Readme

純回答問題有用紅字建議搜尋方式。

(40pt) Classification Task:
Implement the following algorithms manually, without relying on libraries like scikit-learn.After implementation, compare the performance of these classifiers. You should also select an evaluation criterion of your choice and justify its selection. Ensure to plot the training/validation loss for each algorithm, and all results should be evaluated on the test set.
1. Linear Classifier (5pt):
2. K-NN Classifier (10pt):
3. Naïve Decision Tree Classifier (10pt):
4. Decision Tree with Pruning (15pt):

模型函式: model.py
執行檔案: train.ipynb
主要回答請搜尋第一題

(40pt) Feature engineering:

- (10pt) Implement an algorithm that can determine the "feature importance" for both linear classifiers and decision trees. Explain the rationale behind your chosen algorithm.
- (10pt) Utilize SHAP (https://shap.readthedocs.io/en/latest/) with your implemented algorithm to assess feature importance. Compare your findings with those obtained using SHAP.
- (20pt) It is known that sometimes the original feature set may not be effective. Designing new features based on the original set is crucial for model performance. Based on your observations and experience, propose an algorithm that can derive new features to enhance model accuracy.

執行檔案: train.ipynb
主要回答請搜尋第二題

(30pt) Cross-Validation: Based on Problem 1, use k-fold cross-validation to verify the stability of each classifier. Note that cross-validation could adopt any existing package. Answer the questions below:
1. (10pt) se k=3,5,10, and make some discussions of your observation.
(可以從 Document 搜尋 Based on Problem 1,use k-fold cross-validation to verify the

stability of each classifier)

2. (10pt) Now you have a test dataset you have partitioned from train.csv. Please design an algorithm that can merge/aggregate the predicted results from k classifiers in k-fold cross-validation. Compare the performance and complexity of the crossvalidation with Problem 1.

3. (10pt) How do we know the performance of one model is really better than another one? Please compare the result in 5-fold cross-validation and the result of Problem 1 to justify which is “REALLY” better. Also show me why.

(可以從 Document 搜尋 compare the result in 5-fold cross-validation and the result of Problem 1 to justify which is “REALLY” better.)

執行檔案: kfold.ipynb、Anova.py 計算 P-value

主要回答請搜尋第三題