سوال اول )

قبل از اینکه به حل سوال بپردازیم بهتر هست به درک الگوریتم حذف کاندید (Candidate Elimination ) بپردازیم و مراحل و Step های متفاوت اون رو هم معرفی کنیم و بعد با دیدگاهی واضح تر به حل سوال بپردازیم.

این الگوریتم در واقع توسعه یافته الگوریتم پیدا کردن S هست.

از اصطلاحاتی که استفاده میشه

Concept Learning: که در بحث ماشین لرنینگ همون حکم train رو داره.

General Hypothesis: علایم به طور کلی مشخص کننده feature برای تعلیم نیست.

G = {‘?’, ‘?’,’?’,’?’…}: Number of attributes

Specific Hypothesis: Specifying features to learn machine

S= {‘pi’,’pi’,’pi’…}: Number of pi depends on number of attributes.

گام های الگوریتم به طور کلی به طور زیر هستند :

**Step1:** Load Data set

**Step2:** Initialize General Hypothesis and Specific Hypothesis.

**Step3:** For each training example

**Step4:** If example is positive example

if attribute\_value == hypothesis\_value:

Do nothing

else:

replace attribute value with '?' (Basically generalizing it)

**Step5:** If example is Negative example

Make generalize hypothesis more specific.

هیمنطور فایل های ژوپیترنوت بوک و همینطور sample فایل اکسل نیز ضمیمه شده کافی هست دیتا رو با دیتای مورد آزمایش خودمون تغییر بدیم و از کد اجرا بگیریم.

سوال 2 و 3)

A Supervised Machine Learning Algorithm, used to build classification and regression models in the form of a tree structure.

A decision tree is a tree where each -

*Node - a feature(attribute)*

*Branch - a decision(rule)*

*Leaf - an outcome(categorical or continuous)*

There are many algorithms to build decision trees, here we are going to discuss ID3 algorithm with an example.

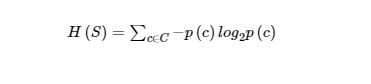
**What is an ID3 Algorithm?**

ID3 stands for Iterative Dichotomiser 3

It is a classification algorithm that follows a greedy approach by selecting a best attribute that yields maximum Information Gain(IG) or minimum Entropy(H).

**What is Entropy and Information gain?**

Entropy is a measure of the amount of uncertainty in the dataset S. Mathematical Representation of Entropy is shown here –



Where,

S - The current dataset for which entropy is being calculated(changes every iteration of the ID3 algorithm).

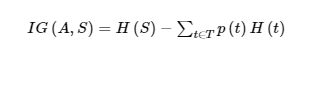
C - Set of classes in S {example - C ={yes, no}}

p(c) - The proportion of the number of elements in class c to the number of elements in set S.

In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set S on that particular iteration.

Entropy = 0 implies it is of pure class, that means all are of same category.

Information Gain IG(A) tells us how much uncertainty in S was reduced after splitting set S on attribute A. Mathematical representation of Information gain is shown here –



**What are the steps in ID3 algorithm?**

1. Calculate entropy for dataset.
2. For each attribute/feature
   * Calculate entropy for all its categorical values.
   * Calculate information gain for the feature.
3. Find the feature with maximum information gain.
4. Repeat it until we get the desired tree.

According to explanations above the main python file is attached at directory (ID3-trees)

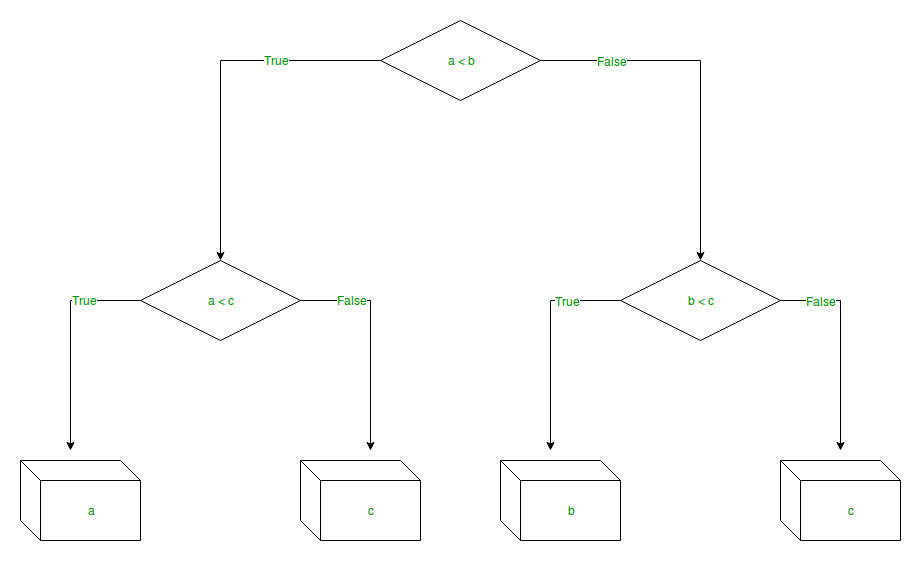
سوال 4)

1. Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances. Decision trees are the most susceptible out of all the machine learning algorithms to overfitting and effective pruning can reduce this likelihood.
2. In machine learning and data mining, pruning is a technique associated with decision trees. Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances.
3. These methods include Reduced Error Pruning (REP), Minimum Cost Complexity Pruning (MCCP), or Minimum Error Pruning (MEP).

سوال6)

The primary difference between classification and regression decision trees is that, the classification decision trees are built with unordered values with dependent variables. The regression decision trees take ordered values with continuous values.

Regression Tree)



Pseudo Code:

*# import the regressor*

*from sklearn.tree import DecisionTreeRegressor*

*# create a regressor object*

*regressor = DecisionTreeRegressor(random\_state = 0)*

*# fit the regressor with X and Y data*

*regressor.fit(X, y)*

Classification Tree)



Pseudo Code:

*def tarin\_using\_entropy(X\_train, X\_test, y\_train):*

*# Decision tree with entropy*

*clf\_entropy = DecisionTreeClassifier(*

*criterion = "entropy", random\_state = 100,*

*max\_depth = 3, min\_samples\_leaf = 5)*

*# Performing training*

*clf\_entropy.fit(X\_train, y\_train)*

*return clf\_entropy*

سوال 7)

In machine learning, we couldn’t fit the model on the training data and can’t say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique.

**Cross-Validation**

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows:

* + Reserve some portion of sample data-set.
  + Using the rest data-set train the model.
  + Test the model using the reserve portion of the data-set

**Methods of Cross Validation**

**Validation**

In this method, we perform training on the 50% of the given data-set and rest 50% is used for the testing purpose. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e higher bias.

**LOOCV (Leave One Out Cross Validation)**

In this method, we perform training on the whole data-set but leaves only one data-point of the available data-set and then iterates for each data-point. It has some advantages as well as disadvantages also.

An advantage of using this method is that we make use of all data points and hence it is low bias.

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation. Another drawback is it takes a lot of execution time as it iterates over ‘the number of data points’ times.

**K-Fold Cross Validation**

In this method, we split the data-set into k number of subsets(known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

Part 2 Practical Problems

Q1)

The first common **strategy** for dealing with **missing data** is to delete the rows with **missing values**. Typically, any row which has a **missing value** in any cell gets deleted. ... Finally, you can use classification or regression models to predict **missing values**.

Q2&Q3)

let’s understand the types of categorical data:

* Nominal Data: The nominal data called labelled/named data. Allowed to change the order of categories, change in order doesn’t affect its value. For example, Gender (Male/Female/Other), Age Groups (Young/Adult/Old), etc.
* Ordinal Data: Represent discretely and ordered units. Same as nominal data but have ordered/rank. Not allowed to change the order of categories. For example, Ranks: 1st/2nd/3rd, Education: (High School/Undergrads/Postgrads/Doctorate), etc.Ways to handle categorical features:

**Implementation:**

Step 1. Replace original cabin value with the first character of the cabin name.

Step 2. Find percentage (%) of people survived in a particular cabin and store into new dataframe.

Step 3. Create a new column into probability survived dataframe with the probability of people getting dead in a particular cabin.

Step 4. Create one more new column into probability survived dataframe i.e ratio of survived and died probability.

Step 5. Create a dictionary with probability ratio column.

Step 6. Create a new column in Data and replace by mapping cabin column categories with its encoded probability ratio dictionary.

Step 7. Drop original cabin column.

Files are attached in the “Practical Content” Directory

Files to Q1,Q2,Q3 🡪 *Handle\_Categorical\_Data.ipynb , Handle\_Categorical\_Missing\_Data.ipynb , Handle\_Continous\_Missing\_Data.ipynb.*