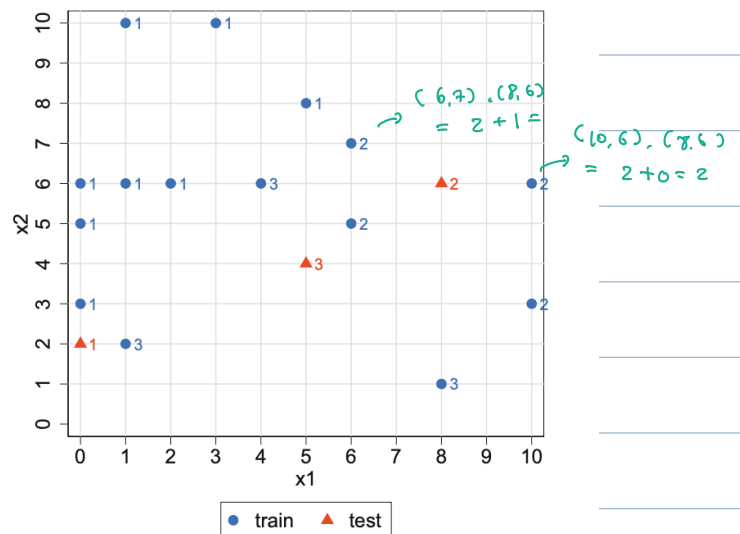
7.4 Because there are ties in classification when $k=8$.

7.3 Figure 7.7 shows three different kNN fits using different values for k . Order the three graphs in ascending order of k .

7.4 Figure 7.8 shows the boundary for 3 kNN fit for $k = 7, 8, 9$. The boundaries are computed from predictions on a 40 by 40 grid. Why does the boundary look messy (showing the underlying grid) for the graph in the middle?

7.5 A small data set has 18 observations and 3 classes ($y = 1, 2, 3$) and 2 x -variables (x_1, x_2). The 18 observations were divided into 15 observations in the training data (blue circles) and three test data observations (red triangles). The data are plotted in Fig. 7.9. Each point is labeled with its class ($y = 1, 2, 3$). The coordinates of all points are integers. The Manhattan distance between two vectors (or points) x_1 and x_2 is defined as $\sum_{i=1}^2 |x_{1i} - x_{2i}|$ over the dimensions of the vectors. This is also known as the **L1 norm**. The three test observations are labeled as follows: test obs 1: $x = (0, 2)$, test obs 2: $x = (5, 4)$, and test obs 3: $x = (8, 6)$. Tie breaking rule: whenever there is a tie, break the tie by increasing k until there is a clear winner.

- (a) Using kNN with $k = 1$, classify the three test observations. Each time list the coordinates of the corresponding nearest neighbor. If there is a tie, give the coordinates of the point that breaks the tie. What is the classification accuracy?
- (b) Using kNN with $k = 3$, classify the three test observations. Each time give the vote across the three classes. For example, 2:1:0 means among the three neighbors $y = 1$ occurred twice, $y = 2$ once, and $y = 3$ did not occur.
- (c) The previous two parts were about classification on the test data. This part is about classification on the training data. For kNN with $k = 1$, how is the point with coordinates (4,6) classified?
- (d) This part is about LOO cross-validation. For kNN with $k = 1$, how is the point with coordinates (4,6) classified using LOO cross-validation?



Manhattan Distance: $\sum_{i=1}^2 |x_{1i} - x_{2i}| \sim d = |x_1 - x_2| + |y_1 - y_2|$

① When $k=1$, $x_1 (0,2) = \text{Tie}$, $(0,5)$ breaks the tie into class 2

$x_2 (5,4) = 2$, $x_3 (8,6) = 2$

classification accuracy = $\frac{2}{3} = 0.67$

⑥ When $k=3$, $x_1 (0,2) = 1$, $x_2 (5,4) = 2$, $x_3 (8,6) = 2$

⑦ Using the full data set, the nearest neighbor of $(9,6)$ is itself, class 3

⑧ Loc. $k=1$. This removes the $(9,6)$ point itself. Then it is classified as 1.