# Assignment 4: Classifier Comparison

Pascal Schlachter & Mario Döbler

Institute for Signal Processing and System Theory
University of Stuttgart

## 1 Solution

### 1.1 Theory

#### 1.1.1 Nearest mean classifier

1. Since the dataset is binary, two means have to be calculated. For each mean, $1 \cdot 2 = 2$ multiplications and $10,000 \cdot 2 = 20,000$ summations are necessary (the factor 2 comes from the number of dimensions). In total, $2 \cdot 2 = 4$ multiplications and $20,000 \cdot 2 = 40,000$ summations have to be performed.

2. 2 distance calculations (1 for each mean) are necessary for each test sample. Hence, $2 \cdot 20,000 = 40,000$ distance calculations are necessary for the whole test set

#### 1.1.2 k-nearest neighbor classifier

3. No training is necessary for the $k$-nearest neighbor classifier

4. To find the nearest neighbors, the distances to all training samples have to be calculated for each test sample. This corresponds to $20,000$ distance calculations per test instances. Consequently, $20,000 \cdot 20,000 = 400,000,000$ distance calculations are necessary for the whole test set.

5. For training, the relation of the computational complexity is given as

$$\text{nearest mean classifier} > k\text{-nearest neighbor classifier}.$$

However, for inference the relation of the computational complexity is as follows

$$\text{nearest mean classifier} \ll k\text{-nearest neighbor classifier}.$$

A high computational complexity is more costly during inference. Typically, training is only done once, while inference is performed multiple times. If the trained classifier is implemented in a productive environment (possibly even a real-time application) or should be applied on a large dataset, a slow inference cannot be tolerated.

Table 1: Suitable model orders of the Gaussian mixture model for all datasets

| Dataset | $M_0$ | $M_1$ |
|---|---|---|
| Two moons | 2 | 2 |
| Four parallel | 2 | 2 |
| Four Gaussian | 2 | 2 |
| Circular | $\geq 3$ | 1 |

Table 2: Time for inference of each classifier.

| | nearest mean classifier | $k$-nearest neighbor classifier | Gaussian mixture model classifier |
|---|---|---|---|
| Times | 0.786 s | 49.634 s | 0.008 s |

### 1.1.3 Gaussian mixture model classifier

6. See Table 1.

7. LRT of ML decision:

$$\frac{p(\underline{x}|\omega_2)}{p(\underline{x}|\omega_1)} \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \gamma = 1$$

LLRT of ML decision:

$$\log p(\underline{x}|\omega_2) - \log p(\underline{x}|\omega_1) \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \log \gamma = \log 1 = 0$$

## 1.2 Practice

1. See solution of the code.

2. (a) The times can be found in Table 2. The $k$-nearest neighbor classifier is the slowest, while the Gaussian mixture model classifier is the fastest. (The exact values may differ but the relations should be the same)

   (b) The exact results are individual. However, the following formulas should be applied.

$$\begin{aligned} \text{FP rate} &= \frac{n_{21}}{n_{11} + n_{21}}, \\ \text{TN rate} &= 1 - \text{FP rate}, \\ \text{TP rate} &= \frac{n_{22}}{n_{12} + n_{22}}, \\ \text{FN rate} &= 1 - \text{TP rate}, \end{aligned}$$

   where $n_{ij}$ are the elements of the confusion matrix.

   For all datasets, the nearest mean classifier performs worst. The $k$-nearest neighbor classifier and Gaussian mixture model classifier perform equally well (if the model orders are chosen suitably).