INFO 1998 Final Project Deliverable

Clustering, Meta-Learning, Cross Validation

Guideline and Rubric

Release Date: April 16th
Due Date: May 7th
Submit Through: CMS

Overview

For this project, we will further expand on our repertoire of machine learning algorithms. Specifically, we will focus on clustering (different from regression and classification), meta-learning (ensemble learning), and cross validation. You will be using the same train.csv from the housing dataset for this project as well and you can reuse your cleaned dataset again. Remember that clustering is an unsupervised machine learning algorithm, so there should be no labels used to train the models.

Because you are doing unsupervised learning, your goal is for the clustering models you create to learn latent variables. You should also produce one or more visualizations for each model. There is no graded baseline or accuracy measure since this is unsupervised learning, so your paragraph to explain your thinking is important for this project. Some things to consider when trying to show that the clustering models extract meaningful results are checking whether the clusters are ones you would expect and if they make sense to you. Additionally, try different hyperparameters to see how these affected your model results.

Don't forget to submit a **brief** paragraph on each of your models. This paragraph should discuss what kind of conclusions you can draw from your dataset, as well as the motivation for your procedure. Though there is no baseline for unsupervised learning, we expect that your paragraphs will demonstrate that the clusters formed are meaningful in some way.

In addition to clustering, you will implement ensemble models and use cross validation to check model accuracy. There were three types of meta-learning taught: bagging (models are applied without knowledge of each other and then averaged), boosting (models are applied sequentially on top of each other), and stacking (take the weighted average of many different models). For this assignment you will have to build **one ensemble model** out of the three options. For this model you will be **predicting the SalePrice column** (remember in project B you created a linear regression model for SalePrice). You can either predict the continuous sale price or you can split the data into strata ($50,000 - $100,000, $100,000 - $150,000, etc.) if you want to build a categorical model).

Each layer of your model should show improvement to ensure your ensemble model works. This means, for a bagging model, each base layer model should show improvement over the baseline and the final layer which averages all of the models should show improvement over all the base models. For a boosting model, the first model should show improvement over the baseline and each successive model should improve on the previous. For a stacking model, each base layer model should show improvement over the baseline and the final model should improve over all the base models. We recommend setting the **random seed** to ensure this is reproducible.

For your ensemble model, **explain which type of cross-validation** you did for every model in your ensemble. You must use **at least two different types of cross-validation**. Additionally, **explain one pro and one con** of that type of validation. Note that you can use the same explanation for multiple models if applicable (ex. model A, B, and C use __ type of cross-validation).

Course Evaluation

There will be a course evaluation form available on CMS shortly (different from the mid semester evaluation). It really helps the team as a whole to fix, improve, and prepare for the next semester's training program. The submission of course evaluation will be 5% of the total grade. However, because the evaluation will be completely anonymous, we can't keep track of who submitted on our record, so please indicate in your final project write-up whether each member has submitted one or not. The overall grade breakdown will be:

- Quiz: 15%
- Project A: 20%
- Project B: 20%
- Project C: 20%
- Final Project: 20%
- Course Evaluation: 5% (this will be individually applied)

What to Submit:

A Jupyter Notebook containing:

- Hierarchical Clustering
  - Code for the hierarchical clustering
  - Appropriate dendrogram visualization
  - Paragraph on your hierarchical clustering algorithm
- K-Means Clustering
  - Code for the K-Means Clustering
  - Appropriate clustering visualization
  - Paragraph on your K-Means clustering algorithm
- Code for ensemble model (including accuracy of each layer)
  - Code for the model
  - Code that shows accuracy improvement over each layer using cross validation
- Cross-validation explanation paragraph
- Course Evaluation (on CMS)

Grading Rubric

Criteria	Points
Dendrogram (Hierarchical Clustering)	
Correct algorithm used	3
Dendrogram is accurate and interpretable - Parameter selection - Appropriate visualization(s) produced	12
Paragraph explanation - Should include the reasoning behind your codes and what insights you gained from the process/results	5
K-Means Clustering	
Correct algorithm used	3
Clustering is accurate and interpretable - Appropriate variable chosen - Parameter selection - Appropriate visualization(s) produced	12
Paragraph explanation - Should include the reasoning behind your codes and what insights you gained from the process/results	5
Ensemble Model	
Model is an ensemble model predicting SalePrice	10
Baseline and accuracy calculations for each layer - Each layer improves on the previous	10
Cross-Validation Explanation	
Explanation of cross-validation for each model - Identify the type of cross-validation - Explain one pro and one con	13
Use two different types of cross-validation	7
Total	**80**