096202 - Introduction to Data Analysis

HW4 - Classification & Clustering

Guidelines

- Deadline is 16.1.20 at 23:55
- Submission in singles. Each student is expected to work on the notebook independently.
- Discussing the homework (but not working on them) with other students is allowed, but any student who does so must add a markdown cell at the top of the notebook explicitly naming the person with whom s/he discussed the homework.
- Submission will include a single Jupyter notebook file named '**HW4_ID.ipynb**'
- Present all supporting calculations and analyses in **code cells**. Document your code with comments and clearly specify the final answers in **markdown cells**
- Any visualization **must include** axis labels.
- A particularly thoughtful and careful work may result in bonuses.

Classification - kNN

Description

For the following questions we will use the 'breast cancer' dataset from 'sklearn'. Use the following command to load the dataset: *datasets.load_breast_cancer()*.
In this exercise we'll use the kNN algorithm to build a model that predicts whether a tumor is benign (not dangerous) or malignant (dangerous) based on the features provided in the dataset. To optimize our model we'll try to find the optimal number of neighbors for the kNN algorithm that provides the highest accuracy. We'll use the K-fold cross-validation method (K=5) to evaluate the accuracy scores for each kNN variation.

Use the first 500 records for training+validation, and the remaining 69 records as test data.

Tasks

1. Run the kNN model with different values of k ranging from 1 to 20. Use the K-fold cross validation method using K=5 to evaluate the accuracy of each iteration.
2. Present the results using a plot showing the model accuracy scores for each k.
3. What is the optimal k for the model?
4. What is the accuracy of the optimal model on the test set?

Clustering - K-Means

Description

In this section you will perform a clustering task using the K-Means algorithm. Download the datafile: *'kmeans_data.csv'* from the course's Moodle site.

Tasks

1. Find the optimal number of clusters (K) based on the elbow method (present SSE vs. K plot). What is the optimal K?
2. Use the optimal K from the previous task to cluster the data using K-Means. Plot the clustering results: Use a different color for each class, and use a unique visualization for the centroids. Explain the results
3. Based on the visualization, are the clusters found by the algorithm reasonable? If so, explain why (1-2 sentences). If not, suggest a way to improve the clustering task and present your new results (visualization guidelines as in the previous task)