Biomedical Data Science & AI

Exercise sheet 6 - Introduction

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

Due date: June 8th

Questions

Exercise 1 - NMF Clustering (13 points)

1. Write an algorithm to showcase the working of Non-negative matrix factorization (NMF) **(2 points)**
2. Mention the pros and cons of NMF as well as one of its applications. **(1 point)**
3. Use the nimfa package for NMF clustering on gene expression data to cluster genes into groups. Use the parameters (10 ranks, 50 maximum iterations and 25 runs) to compute the following:
    a. From the average connectivity matrix across multiple runs compute consensus matrix. **(1 point)**
    b. Produce a heatmap with a dendrogram from the clustering results you obtained. **(1 point)**
    c. What are the consequences of selecting a rank value that is too small or too large? Implement a method showing how you can optimize the value of the rank to be used. **(2 points)**
4. Inform yourself about Non-Negative Matrix Tri-Factorization (NMTF). What is the primary difference between NMF and NMTF and what does it achieve?
   (3 points)
5. PCA and NMF are both matrix factorization methods, how do they differ from each other? Describe a situation where PCA is favored over NMF. **(3 points)**

Exercise 2 - Machine Learning (12 points)

1) The type of machine learning (e.g. supervised learning, unsupervised learning, etc.) applied depends on the problem at hand. Assume that we have an Alzheimer's disease (AD) dataset where rows represent 500 participants and columns represent 100 different collected measurements for each participant.
    a) You are asked to train a model that can predict whether a participant is healthy or AD. Mention the type of machine learning you would use for this case scenario and elaborate. **(1 point)**

   b) Assume that we do not have any information about the diagnosis of each participant. This time we would like to divide our participants into groups based on the features that we have in hand. What type of machine learning would be appropriate for this scenario and elaborate? **(1 point)**
   c) Imagine that the shape of our dataset is (100, 600), mention one pre-processing step that you would take to carry out the tasks (a) and (b)? **(1 point)**

2) Generate a pipeline in scikit learn using the following code snippet,

```
polynomial_features = PolynomialFeatures(degree=15, include_bias=False)

linear_regression = LinearRegression()

pipeline = Pipeline([
      ("polynomial_features", polynomial_features),
      ("linear_regression", linear_regression)
])
```

   a) Using the Fish dataset provided, identify the quality of fit of the pipeline for the dataset (use the weight as the response variable). **(2 points)**
   b) If the pipeline produces a badly fit model for the dataset, list some methods to improve the model. **(1 point)**

3) In this exercise we will compare the accuracy of different methods on a high-dimensional (p>>n) dataset. Load the leukemia_small.csv and extract the class labels from the column names (2 classes, “AML” and “ALL”)
   Randomly split the data into 70% training and 30% test.
   Hint: Use the train_test_split function from scikit-learn to define the test_size and set random_state=1 for better reproducibility.
   a) Fit a logistic regression (no penalization) **(1 point)**
   b) Fit multiple l1-penalized logistic regressions (lambdas = 0.001, 0.01, 0.1, 1, 10, 100) **(1 point)**
   c) Fit multiple l2-penalized logistic regressions (lambdas = 0.001, 0.01, 0.1, 1, 10, 100) **(1 point)**
   d) For the models from (a), (b), and (c) measure the performance on the training and test set **(1 point)**
   e) Using (d) report the performances with one scatterplot for each approach (1 scatterplot for unpenalized, l1, l2), with the regularization constant on the x-axis and the accuracy on the y-axis, train and test set colored differently, proper axis labels and a legend. **(1 point)**

f) Which method in combination with which parameter gives the best results on the test set? **(1 point)**