NYCU Pattern Recognition, Homework 2

**311552034, 林柏翰**

**Part. 1, Coding (70%)**:

1. (5%) Compute the Entropy and Gini index of the array provided in the sample code, using the formulas on page 6 of the HW3 slide.

一張含有 文字 的圖片

自動產生的描述

1. (10%) Show the accuracy score of the validation data using criterion='gini' and max\_features=None for max\_depth=3 and max\_depth=10, respectively.

一張含有 文字 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

1. (10%) Show the accuracy score of the validation data using max\_depth=3 and max\_features=None, for criterion='gini' and criterion='entropy', respectively.

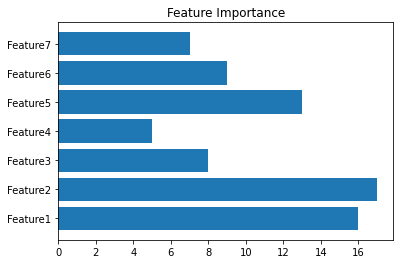
一張含有 文字 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

1. (5%) Train your model using criterion='gini', max\_depth=10 and max\_features=None. Plot the [feature importance](https://sefiks.com/2020/04/06/feature-importance-in-decision-trees/) of your decision tree model by simply counting the number of times each feature is used to split the data.



1. (10%) Show the accuracy score of the validation data using criterion='gini', max\_depth=None, max\_features=sqrt(n\_features), and bootstrap=True, for n\_estimators=10 and n\_estimators=50, respectively.

一張含有 文字 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

1. (10%) Show the accuracy score of the validation data using criterion='gini', max\_depth=None, n\_estimators=10, and bootstrap=True, for max\_features=sqrt(n\_features) and max\_features=n\_features, respectively.

一張含有 文字 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

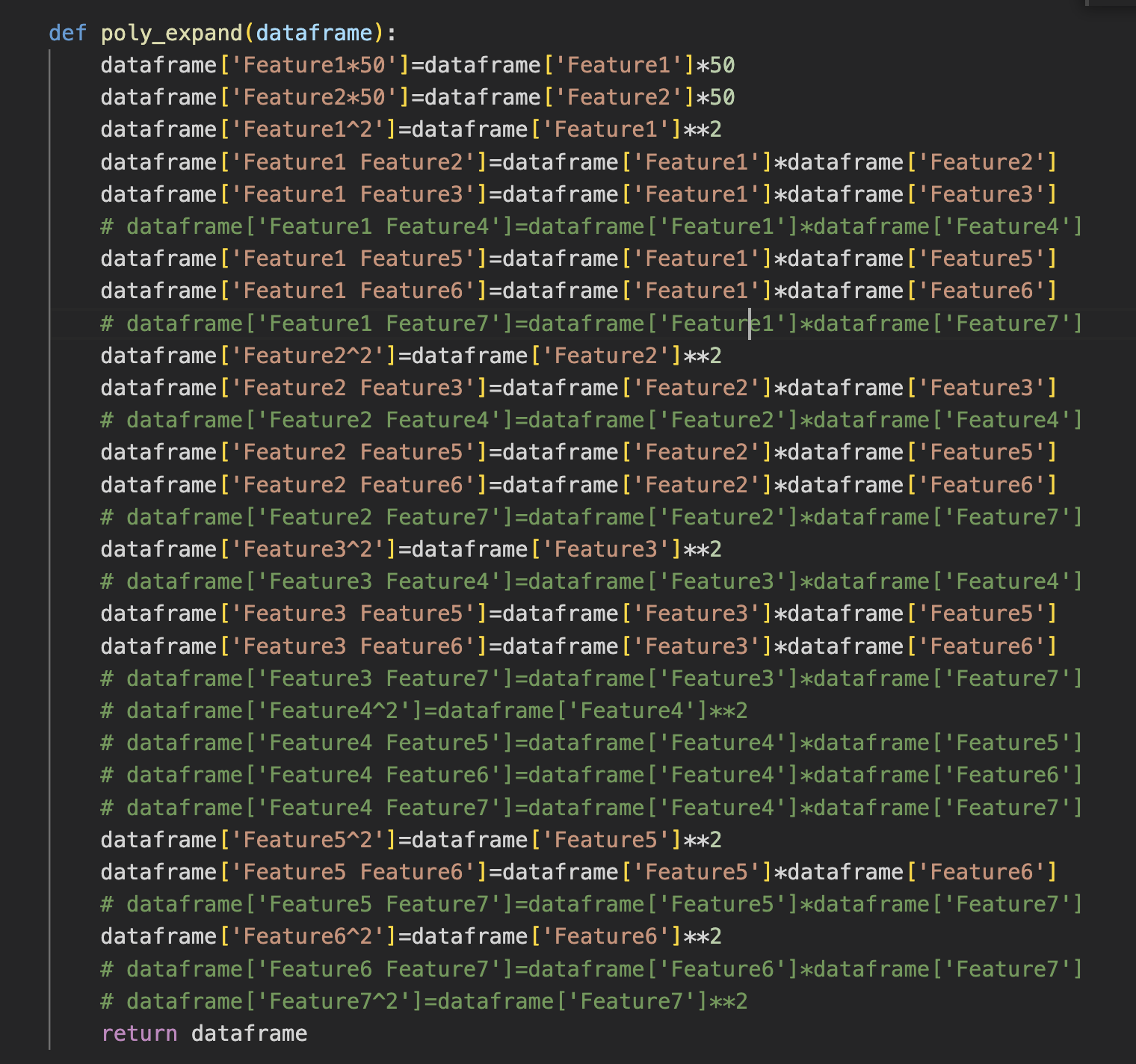
1. **Train your own model**

(20%) Explain how you chose/design your model and what feature processing you have done in detail. Otherwise, no points will be given.

Ans:

Just like HW1 and HW2, I decided to use polynomial feature expansion to create feature.

According to the feature importance graph. I decided to skip the expansion of feature 4&7.



After creating total 24 features, I decide to use random forest to train my model. To figure

Out how many estimators I will need to obtain best accuracy, I use brute method to try many different number of estimator and the result shows that the model reached its best performance with 560 estimators.

一張含有 文字 的圖片

自動產生的描述

and the best accuracy I got on validation set is 0.90375. Besides, I decided to use random seed 2 because 2 is my lucky number.

**Part. 2, Questions (30%):**

1. Answer the following question in detail:
   1. Why does a decision tree tend to overfit the training set?

A decision tree tends to overfit the training set because it can create a complex decision boundary that perfectly fits the training data. As a result, the model may memorize the training data, and fail to generalize well to new, unseen data. This is because the decision tree may split the data too finely, and include features that are not relevant to the target variable. Additionally, decision trees are sensitive to small changes in the training data, which can lead to different decision trees being generated for each set of training data. Overfitting can also occur when the decision tree is too deep, resulting in overly specific rules and decisions that only apply to the training data but not to new data.

* 1. Is it possible for a decision tree to achieve 100% accuracy on the training set?

It is possible for a decision tree to achieve 100% accuracy on the training set, especially if the tree is allowed to grow without any constraints or pruning. However, such a tree may not generalize well to new data. That is, such a model is likely to overfit the data and not generalize well to new, unseen data. And therefore may not perform as well on the test or validation set.

* 1. List and describe at least three strategies we can use to reduce the risk of overfitting in a decision tree.
     1. Pruning: Pruning is a technique used to simplify a decision tree by removing some of its branches or nodes that do not add much to its overall predictive power. This can help to prevent overfitting by reducing the complexity of the tree and making it more generalizable to new data.
     2. Setting constraints: Setting constraints such as minimum sample size per leaf node, maximum depth of the tree, and minimum information gain can help to prevent overfitting. These constraints limit the complexity of the tree, forcing it to focus on the most important features and relationships in the data.
     3. Ensemble methods: Ensemble methods such as Random Forest and Gradient Boosted Trees can also help to reduce the risk of overfitting. These methods use multiple decision trees to make predictions, each with a different subset of features and data samples. By combining the predictions of multiple trees, the ensemble can reduce the risk of overfitting while still achieving high accuracy on both the training and test sets.

1. For each statement, answer True or False and provide a detailed explanation:
   1. In AdaBoost, weights of the misclassified examples go up by the same multiplicative factor.

True, In AdaBoost, the weights of the misclassified examples are increased by the same multiplicative factor, which is given by the formula:

w\_i' = w\_i \* exp(α)

where w\_i is the weight of the ith example in the current iteration, w\_i' is the updated weight of the ith example in the next iteration, and α is the weight of the current weak classifier.This formula ensures that the misclassified examples are given higher weights than the correctly classified examples, and the subsequent weak classifiers focus more on these examples in the next iteration. This process is repeated iteratively, with each iteration producing a new weak classifier and updated weights for the training examples.

* 1. In AdaBoost, weighted training error 𝜺t of the tth weak classifier on training data with weights Dt tends to increase as a function of t.

True ,In the course of boosting iterations the weak classifiers are forced to try to classify more difficult examples. The weights will increase for examples that are repeatedly misclassified by the weak classifiers. The weighted training error ǫt of the tth weak classifier on the training data therefore tends to increase.

* 1. AdaBoost will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

False, not true if the data in the training set cannot be separated by a linear combination of the specific type of weak classifiers we are using .

While the AdaBoost algorithm is designed to improve the classification accuracy of the weak classifiers, it is possible that some examples may be impossible to classify correctly, even with an infinite number of iterations. In such cases, the training error of the AdaBoost ensemble may reach a minimum nonzero value, and cannot bereduced any further.

III. Consider a data set comprising 400 data points from class C1 and 400 data points from class C2. Suppose that a tree model A splits these into (200, 400) at the first leaf node and (200, 0) at the second leaf node, where (n, m) denotes that n points are assigned to C1 and m points are assigned to C2. Similarly, suppose that a second tree model B splits them into (300, 100) and (100, 300). **Evaluate the misclassification rates for the two trees and hence show that they are equal**. Similarly, **evaluate the cross-entropy Entropy= -**and **Gini index** Gini=1- **for the two trees**. Define to be the proportion of data points in region R assigned to class k, where k = 1, . . . , K.

一張含有 文字, 信 的圖片

自動產生的描述

一張含有 文字, 信 的圖片

自動產生的描述