Machine Learning : Supervised and Unsupervised Learning

Deepak Narayan
University of New Hampshire,
Durham, New Hampshire, USA.

Model Testing and Validation

- It is not enough that the model works only on the data we trained it on.
- It is essential that it works on new data. Else it fails miserably in the real world.
- Lets look at two commonly used approaches :
  - Validation set approach : widely used
  - Cross validation : also widely used

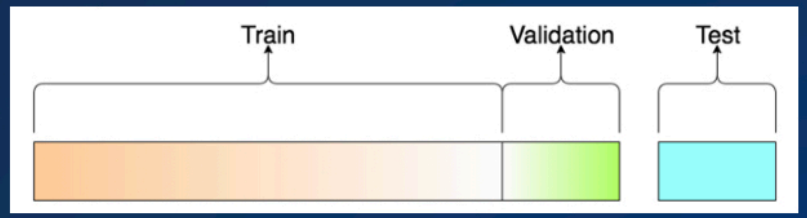
Validation Set Approach

- Training, Validation and Test sets
  - Training Set :
    - The data on which the model is built on.
    - ‘Learns’ from this data.
  - Validation Set :
    - Used to validate the model, often used to tune the parameters
    - Model does not directly ‘learn’ from this data.
    - As I said, what happens when a new input comes in, which is not in our dataset? To validate this, lets use the validation set and then validate it properly.
  - Test Set :
    - The gold standard for evaluating the model.
    - Ideally, this is only used in the final stage, and the model is evaluated how good it is, to give an idea if it would work or not, in the real world.
    - Often this contains carefully ‘sampled’ or carefully selected data which contains various data points that would be possible in a real life scenario.

Validation Set Approach

Training, Validation and Test sets splitting :

- How we split depends on the dataset.
- A common practice :
  - Data initially split into 2:
    - First part : 70-80%
    - Second part : 20-30%
    - Second part is considered as the Test set.
  - First part is split into 2:
    - First subpart : 60-70%
    - Second part : 30-40%
    - First subpart is the Training set and second part is the validation set.
- Of course, in cases where we do not have enough data, we may have to change the splitting accordingly.
- Often in Kaggle competitions, the training and validation datasets are provided to the participant, however the test set is released when the competition is about to close. It is the result on the test set that decides the winner.



Cross Validation Approach

- LOOCV and k-fold CV
- K-fold CV :
  - Split into k folds, say 4.
  - Use 3 folds for training, 1 for validation.
  - In the next iteration, use the other another 3 for training, 1 for validation and so on.
  - Finally take average of errors. $$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$
  - For classification, CV can be used. $$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i,$$
  - K fold CV reduces variance by taking the average of multiple models created, so might be better than most cases.