→ CV is model validation for assessing how the result of a statistical analysis will ~~generate~~ generalize ~~to an~~ independent data set.
→ used in order to limits problems like overfitting

Cross Validation :-

→ Cross Validation is a technique which involves reserving a particular sample of a data set on which we don't train the ~~data~~ Model. Later we test the Model on this sample before finalizing the Model.

→ steps in Cross validation :-

① We reserve a sample data set
② Train the Model using the remaining part of data set.
③ Use the reserve sample of the data set test (validation) set. This will help us to know the effectiveness of Model performance. If our Model deliver +ve result on validation set, go ahead with current model. It rocks!

⇒ Common Method used for Cross-Validation :-

① Validation Set Approach :-

→ In this, we divided data set 50% for validation
50% for training

→ Dis :-
We train Model only on 50% data whereas may be possible we are leaving some interesting info i.e. high bias.

②

Scanned by CamScanner

② Leave one out cross validation (LOOCV) :-

→ In this approach, we reserve only one data point and train Model on rest data points. This process iterate for each data points.

→ Adv & Dis :-

- We make use of all data points i.e low bias
- We repeat the cross validation process iterate n times means high execution time.
- This approach leads to higher variation in testing Model ∵ we testing against one data point. If the data point turns out to be an outlier - It can lead to higher variation i.e high variance.

③ K-fold Cross Validation :-

- From above 2 methods - i) high bias
  ii) high variance.

So, it will take Care of both.

Steps :-

(i) Randomly split our entire data set into k "folds"
(ii) For each k folds in our data set, build our model on (k-1) folds of data set. Then test the model to check effectiveness of k^{th} fold.
(iii) Record the error we see on each predictions.
(iv) Repeat this until each of k folds has served as the test set.
(v) The avg. of our k recorded errors is called the cross validation error & will serve as our performance metric for the Model.

Scanned by CamScanner

Date ______
Page ______

⇒ How to choose right value of K :

→ lower value of K ⇒ more bias ⇒ undesirable
large " " " ⇒ more variability.

Small value of K leads to validation set approach
higher " " " " " LOOCV approach

So, always suggested K = 10

⇒ How to measure Model's bias-variance ?

→ After K-fold cross validation, we will get K- different model estimation errors ($e_1, e_2, \dots e_k$). In ideal scenerio, these error values should add to zero.

→ For Model's bias, take avg of all errors,
Lower avg value, better Model

→ For Model's variance, take standard deviation of all error.
Lower value of std, our model doesn't vary a lot with different subset of training data.

→ So, our focus on acheiving balance b/w bias & variance
So, for better predictive Model.
reduce variance, controlling bias to an extent

→ This trade-off usually leads to building less complex predictive models.

Model is too simple ⇒ Suffer from underfitting
Model is too Complex ⇒ Suffer from overfitting

Scanned by CamScanner