Machine Learning Assignment

(2) (i) Total Sum of Squares

TSS is a variation of the values of a dependent variable from the sample mean of the dependent variable.

Formula:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Where:

- y_i: value of the sample
- $\bar{y}$: mean value of the sample

(ii) Explained ~~Regression~~ Sum of squares

It describes how well a regression model represents the modelled data.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\hat{y}_i$ - value estimated by regression line

$\bar{y}$ - mean of sample.

(iii) Residual Sum of squares

It essentially measures the variation of modelling errors. In other words, it depicts how the variation in ~~dependable~~ dependent variable in a regression model cannot be explained by the model.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i — observed value

$\hat{y}_i$ — value estimated by regression line

(3) Regularization overcomes the following problems that occur in a model:

- Overfitting.
- Multicollinearity
- Computationally Intensive

In order to take care of the above problems, one goes for applying one of the regularization techniques.

(4). The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node and subsequent split

- Gini Impurity tells us what is the probability of misclassifying an observation.

(5). Yes decision trees are prone to overfitting especially when a tree is particularly ~~deep~~ deep.

- This is due to the amount of specifity we look at leading to smaller sample of events that meet the previous assumptions
The small sample could lead to unsound conclusion.

(6) Ensemble methods is a machine learning technique that combines several base models in order to produce ~~an~~ one optimal predictive model.

Scanned with CamScanner

(7)

Bagging	Boosting
• Method of merging same type of predictions	method of merging different type of predictions
• decreases variance, no bias & solves over-fitting issues in a model.	decreases ~~variance~~ bias but not variance
• each model receives an equal weight	models are weighed base on their performance
• Models are build independently	New models are affected by previous build model performance
• applied when classifier is unstable & have high variance	applied when classifier is stable & has high bias

(8) • Out-of-bag error also called out-of-bag estimate is a method of ~~prediction~~. measuring the prediction error or ~~loss~~ of random forest model utilizing bootstrap aggregating (bagging).

• Bootstrap Aggregating allows one to define an out-of-bag estimate of the prediction performance improvement by evaluating predictions on those

Scanned with CamScanner

observations which were not used in building of test base

9. K-fold cross validation.

- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

- The procedure has a single parameter called K that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called K-fold cross validation.

- When a specific value of K is chosen, it may be used in place of K in the reference to the model.

10. - A hyper parameter is a parameter whose value is set before the learning process begins.

- Hyper parameter tuning is crucial as they control the overall behaviour of a machine learning model.

- Every machine learning model have different hyper parameters that can be set.

6

DATE
PAGE

(11). A learning rate that is too large, will result in weight updates that will be too large and the performance of the model will oscillate over training epochs.

- When learning rate ~~is~~ is too large gradient descent can inadvertently increase rather than decrease the training error.

- When using high learning rates, it is possible to encounter a the feedback loop in which large weights induce large gradients ~~to~~

(12). No, Logistic regression is considered ~~as~~ generalized linear model because the outcome always depend on the sum of the inputs and the parameters.

- Or in other words, the output cannot depend on the product of its parameters.

- The decision boundary Ex: $\hat{y}=0.53$ is linear: It's the solution to $\theta \cdot x = 0$.

(13) • The technique of boosting uses various loss functions. Incase of ADABoost, it minimises the exponential loss function that can make the algorithm sensitive to outliers. With Gradient Boosting, any differentiable loss function can be utilised. Gradient Boosting algorithm is more robust than Adaboost.

• In the case of ADaBoost, the shifting is done by up-weighting observations that were misclassified before, while Gradient boosting identifies difficult observation by large residuals computed in previous iteration

(14) Bias variance Tradeoff

If our model is too simple and has very few parameters then it may have high bias and low variance.

On the other hand if our model has large number of parameters then it's going to have high variance & low bias.

So we need to find the right/good balance without overfitting and underfitting the data.

The tradeoff in complexity is why there is tradeoff between bias & variance. An algorithm can't be more complex & less complex at the same time.

DATE
PAGE

(15) i) Linear Kernel

It is used when data is linearly seperable, ie it can be seperated using a single line. It is one of the most common kernels to be used. It is mostly used when there are a large number of features in a particular Dataset.

ii) Polynomial Kernel

In general, it is defined as:

$$K(x_1, x_2) = (a + x_1^T x_2)^b.$$

Where
b = degree of Kernel
a = constant term

In the polynomial kernel, we simply calculate the dot product by increasing the power of kernel.

iii) Radio Base function Kernel (RBF).

RBF ~~is another~~ kernel is a function whose value depends on the distance from the origin or from some point.

Gaussian Kernel is of the following format:

$$K(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2)$$

$\|x_1 - x_2\|$ = Euclidean distance between x_1 & x_2.

Using the distance in the original space we calculate the dot product of x_1 & x_2.