Kaggle Master Week-1 Q&A

Q1- After training our decision tree model, we saw that the model is overfitted on the training data and it has bad performance on the test data. Which hyper-parameter could help us to get rid of this problem? Note: You can use sklearn.tree.DecisionTreeClassifier documentation.
https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

- Criterion
- max_depth ✓
- random_state
- splitter

A1- Yanıt B seçeneği. "criterion" parametresi nodeların hangi feature'a göre dağılacağını hesaplayan fonksiyonları belirlemede kullanır,overfitting ile direkt olarak bağlantısı yoktur. "max_depth" karar ağacımızın maksimum derinliğidir ve çok derin ağaçlar train datasına overfit olurlar."random_state" algoritmadaki rastgeleliğin sabitlenmesi içindir, overfitting konusu ile alakası yoktur."splitter" parametresi ise her bir node'un nasıl ayrışacağına karar verir.

Q2- Which of the below can be said definitely according to the results table taken from the data.describe() method?

I- 75% of the values in the Rooms column are greater than 2.
II- There are some houses with a land size of 0.
III- There are missing values in the BuildingArea column.
IV- There is no house with 9 rooms in the data set

In [2]:
```
import pandas as pd

data = pd.read_csv("/home/fatih/Desktop/melb_data.csv")

data.describe()
```

Out[2]:

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	1
count	13580.000000	1.358000e+04	13580.000000	13580.000000	13580.000000	13580.000000	13518.000000	13580.000000	7130.000000	8205.000000	
mean	2.937997	1.075684e+06	10.137776	3105.301915	2.914728	1.534242	1.610075	558.416127	151.967650	1964.684217	
std	0.955748	6.393107e+05	5.868725	90.676964	0.965921	0.691712	0.962634	3990.669241	541.014538	37.273762	
min	1.000000	8.500000e+04	0.000000	3000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1196.000000	
25%	2.000000	6.500000e+05	6.100000	3044.000000	2.000000	1.000000	1.000000	177.000000	93.000000	1940.000000	
50%	3.000000	9.030000e+05	9.200000	3084.000000	3.000000	1.000000	2.000000	440.000000	126.000000	1970.000000	
75%	3.000000	1.330000e+06	13.000000	3148.000000	3.000000	2.000000	2.000000	651.000000	174.000000	1999.000000	
max	10.000000	9.000000e+06	48.100000	3977.000000	20.000000	8.000000	10.000000	433014.000000	44515.000000	2018.000000	

- I, II
- II, III
- II, III, IV
- I, II, III ✓

A2- Rooms sütunu %25'lik çeyreklik değeri 2 olduğundan verilerin %75'inin 2 değerinden büyüktür. Land size sütununun minimum değeri 0 olduğundan bu sütunda 0 değerinin olduğu kesinlikle söylenebilir. BuildingArea count değeri diğer sütunların count değerinden küçük olduğu için bu sütunda eksik değerlerin olduğu kesinlikle söylenebilir. Rooms sütunun maksimum değeri 10 olduğundan ve verilerin %25 lik kısmı 3-10 arasında olduğundan 9 odalı bir ev olduğu söylenebilir.

Q3- Which one is false about overfitting and underfitting?

- Insufficient training (less epoch less batch size), causes underfitting.
- Training on too much epoch and batch size causes overfitting.
- Splitting dataset as train and test datasets will always be enough to prevent overfitting, no need for validation datasets. ✓
- In overfitting accuracy will be very good at train data but will be very bad at unseen data.

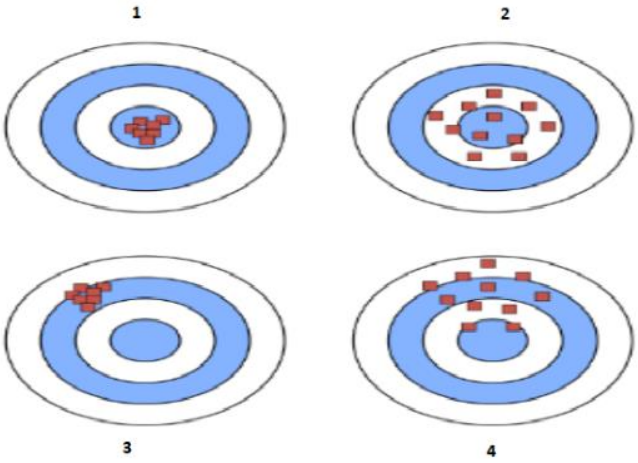
A3- Overfitting ve underfitting eğitim sırasında modelin yanlış eğitiminden kaynaklanan bir problemdir. Underfitting'de modelimiz verisetindeki veriler arasındaki patterni öğrenemeyecek kadar az eğitilir. Bu azlık hiperparametrelerin azlığı da olabilir. (Epoch, batch size vs.) Overfit ise modelin patterni öğrenemsinden çok ezberelemesine neden olacak kadar çok eğitilmesidir. Bu da yüksek hiperparametrelerin buna neden açabileceği olarak yorumlanabilir. Overfit eğitim setindeki patterni ezberleyeceği için görülmemiş verileri tahminde çok da iyi sonuç elde edemez. Öte yandan bunun önüne geçebilmek için hiperparametre ayarının yanı sıra eğitim setimiz üçe bölünebilir. Train, test ve validation sets. Burada train seteğitilir, test set'le kontrol edilip modelimiz optimize edilir. Validation set ile de modelimizin accuracy'si hesaplanır.

Q4- Which of the following is false regarding pandas and scikit-learn methods?

- DataFrame.head(x) shows x samples in the DataFrame from the beginning.
- DataFrame.describe() shows summary of the data.
- model.predict() determines how accurate the model's predictions are. ✓
- DataFrame.dropna(axis=0) drops missing values.

A4- model.predict() kurmuş olduğumuz modeli kullanarak tahminde bulunur. model.evaluate() ise bu tahminlerinin ne kadar isabetli olduğunu bulmaya yarar. Dolayısı ile C şıkkında belirtilen fonksiyonu gerçekleştiren evaluate() metodudur.

Q5- According to the shooting clusters scheme above, for each figure which statements are true? Notice that, shooting targets are the centers.



- 1:Low Bias- Low Variance 2:Low Bias-High Variance 3:High Bias-Low Variance 4: High Bias-High Variance ✓
- 1:Low Bias- High Variance 2:Low Bias-Low Variance 3:High Bias-High Variance 4: High Bias-Low Variance
- 1:High Bias- Low Variance 2: High Bias-High Variance 3:Low Bias-Low Variance 4:Low Bias-High Variance
- 1:High Bias- High Variance 2:High Bias-Low Variance 3:Low Bias-High Variance 4:Low Bias-Low Variance

A5- Bir model düşük variance ve bias değerlerine sahipse, hedefe çok yakın değerler etrafında tahmin yapar. Tam tersine, yüksek bias ve variance değerlerine sahipse hem hedefi kaçırır hem de hedef merkezi çevresinde yayılma gösterir.

Q6- According to the random forest algorithm, which of the below statements are true?

I - It is an algorithm that aims to increase the classification value by producing multiple decision trees.

II - It was created by combining Bagging and Random Subspace methods.

III - While creating the tree, it is made performance evaluation with 2/3 of the data set.

- I, III
- II, III
- I, II ✓
- I, II, III

A6- Soruya göre biliniyor ki rassal ormanlar algoritması,sınıflandırma işlemi esnasında birden fazla karar ağacı üreterek sınıflandırma değerini yükseltmeyi hedefleyen bir algoritmadır. Bu algoritma iki methodun birleştirilmesinin sonucudur. Bunlar Bagging(Boostrap aggregation) ve Random Subspace dir. Algoritmayı kullanırken veri setini 1/3'e 2/3 olarak böleriz. 2/3 olan kısım eğitim için ayrılır. geride kalan kısım test için kullanılır. Test için kullanılan kısım elbette performans değerlendirmesine yarayacaktir. Soruda 2/3 olan kısım test için verildiği gösterildiği için 3. madde yanlıştır.

Q7- What do you think about train_X when line 1 and line 2 are executed separately? The rest of the code is exactly the same.

```
Line 1. train_X, val_X, train_y, val_y = train_test_split(X, y, random_state = 2,shuffle=False)
Line 2. train_X, val_X, train_y, val_y = train_test_split(X, y, random_state = 1,shuffle=False)
```

- They generate different random number so the train_X differs from each other.
- They generate different same number and the train_X is equal to each other.
- They generate different random number so the train_X is equal to each other.
- They generate different random number, but the train_X is equal to each other. ✓

A7- random_state fonksiyonuna atanan tam sayi degerleri farkli oldugu icin farkli rastlantisal sayi uretilecektir ama shuffle'a False mantiksal ifadesi atandigi icin veri seti sirali olacaktir. Bu yuzden train_X her ikisinde ayni kalacaktir.

Q8- Trees have their length and we call that the depth of the tree. RandomForestRegressor, in scikit-learn library, has a maximum leaf (max_depth) parameter which is None as default which means nodes are expanded until all leaves are pure. What can be said if we change the number of maximum leaf nodes of a random forest?

- Length of a tree does not affect any of the results.
- Model may overfit for large depth values. ✓
- The longer tree is the better tree.
- Short trees more precise than long trees.

A8- max_depth parametresi ormandaki herhangi bir ağacın ulaşabileceği maksimum derinliği temsil eder ve bu değer her durum için değişkendir. Buna binaen ne uzun ağaçlar daha iyi ne de kısa ağaçların daha kesin sonuçlar verdiğini söylebiliriz. Söyleyebileceğimiz tek bir şey vardır ki o da gereğinden uzun bir ağaç mutlaka overfit olacaktır. Cevaptan da anlayabileceğimiz üzere ağaçların uzunluğu sonuçlarımızda rol oynuyor.

Q9- Let assume, we have a data set called home_data with 3 features names; LotArea, YearBuilt, PoolArea. How do you define non-missing values for the feature LotArea?

- non_missings = home_data["LotArea"].mean()
- non_missings = home_data.count()
- non_missings = home_data["LotArea"].count() ✓
- non_missings = home_data.mean()

A9- count() method, istenen özelliğin eksik olmayan değer sayısını döndürür.

Q10- What is the aim of the below code pieces?

```
from sklearn.metrics import mean_absolute_error

predicted_home_prices = melbourne_model.predict(X)
mean_absolute_error(y, predicted_home_prices)
```

- For splitting the data as test and train
- For interpreting the data description
- For summarizing model quality ✓
- For data modelling

A10- Mean absolute error (MAE), (ortalama mutlak hata), tahmin edilen değer ile gerçek değer arasındaki farkın (buna hata değeri diyoruz) mutlak değerini ölçmeye yarayan bir validasyon metriğidir. Bu hata ne kadar küçük ise, modelimiz o kadar iyi, yüksek doğrulukta çalışıyor diyebiliriz.