Diagnosing breast tumour using *k*-NN algorithm

Tomáš Chobola; February 19, 2019

1. Abstract

The main purpose of the program is to predict breast cancer based on features computed from a digitised image of a fine needle aspirate (FNA) of a breast mass. For the prediction the program uses the *k*-nearest neighbours algorithm (*k*-NN) and dataset from the UCI Machine Learning Repository, specifically Breast Cancer Wisconsin (Diagnostic) Data Set.[1]

2. Dataset description

Dataset consists of 569 instances of cell nuclei from various breast tumours. Every instance describes features of the cell nuclei which are computed from a digitised image of a fine needle aspirate (FNA) of a breast mass.

The dataset has 32 columns - id number, diagnosis (M for malignant, B for benign), radius (mean of distances from centre to points on the perimeter), texture (deviation of grayscale values), perimeter (mean size of the core tumour), area, smoothness, compactness, concavity (mean of severity of concave portions of the contour), concave points (mean of severity of concave portions of the contour), symmetry, fractal dimension (mean for "coastline approximation" - 1), standard error of the radius, standard error of the texture, standard error of the smoothness, standard error of the compactness, standard error of the concavity, standard error of the concave points.[2]

Application uses only the first 12 columns. First two columns are used merely for identifying the results of the cell analysis. Remaining 10 columns are used for training and validating.

3. Dataset example

853201	M	17.57	15.05	115	955.1	0.0984	0.1157	0.0988	0.0795	0.1739	0.06149
853401	M	18.63	25.11	124.8	1088	0.1064	0.1887	0.2319	0.1244	0.2183	0.06197
853612	M	11.84	18.7	77.93	440.6	0.1109	0.1516	0.1218	0.0518	0.2301	0.07799
854253	M	16.74	21.59	110.1	869.5	0.0961	0.1336	0.1348	0.0602	0.1896	0.05656
854002	M	19.27	26.47	127.9	1162	0.0940	0.1719	0.1657	0.0759	0.1853	0.06261

4. Learning process and making predictions

Application is using the *k*-nearest neighbours algorithm that stores all available cases and classifies new instances based on similarity (distance function). Before the actual process of learning the whole dataset is split into two parts. The first part contains two thirds of the dataset and is used for learning and the remaining part is used for validation. The training data are then fit into a classifier that is an instance of Python class `KNeighborsClassifier` from `scikit-learn` library. Application then finds the best *k* value for the inserted data by scoring the precision of the classifier using the validation part of the dataset. The classifier uses Euclidean distance as the metric function.

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

After picking the right *k* value the program is ready to make predictions based on values set by the user. The result is either that the breast tumour is malignant or benign. The program is able to predict that with precision around 92 % based on testing using the validation part of the original dataset, results may differ when the program is used for actual predictions.

The accuracy could be increased by using much larger dataset for learning and validating processes which would describe almost all possible situations, and taking measuring error into account. Another improvement in accuracy could be made by using different distance functions and comparing the results or using a simple heuristic that would assign weights to the values based on the importance of that particular value in recognising the malignancy of the breast tumour.

5. Sources

[1] Wolberg, William H. “Breast Cancer Wisconsin (Diagnostic) Data Set.” Edited by Nick Street and Olvi Mangasarian, *UCI Machine Learning Repository: Flags Data Set*, archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[2] Uci. “Breast Cancer Wisconsin (Diagnostic) Data Set.” *RSNA Pneumonia Detection Challenge | Kaggle*, 25 Sept. 2016, www.kaggle.com/uciml/breast-cancer-wisconsin-data.