Machine Learning

ML Quiz #4

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

Solutions Submissions

- Only submit BEFORE you listen to the answers during the lecture
- Submission form

Questions A: True or False

1. The best model returned by GridSearchCV is trained on the whole dataset
2. High bias can cause an algorithm to miss the relevant relations between features and target outputs
3. Reducing the bias will always reduce the variance
4. The Bias-Variance tradeoff is irrelevant when using large amounts of data
5. Bias is the difference between the expected **prediction** and ground truth
6. Variance refers to the amount by which our prediction would change if we estimated it using a **different training** dataset
7. As model complexity increases, the variance decreases and the bias increases

Questions B: True or False

1. Cross-validation can help us find the optimal balance between bias and variance
2. When deploying a model, the same preprocessing steps used during training need to be applied to new incoming data.
3. In SKlearn, a **pipeline** includes several steps like data preprocessing, feature extraction and model training
4. Label/Ordinal encoding may result in incorrect **order** assumptions
5. Scaling data will always improve the performance of a machine learning model
6. Feature crosses are a technique that **combines** features in a way that they can have a combined effect on the **target** variable

Questions C: True or False

1. Adding **interaction features** always improves model performance
2. Feature engineering can be performed before or after data cleaning
3. Missing value imputation should always be performed on the entire dataset before splitting it
4. Scaling data changes the distribution nature/type of the data
5. RobustScaler in sklearn scales features using statistics that are robust to outliers.
6. Google **Hash Encoding** and learn it. Hash encoding always generates a **unique representation** for each category
7. Hash encoding requires the number of **unique categories** to be known beforehand
8. Hash encoding **increases the dimensionality** of the data
9. Hash encoding is particularly useful for **high cardinality** categorical features

Questions D: Code fix

- What is the bug in the below code?
- Write a fixed version of it

```
import pandas as pd
import numpy as np

df = pd.DataFrame({   'value': [1, -5, 10, -37, 60]   })
df['log_value'] = np.log1p(df['value'])

print(df)
```