DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Project 2: Dimensionality Reduction in Classification

Mathematics for Machine Learning

(MECD & MMAC, 2^{nd} Semester, 2022/2023)

Handed out on March 22, 2023.

To be handed back by **May** 2, 2023.

This project could be done in R or Python.

Consider the dataset *Turkish Music Emotion* and associated description available at: https://archive.ics.uci.edu/ml/datasets/Turkish+Music+Emotion+Dataset (see [3] for additional information).

1. Make a preliminary analysis of the data, identify potential problems. Discuss your findings.
2. Fix a seed, and randomly divide your dataset into 75% for training and 25% for testing.
3. Use the training dataset to run the forward feature selection methods based on information theory to sort the features (see [1,5,6]).
4. Consider the first s features from the sorted list of features obtained in (3), with $s = 15, 25$, as input of the classification process discussed below.
5. Consider two classifiers: k-Nearest Neighbour (kNN) with $k = 5$ neighbours and another classifier at your own choice (e.g. see [2,4]).
   (a) Train the classifiers using the training dataset.
   (b) Predicting the class for each test set observation.
   (c) Compare the true class and the assigned class for the test set, and estimate the following measure of performance (e.g. see [4]):
      i. Accuracy (Acc): percentage of corrected assigned observations.
      ii. Macro_Recall (Macro_Re): arithmetic mean of the classes recall, where the recall of the i-th class ($Re(i)$) is the percentage of the observations of the i-th class correctly assigned to that class.
      iii. Macro_Precision (Macro_Pr): arithmetic mean of the classes precision, where the precision of the i-th class ($Pr(i)$) is the percentage of observation assigned to the i-th class that truly belong to the i-th class.
      iv. Macro_F_1 measure (Macro_F_1): arithmetic mean of the classes F_1-measure, where the F_1-measure of the i-th class is the harmonic mean of the i-th class recall and precision.

6. Apply principal component analysis to your dataset and decide the number of features to be retained. Use the scores as input variables and repeat (5).

7. Apply another dimensionality reduction method and use the projected data as input features to the classification process described in (5).

8. Compare and discuss the results resulting from the different strategies. Include in your discussion all options that you have made, the advantages, and disadvantages of each alternative.

References:

[1] G. Brown, A. Pocock, M.-J. Zhao, M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, *J. Mach. Learn. Res.*, 13, 27–66, 2012.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, NY, 2013 (Printing 2021).

[3] M. E. Bilal and I. B. Aydilek. Music emotion recognition by using chroma spectrogram and deep visual features. *International Journal of Computational Intelligence Systems*, 12(2), 1622–1634. , 2019. doi:http://dx.doi.org/10.2991/ijcis.d.191216.001

[4] M. Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1–26, 2008. doi:http://dx.doi.org/10.18637/jss.v028.i05. See A Short Introduction to the caret Package or The caret Package.

[5] C. Pascoal, M. R. Oliveira, A. Pacheco, and R. Valadas. Theoretical evaluation of feature selection methods based on mutual information, *Neurocomputing*, 226(1) 168–181, 2017.

[6] H. Zhou, X. Wang, R. Zhu. Feature selection based on mutual information with correlation coefficient, *Applied Intelligence*, 52(5), 5457–5474, 2022.

- **Progress check:**
  - On **18 April** 2023, each group should do a 5 min presentation, reporting their progress, difficulties, and future work on the Project 2.

- **About the groups:**
  - Students should organize themselves in groups of 2 persons.

- **About the report:**
  - You have to deliver a report including an explanation of what you have done, the visualization of our results, and a critical discussion of your results.
  - The report should not exceed 8 pages.
  - The **commented R/Python code** and the **report** must be uploaded to the Fénix webpage.