# $k$ nearest neighbours

Computing Workshop: Software

January 19, 2019

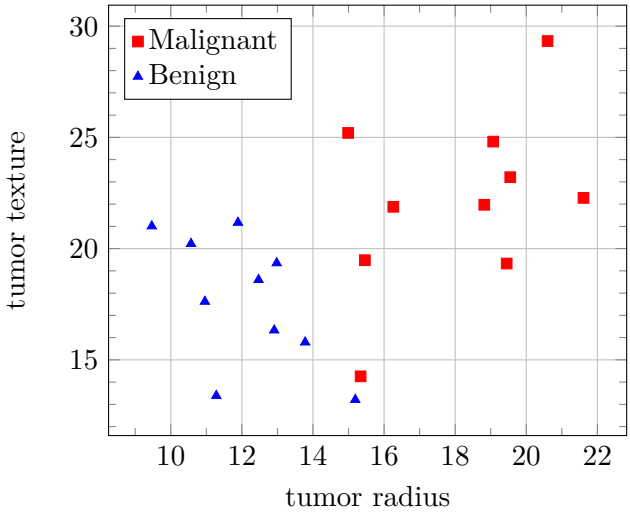


Figure 1: This plot shows 20 data points from the UCI Breast Cancer Dataset. The dataset contains measurements from people with breast tumors, some benign and some malignant. Malignant tumors are dangerous, so it's important to be able to predict whether a tumor might be malignant. The full dataset contains 10 different measurements per person, so the data is 10-dimensional! To make things printable on a sheet of paper, we focus on only two measurements: the tumor radius ($x$ axis) and the tumor texture ($y$ axis).Malignant tumors are shown as squares (red) and benign tumors are shown as triangles (blue).

## Questions

Suppose we want to classify a new patient $p$ whose measurements are tumor radius = 15 and tumor texture = 18.

1. By visual inspection of figure 1, what classification would you give the point $p$?

2. By using the table in figure 2, decide on a classification using the $k$ nearest neighbours of $p$ for $k = 1, 2, 3, 5$. For example, if the majority of the nearest neighbors are malignant, then classify $p$ as malignant.

| ID | Diagnosis | Radius | Texture | Distance |
|---|---|---|---|---|
| 927241 | M | 20.6 | 29.33 | 12.64 |
| 90944601 | B | 13.78 | 15.79 | 2.52 |
| 875878 | B | 12.91 | 16.33 | 2.68 |
| 87930 | B | 12.47 | 18.6 | 2.6 |
| 901315 | B | 10.57 | 20.22 | 4.96 |
| 896864 | B | 12.98 | 19.35 | 2.43 |
| 855133 | M | 14.99 | 25.2 | 7.2 |
| 855625 | M | 19.07 | 24.81 | 7.93 |
| 868871 | B | 11.28 | 13.39 | 5.92 |
| 895633 | M | 16.26 | 21.88 | 4.08 |
| 8511133 | M | 15.34 | 14.26 | 3.76 |
| 903516 | M | 21.61 | 22.28 | 7.87 |
| 8670 | M | 15.46 | 19.48 | 1.55 |
| 859464 | B | 9.47 | 21.01 | 6.3 |
| 90312 | M | 19.55 | 23.21 | 6.92 |
| 886226 | M | 19.45 | 19.33 | 4.64 |
| 9012568 | B | 15.19 | 13.21 | 4.79 |
| 919537 | B | 10.96 | 17.62 | 4.06 |
| 905686 | B | 11.89 | 21.17 | 4.44 |
| 908445 | M | 18.82 | 21.97 | 5.51 |

Figure 2: The table of data used to generate the preceding scatter plot. $M$ indicates a malignant diagnosis and $B$, benign. Included in this table is also the distance of each row to the new patient $p$.

## More Questions

Suppose we use a 3 nearest neighbors algorithm to classify 20 points as either benign or malignant. The table below lists the points to classify as well as their predicted classification and actual classification.

1. Fill in the *confusion matrix* table below to contrast the classification errors made by this classifier.
2. Discuss in your groups the consequences of false positives and false negatives in the context of breast cancer screening. Are some errors more dangerous than others?

| Radius | Texture | Actual | Prediction |
|---|---|---|---|
| 15.28 | 22.41 | M | B |
| 16.74 | 21.59 | M | B |
| 12.36 | 21.8 | B | B |
| 14.4 | 26.99 | B | B |
| 23.51 | 24.27 | M | M |
| 8.22 | 20.7 | B | B |
| 12.47 | 18.6 | B | B |
| 15.37 | 22.76 | M | B |
| 12.43 | 17 | B | B |
| 9.79 | 19.94 | B | B |
| 17.19 | 22.07 | M | B |
| 9.85 | 15.68 | B | B |
| 11.52 | 14.93 | B | B |
| 11.67 | 20.02 | B | B |
| 12.3 | 15.9 | B | B |
| 14.8 | 17.66 | B | M |
| 13.17 | 21.81 | M | B |
| 13.37 | 16.39 | B | M |
| 12.06 | 12.74 | B | B |
| 12.36 | 21.8 | B | B |

Figure 3: The data generated by the 3 nearest neighbours classifier. Sometimes, it doesn't give the right answer!

| | **Predicted** $B$ | **Predicted** $M$ |
|---|---|---|
| **Actually** $B$ | | |
| **Actually** $M$ | | |

Figure 4: This is the *confusion matrix* of the classifier. It describes what types of correct and incorrect classifications were made.In the cell **Actually** $B$ **- Predicted** $M$, you should fill in the number of times the classifier reported an $M$ when it should have reported a $B$. That's the number *false positives* the classifier made. Fill in the other cells similarly.