Machine Learning worksheet - 8

1. (B)
2. (A)
3. (c)
4. (c)
5. (D)
6. (B)
7. (c)
8. (B) & (c)
9. (c) & (D)
10. (B)

11. The disadvantage of One Hot encoding is that for high cardinality, the feature space can really blow up quickly and you start fighting with curse of dimensionality.

- In these cases, One hot encoding is employed followed by PCA for dimensionality reduction.

(15) Techniques for handling imbalanced datasets:

→ Under Sampling.

- Random Under Sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until majority & minority classes are balanced out.

→ Over Sampling.

- It increases the number of instances in minority classes by randomly replicating them in order to present a higher representation of the minority class in sample.

→ Cluster based Over Sampling

In cluster based over sampling, K means algorithm is independently applied to minority and majority class instances.
Subsequently, each cluster is oversampled. Such that all clusters of the same class have an equal number of instances & all classes have same size.

(13)

- The key difference between ADASYN and SMOTE is that the former uses a density distribution as a criteria to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the minority samples to compensate for the skewed distributions.
- The latter generates the same number of synthetic samples for each original minority sample.