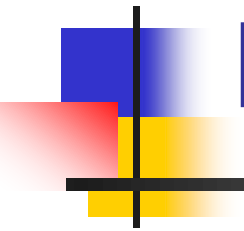


Machine Learning

Term Project Specification

Ok-Ran Jeong and Woong-Kee Loh

2021

Term Project Requirements (1/2)

- The term project is a team project.
- Prepare a proposal and submit it next week.
  - Your proposal must include a statistical description of the dataset, objective, and algorithms to use.
  - There will be no presentation of proposals.
  - I will give you my comments next week, and, if needed, your modified proposal must be re-submitted in the next week.
- Final presentation will be made in the last week before Final Exam.
  - For each presentation I will give you my comments, and you should reflect them in your final reports.
  - Write the manual of your entire program framework in a scikit-learn style (especially, auto machine learning in p.10-12)

Term Project Requirements (2/2)

- You must apply every step of end-to-end Big Data process.
  - ≥3 classification and ≥3 clustering algorithms you studied in this lecture
  - ≥1 clustering algorithms that you studied for active learning (should be mentioned in your proposal and final report)
- You should find a suitable dataset.
  - For educational purposes, the dataset must include a reasonable number of records and features (attributes) and also a reasonable amount of dirty data and categorical data.
  - Use the same dataset for both classification and clustering as in PHW #1 and #2. See p.9 for more explanation.

Term Project Proposal

- Your proposal should include the following:
  - Project title
  - Dataset - one paragraph description and source
    - The dataset should include categorical attribute(s)
  - ** (for classification) Provide a list of features that you think will most influence the classification accuracy and your explanation on why.
  - Project idea, including a clear description on the problem and your approach to solving it
  - Your estimated schedule and collaboration plan
  - Due: 9PM on Oct. 19 (Wed. class) and Oct. 20 (Thur. class)

Dataset (1/2)

- You may select a dataset from the provided list or find a suitable dataset on your own
- Requirements
  - Columns (number of attributes): 15+
  - Rows (number of data instances/records): 10,000+

Dataset (2/2)

- Google dataset search (https://toolbox.google.com/datasetsearch)
- Kaggle (https://www.kaggle.com/datasets)
- UCI Machine Learning Repository (http://mlr.cs.umass.edu/ml/)
- VisualData (https://www.visualdata.io)
- CMU Libraries (https://guides.library.cmu.edu/machine-learning/datasets)
- data.gov (https://www.data.gov)
- The US National Center for Education Statistics (https://nces.ed.gov)
- The UK Data Service (https://www.ukdataservice.ac.uk)
- Data USA (https://datausa.io)
- Others

Classification

- What to consider in this project
  - Attribute information analysis
  - Categorical-to-numerical encoding
  - Machine learning algorithms
    - Logistic regression, KNN, SVM, decision tree, random forest, gradientboostingclassifier, xgbclassifier, gaussiannb, votingclassifier, etc.
  - Confusion matrix, ROC curve
  - Precision, recall, F1, avg_total analysis

Clustering

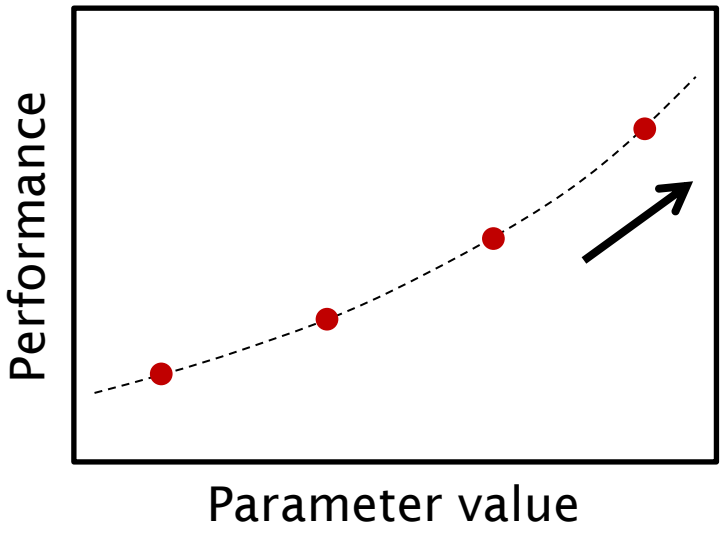
- What to consider in this project
  - Merging related attributes (based on your objective)
  - Attribute information Analysis
  - Data preprocessing
    - Avoid removing NaN value as much as possible
  - Categorical-to-numeric encoding
  - Machine learning algorithms
    - k-Means, EM, DBSCAN, etc.
    - Similarity measures
  - Visualization – if needed, you may use PCA
  - Evaluation
    - Silhouette
    - Purity

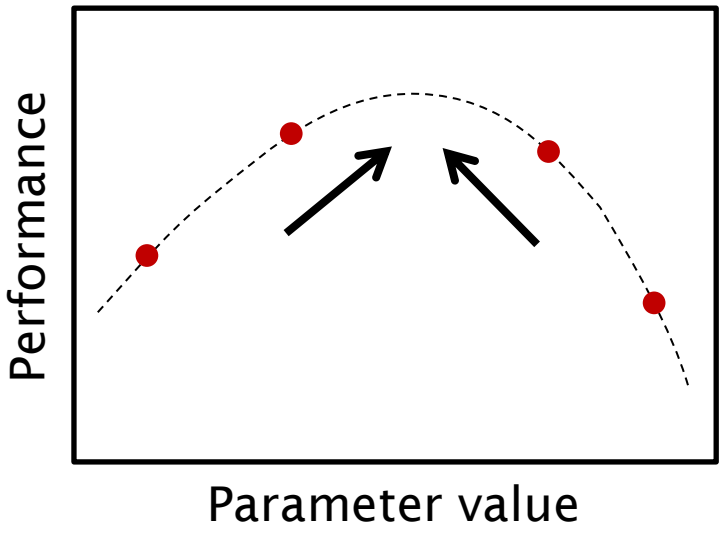
Dataset Application

- Apply the same dataset for classification and clustering
  - Use the same set of features; exclude the target attribute in clustering
  - In clustering, you must not use *any* information that could be obtained from the target attribute
  - Perform and evaluate classification and clustering independently according to their own points of pursuit
  - Compute silhouette and purity for clustering results; evaluate the clustering quality based on the measures

Auto Machine Learning (1/3)

- Parameter tuning
  - You find the parameter combination as best as you can (local optima).
  - Cases for obtaining best performance
    - With a larger/lesser parameter value
    - With a parameter value in the middle of previous ones





Auto Machine Learning (2/3)

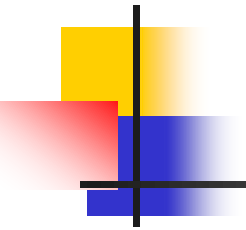
- Simple pseudocode
  - Set randomly a default value v_i for each parameter p_i.
  - Measure accuracy.
  - Repeat
    - Repeat for each parameter p_i
      - For a few values v_i, measure accuracy. Use default values for the remaining parameters.
      - Set the value with the best performance as default.
      - Find the case of performance trend.
    - Until only a trivial improvement in accuracy
  - Until only a trivial improvement in accuracy
  - Print all default parameter values and the accuracy.

Auto Machine Learning (3/3)

- Rationale of outer repeat
  - Since the 'best' values for parameters are obtained in a certain order of parameters, the values obtained in the front might not be the best.
  - For every iteration of outer repeat, the accuracies before and after the inner repeat are compared.
  - If the difference is trivial, e.g., for a pre-specified small ε, $(acc_{after} - acc_{before})/acc_{before} \leq \varepsilon$, then exit the outer repeat.
    - You should be careful when deciding ε, or you might fall into an infinite loop. At start, try using a rather large value for ε, and keep monitoring intermediate results.

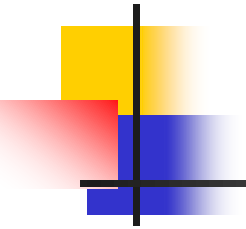
Wise Prophet

- http://prophet.wise.co.kr
- TBA – negotiating with WISEiTECH



Team Members

- Announced in a separate pdf file
- Different from those for the previous homework



End of Specification