Aston University
Machine Learning

Portfolio Task 4: Dimensionality Reduction

Released: 07/12/20
Due: 18/12/20, before 23:59

Instructions:
In this assessed task, you will be applying some of the dimensionality reduction algorithms covered in this module. The aim of this task is to test your ability to apply machine learning algorithms to well-specified tasks and to evaluate the performance of these algorithms and to use this evaluation to improve performance.

Details:
Follow the instructions below to complete the portfolio task. The task requires you to carry out some implementation in Python and to provide a short written justification of your choices, of maximum 250 words. The recommended format for submission is a Jupyter notebook, integrating your code and written justification.

Marking:
This portfolio task is worth 15% of the overall module mark.

The mark scheme for the task is as follows:

- **50-59** Solution approaches have been applied to both sub-tasks and, where requested in the task, their performance measured. The approaches taken are broadly correct but may have some flaws in application or methodology. Model evaluation and a justification of chosen approaches have been attempted but shows limited understanding.
- **60-69** Justification in sub-task 1 shows clear understanding of the properties of the chosen algorithms. Multiple solution approaches (algorithms/models/parameter sets) have been applied to the problem in sub-task 2 and have undergone evaluation. Justification for the selected approach is evidence-based and well presented.
- **70-79** The methodology used to compare solution approaches for sub-task 2 is carefully designed and leads to well-supported conclusions. Clear understanding of experimental design is demonstrated.
- **80+** As above, but with additional evidence (for both sub-tasks) of some or all of: attention to quality throughout the implementation, thorough understanding in experimental design, excellent justification.

No specific descriptors are provided for marks below the threshold of 50. Marks in the range **0-49** are allocated where the submitted work has not reached the expectation for the threshold descriptor.

Sub-task 4.1:
Download the file kc_house_data_reduced.csv from Blackboard. It contains approximately 21k data points, each of which contains data on houses in King County USA (adapted from this dataset). For each house, it lists the house's predicted sale price, condition, grade (quality of construction) and three measures of size in square feet.

One of your colleagues suggests that the non-price aspects of this dataset could be modelled using two latent variables: quality (capturing condition and grade) and size (capturing the three size measures) and that, taken together, these should capture the variability in price.

Use scikit learn's implementation of factor analysis to fit a model with two components to the data.

Print out the components found through factor analysis. Are they easily interpretable (e.g. as the model we proposed for car data during lectures was)? If so, provide an interpretation of the components in plain English. If not, suggest why the factor analysis algorithm didn't result in an interpretable model.

Sub-task 4.2:
Download the file kc_house_data.csv from Blackboard. It contains the full dataset of 21k data points, each with 19 features.

Use a dimensionality reduction method of your choice to visualise the data in three dimensions. Evaluate the success of your dimensionality reduction procedure to decide whether the three dimensional projection of your dataset captures the important details of the original dataset. Justify your answer and, if you don't feel that a three dimensional projection captures these details, design and implement a methodology to determine the appropriate minimum dimensionality to project the data to.