Performance Evaluation

1 Introduction

Learning Goals

- set parameters for the k-nearest neighbors (k-NN) classifier
- understand performance measures other than accuracy
  - precision, recall, F_1-score
- read and interpret confusion matrices and scikit-learn's classification report

2 k-nearest Neighbors Classifier

Training

Store training data.

Testing

For each data $\boldsymbol{x}$ in test, find the k closest (w.r.t. Euclidean norm) data in the training set train by letting $d_i = \left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}\right\|$, $i \in \text{train}$ and sorting those such that $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(k)} \leq \ldots$. Find the indexes of the k nearest training data,

$$\mathcal{N} = \left\{i \in \text{train} \ : \ d_i \leq d_{(k)}\right\} \qquad (k \text{ nearest neighbors})$$

and decide for the most frequent class therein,

$$\pi\left(\boldsymbol{x}\right) = \underset{c \in \mathcal{L}}{\arg\max} \, \# \left\{\tau\left(\boldsymbol{x}^{(i)}\right) = c \ : \ i \in \mathcal{N}\right\}.$$

Recall the principles of the k-nearest neighbors (k-NN) classifier from the lecture. Take a look at scikit-learn's k-NN documentation to make yourself familiar with its usage. Find out about the classifiers **tie-breaking rule**, i.e. what happens if two data points have the same distance.

3 Tasks

Load Dataset

1. Create a new jupyter notebook or reuse your earlier work.
2. Load the UCI ML Breast Cancer Wisconsin using scikit-learn's API.
3. Read the information available in the user guide to be able to discuss the questions below with a partner.

Pair work: Answer the following questions

- How was the data obtained?
- How many classes are there?
- How many data points and features does the dataset contain?
- Which kind of features are there?
- Which feature(s) has/have the highest absolute values?
- Which two features exhibit the highest correlation coefficient?

Classification and Evaluation

1. Split data into training and test set (85:15).
2. Apply scaling to the dataset.
3. Train a `KNeighborsClassifier` and report the accuracy score on the test set.
   - Seed the random generator.
   - Set $k = 7$ for the k-NN.
4. Using your classifier (or its predictions), display a confusion matrix with the help of sklearn's `ConfusionMatrixDisplay`.

Pair work: Answer the following questions

- Which class numbers represent benign resp. malignant data points?
- How many data points are correctly classified as benign?
- How many data points are correctly classified as malignant?
- How many data points are classified as malignant, although being benign?
- How many data points are classified as benign, although being malignant?
- Which of the two last two cases is more favorable from a medical point of view?

Evaluation cont.: Performance Measures

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

1. Given the output from the confusion matrix, compute precision, recall and F_1-score for both labels (benign and malignant).
2. Afterwards, use the methods implemented in scikit-learn (`precision_score`, `recall_score`, `f1_score`) to verify your results for both classes.
3. A compact form of displaying performance measures is given to you via the classification report. Make sure you are able to read and interpret this report!
4. Lastly, you may experiment with different settings to see how these affect performance measures, e.g. using a different k, select a subset of features for classification or use unscaled data.

4 Homework and Quiz

Apply the knowledge gained in this session (Lab 3) to the wine dataset. Compare the k-NN classifier's performance to that of MinDist by entering its results in your table.

Take the quiz in Moodle. Make sure to have your jupyter notebook open, code up and running and performance metrics already pre-computed.

Attention: As we want you to enter some float values in this quiz and have your answers graded automatically, we need to first check if your browser uses comma or dot notation to denote decimals. This is why you need to do **Quiz 3 | Test your formatting** first, before the actual quiz becomes available to you.