Performance Comparison between k-Nearest Neighbor and Naive Bayes Classifier in Breast Cancer Classification

Izabela Ryś

Contents

1 Introduction

The aim of this project is to carry out a performance comparison between two machine learning methods: k-Nearest Neighbor and Naive Bayes classification.

2 Data set description

Data set contains 699 instances. Each instance consists of:

1. ID number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)

All attribute values are Integer numbers. Missing values have been replaced by an arithmetic mean of the existing values of the attribute.

3 k-Nearest Neighbor Classification

3.1 Introduction

k-Nearest Neighbor algorithm is a non-parametric classification method. It is based on an assumption that similar things exist in close proximity.

3.2 Parameter selection

There are a few things to consider while choosing the right k parameter:

1. smaller values of k may cause noise on the classification,
2. bigger values of k make the algorithm significantly slower,
3. in binary classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.

The value of k selected in implementation of this project is equal to 13.

3.3 Training phase

The training phase consists only of storing the feature vectors and class labels of the training samples.

3.4 Classification phase

In the classification phase, a given instance is classified by assigning the label which is most frequent among the k training samples nearest to that query point.
The distance metric used in the implementation of this project is the squared Euclidean distance.

3.5 Results

Testing set consists of 200 instances chosen at random from the initial data set.

	actual malignant	actual benign
predicted malignant	TM = 69	FM = 3
predicted benign	FB = 3	TB = 125

$$\text{accuracy} = \frac{\text{TB} + \text{TM}}{\text{TB} + \text{TM} + \text{FM} + \text{FB}} = \frac{125 + 69}{125 + 69 + 3 + 3} = \frac{194}{200} \approx 0.97$$

$$\text{precision} = \frac{\text{TM}}{\text{TM} + \text{FM}} = \frac{69}{69 + 3} = \frac{69}{72} \approx 0.958$$

$$\text{sensitivity} = \frac{\text{TM}}{\text{TM} + \text{FB}} = \frac{69}{69 + 3} = \frac{69}{72} \approx 0.958$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{sensitivity}}} = \frac{2}{\frac{72}{69} + \frac{72}{69}} \approx 0.958$$

4 Naive Bayes Classification

4.1 Introduction

Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features.

Bayes' theorem is stated mathematically as the following equation:

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

where A,B are events and $p(B) \neq 0$.

4.2 Probabilistic model

Let $x = (x_1, \ldots, x_n)$ be a vector representing a problem instance with n features.
Naive Bayes is a conditional probability model which assigns

$$p(y_k|x) = \frac{p(x|y_k)p(y_k)}{p(x)}$$

for each of k possible classes.

There is interest only in the numerator of that fraction, because the denominator only depends on the given vector x and so it is effectively constant.

Using the joint probability model and the chain rule for repeated applications of the definition of conditional probability, the numerator can be rewritten as:

$$p(x|y_k)p(y_k) = p(x_1, \ldots, x_n|y_k)p(y_k) = p(y_k, x_1, \ldots, x_n) = p(x_1, \ldots, x_n, y_k) =$$

$$= p(x_1|x_2, \ldots, x_n, y_k)p(x_2, \ldots, x_n, y_k) = p(x_1|x_2, \ldots, x_n, y_k)p(x_2|x_3, \ldots, x_n, y_k)p(x_3, \ldots, x_n, y_k) =$$

$$= \cdots = p(x_1|x_2, \ldots, x_n, y_k)p(x_2|x_3, \ldots, x_n, y_k) \ldots p(x_{n-1}|x_n, y_k)p(x_n|y_k)p(y_k)$$

Under the assumption that all features in x are mutually independent, the model can be further expressed as

$$p(y_k|x) \propto p(x|y_k)p(y_k) = p(y_k)p(x_1|y_k) \ldots p(x_n|y_k) = p(y_k) \prod_{i=1}^{n} p(x_i|y_k).$$

The **Naive Bayes Classifier** is a function that assigns a class label $\hat{y} = y_k$ for some k as follows:

$$\hat{y} = \underset{k \in 1,\ldots,K}{\arg\max}\, p(y_k) \prod_{i=1}^{n} p(x_i|y_k)$$

4.3 Training phase

The training phase consists of computing and storing:

1. the probabilities that the class is "malignant" or "benign",
2. the conditional probabilities of attribute distribution given that the class is "malignant" or "benign",

based on the data instances from the training data set.

4.4 Classification phase

In the classification phase, a given data instance is classified by computing the $\hat{y}$ function described in (4.2).

4.5 Results

Testing set consists of 200 instances chosen at random from the initial data set.

	actual malignant	actual benign
predicted malignant	TM = 72	FM = 0
predicted benign	FB = 4	TB = 124

$$\text{accuracy} = \frac{\text{TB} + \text{TM}}{\text{TB} + \text{TM} + \text{FM} + \text{FB}} = \frac{124 + 72}{124 + 72 + 0 + 4} = \frac{196}{200} \approx 0.98$$

$$\text{precision} = \frac{\text{TM}}{\text{TM} + \text{FM}} = \frac{72}{72 + 0} = \frac{72}{72} = 1$$

$$\text{sensitivity} = \frac{\text{TM}}{\text{TM} + \text{FB}} = \frac{72}{72 + 4} = \frac{72}{76} \approx 0.947$$

$$\text{F}_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{sensitivity}}} = \frac{2}{\frac{72}{72} + \frac{76}{72}} \approx 0.972$$

5 Performance comparison

	k-Nearest Neighbor	Naive Bayes
true malignant	69	72
true benign	125	124
false malignant	3	0
false benign	3	4
accuracy	0.97	0.98
precision	0.958	1
sensitivity	0.958	0.947
F_1 score	0.958	0.972

Naive Bayes outperformed k-Nearest Neighbor classifier in most of the parameters. However, the differences between these results were relatively small.

6 Conclusion

Even though both k-Nearest Neighbor and Naive Bayes Classifier are simple classifiers, they exhibited high accuracy, precision and sensitivity.

References

[1] Breast Cancer Wisconsin (Original) Data Set
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[2] Wikipedia: k-nearest neighbors algorithm
https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#Parameter_selection

[3] Wikipedia: Naive Bayes classifier
https://en.wikipedia.org/wiki/Naive_Bayes_classifier