



Evaluating Model Performance

LEARN . GROW . EXCEL

GOPEEKRISHNAN R




Contents

- 1. Evaluating Model Performance**
 - Performance Measures of Models
 - Confusion Matrix
 - Used Performance Measures
- 2. Estimating Future Performance**
 - Cross Validation
 - The holdout Method
 - k -fold Cross Validation
 - Bootstrap Sampling
- 3. Improving Model Performance**
 - Ensembles
 - Bagging
 - Boosting
 - Random Forests

LEARN . GROW . EXCEL


GOPEEKRISHNAN R



SAINTGITS
LEARN.GROW.EXCEL

1. Evaluating Model Performance

LEARN . GROW . EXCEL GOPEEKRISHNAN R



SAINTGITS
LEARN.GROW.EXCEL

Contd...

- In machine learning, there are many classification and prediction algorithms to build models.
- Given a problem, many of these algorithms may be applicable.
- So we need to assess how good a selected algorithm is, for the given problem.
- Also, we need methods / measures to compare the performance of two or more algorithms.

LEARN . GROW . EXCEL GOPEEKRISHNAN R



Performance Measures of Models

- **Confusion matrix**
- Confusion matrix **is used to describe** the performance of a classification model or classifier.
- This is **not a** performance measure.
- **Measures of performances**
 - Accuracy and Error Rate
 - Precision and Recall
 - Sensitivity and Specificity
 - The kappa statistic
 - The F – Measure

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Confusion Matrix

- A confusion matrix **is used to describe** the performance of a classification model (aka classifier).
- A confusion matrix is a table that **categorizes predictions according to, whether they match the actual value.**
- Confusion matrices can be built for two-class classifiers as well as for multi-class classifiers.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Confusion Matrix for two-class classifiers:**
- The general representation of a confusion matrix for a two-class classifier is depicted here.

		Predicted Class	
		A	B
Actual Class	A		
	B		

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Dimension – 1 of the table:**
 - It indicates the possible categories of the **predicted values** like *Positive / Negative, Yes / No, A / B* etc.
- **Dimension – 2 of the table:**
 - It indicates the possible categories of the **actual values** like *Positive / Negative, Yes / No, A / B* etc.

		Predicted Class	
		A	B
Actual Class	A		
	B		

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- When the predicted value is the same as that of the actual value, it is a **correct classification**.
 - Correct classifications fall on the main diagonal of the confusion matrix (denoted by **O**).
- When the predicted value differs from actual value, it is an **incorrect classification**.
 - Incorrect classifications fall on the secondary diagonal of the confusion matrix (denoted by **X**).

		Predicted Class	
		A	B
Actual Class	A	O	X
	B	X	O

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Confusion Matrix for multi-class classifiers:**
- The general representation of a confusion matrix for a three-class classifier is depicted here.

		Predicted Class		
		A	B	C
Actual Class	A	O	X	X
	B	X	O	X
	C	X	X	O

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Dimension – 1 of the table:**
 - It indicates the possible categories of the **predicted values** like $A/B/C$ etc.
- **Dimension – 2 of the table:**
 - It indicates the possible categories of the **actual values** like $A/B/C$ etc.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **performance measures** for classification models **are based on the** counts of predictions falling on and off the diagonals in these tables.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **most common performance** measures consider the model's ability
 - to **distinguish** one class versus all others.
- The class of interest is known as **POSITIVE CLASS**.
- All other classes are known as **NEGATIVE CLASSES**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Examples:**
 1. The ability of the classifier to predict unseen test spam emails into the **class of Positive** and, unseen test ham emails into the **class of Negative**.
 2. The ability of a classifier to test unseen test cancer patients into the **class of Malignant** and, unseen test cancer patients into the **class of Non-Malignant (Benign)**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Understanding the Confusion Matrix:**
- We can **better understand** the Confusion Matrix with the help of an example.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Example:** Suppose that, we are given with the data of 165 examples of a particular disease.
- We are tasked with finding
 - how many actual values of the cases **are agreeing** with the prediction values.
 - prediction values \Rightarrow model has been built already and testing is also done.
- For this, we are designing a confusion matrix as follows.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

165	Predicted POSITIVE	Predicted NEGATIVE
Actual POSITIVE	100	5
Actual NEGATIVE	10	50

- The classifier made a total of 165 predictions.
- Out of those 165 patients, the classifier predicted 100 having the disease (POSITIVE); and, it predicted 50 not having the disease (NEGATIVE).

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



165	Predicted POSITIVE	Predicted NEGATIVE
Actual POSITIVE	100	5
Actual NEGATIVE	10	50

- ✓ Actually, 105 patients are having the disease and, 60 are not having the disease.
- ✓ But, the actual and predicted values of POSITIVES are agreeing on 100 examples to have the disease.
 - ✓ There is a misclassification on the prediction of 5 examples. (Out of the 105 POSITIVES, 5 have been falsely - classified NEGATIVE i.e., FALSE NEGATIVE)
- ✓ Also, the actual and predicted values of NEGATIVES are agreeing on 50 examples not to have the disease.
 - ✓ There is misclassification on the prediction of 10 examples. (Out of the 60 NEGATIVES, 10 have been falsely- classified POSITIVE i.e., FALSE POSITIVE).

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Therefore, a confusion matrix contains information about actual and predicted results given by a classifier.

	Predicted POSITIVE	Predicted NEGATIVE
Actual POSITIVE	TP	FN
Actual NEGATIVE	FP	TN

- WHERE:**

- **TP:** the number of positive examples correctly predicted.
- **FN:** the number of positive examples wrongly predicted.
- **FP:** the number of negative examples wrongly predicted.
- **TN:** the number of negative examples correctly predicted.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Prediction Accuracy and Error Rate:**
- With a 2 x 2 confusion matrix, **Prediction Accuracy** (aka. **Success Rate**) is the proportion of correctly classified examples and, is defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- In this formula, the terms *TP*, *TN*, *FP* and *FN* refer to the number of times the model's predictions fell into each of these categories.
- The accuracy is therefore, the proportion that represents the number of true positives and true negatives, divided by the total number of predictions.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **Error Rate** or, the proportion of incorrectly classified examples is defined as:

$$\text{error rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

- In this formula, the terms TP , TN , FP and FN refer to the number of times the model's predictions fell into each of these categories.
- The error rate is therefore, the proportion that represents the number of false positives and false negatives, divided by the total number of predictions.
- It is also computed using **$\text{error rate} = 1 - \text{accuracy}$** .

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- $\text{error rate} = 1 - \text{accuracy}$** tells us that, a model that is correct 95% of the time is incorrect 5% of the time.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Precision and Recall:**
- The **precision** of a model is **defined as** the **proportion** of **positive examples** that are **truly positive**.
 - In other words, it **is the fraction of** relevant instances among the retrieved instances.
- It is computed using

$$\text{Precision} = \frac{TP}{TP + FP}$$

	Predicted POSITIVE	Predicted NEGATIVE
Actual POSITIVE	TP	FN
Actual NEGATIVE	FP	TN

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **recall** of a model is **defined as** the **proportion** of **true positives** divided by the **total number of positives**.
 - In other words, **it tells us** about how complete the results are.
- It is computed using

$$\text{recall} = \frac{TP}{TP + FN}$$

	Predicted POSITIVE	Predicted NEGATIVE
Actual POSITIVE	TP	FN
Actual NEGATIVE	FP	TN

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Problem:** A database contains 90 records on a particular topic of which 55 are **relevant** to a certain investigation. A search was conducted on that topic and 50 records were retrieved. Of the 50 records retrieved, 40 were relevant. Construct the confusion matrix for the search and calculate the precision and recall scores for the search.

90	Predicted RELEVANT	Predicted IRRELEVANT
Actual RELEVANT	40	15
Actual IRRELEVANT	10	25

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **TP** = 40
- **FP** = 10
- **FN** = 15
- **TN** = 25

$$\bullet \text{ Precision} = \frac{TP}{TP+FP} = \frac{40}{40+10} = \frac{40}{50} = 80\%$$

$$\bullet \text{ Recall} = \frac{TP}{TP+FN} = \frac{40}{40+15} = \frac{40}{55} = 72\%$$

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Problem:** Consider a computer **program** for recognizing dogs (the relevant element) in a digital photograph. Upon processing a picture **which contains** ten (10) cats and twelve (12) dogs, the **program identifies** eight dogs (8). Of the eight (8) elements **identified as dogs (Predicted DOGS)**, only five (5) **actually** are dogs, while the other three (3) are cats.
- Find **Precision** and **Recall**.

22	Predicted DOGS	Predicted CATS
Actual DOGS	5	7
Actual CATS	3	7

- Precision = 5 / 8
- Recall = 5 / 12

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Sensitivity and Specificity:**
- The **sensitivity** of a model is **defined as** the proportion of **positive examples that were correctly classified**.
- It is computed using

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- That is, sensitivity is computed as the number of true positives divided by the sum of true positives (number of correct classifications) and false negatives (number of incorrect classifications).
- From the formulation, it is clear that **sensitivity is the same as that of recall**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **specificity** of a model is **defined as** the proportion of **negative examples that were correctly classified**.
- It is computed using

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- That is, specificity is computed as the number of true negatives divided by the sum of true negatives (number of correct classifications) and false positives (number of incorrect classifications).

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **The Kappa Statistic (Cohen's Kappa Statistic)**
- Kappa statistic is used to find an agreement between two dimensions (for instance, between "Predicted" and "Actual").
- Kappa values range from **0** to a maximum of **1**.
 - A **kappa value = 1** **indicates**, there is perfect agreement between the model's predictions and the true values.
 - A **kappa value < 1** **indicates**, there is imperfect agreement between the model's predictions and the true values.
- The interpretation of the Kappa value is briefed in the following table.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

Kappa Statistic Value	Interpretation
Less than 0.20	Poor Agreement
0.20 to 0.40	Fair Agreement
0.40 to 0.60	Moderate Agreement
0.60 to 0.80	Good Agreement
0.80 to 1.00	Very Good Agreement

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The Kappa Statistic is computed using:

$$\kappa = \frac{p_r(a) - p_r(e)}{1 - p_r(e)}$$

- Where:

- $p_r(a)$ — refers to the proportion of actual agreement.
- $p_r(e)$ — refers to the expected agreement between the predictor and the true values.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The F – Measure
- It **gives a measure of** model performance that combines precision and recall into a single number.
- It is computed as

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$= \frac{2 \times TP}{2 \times TP + FP + FN}$$

- The F – Measure combines the precision and recall using harmonic mean.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



2. Estimating Future Performance

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Usually Statistical machine learning packages (created in **R**, **Weka**, **Python**, **Matlab** etc.) construct confusion matrices and other performance measures **during the model building process**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Assume, **we are given with the data** for model building as well as for model evaluation.
 - The **same data is used** for model building and for model evaluation.
 - This process is known as **resubstitution evaluation**.
- The error that is got from such an evaluation is called **resubstitution error**.
 - It is the error of predicted outcome **vs.** the actual value from the **same** training data set.
- Error on the (training) data is ***not*** a good indicator of performance on future data.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- How **to get rid of** resubstitution error?
 - **SPLIT THE DATA.**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- We split the given data into two:
 1. **Training Data Set**
 2. **Test Data Set**
- Generally, **Cross-Validation** is used to achieve this, that we are going to discuss next.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Cross Validation

- Cross-validation is a **technique** to evaluate predictive models by partitioning the original sample into
 - a **training data set** to train the model, and
 - a **test data set** to evaluate it.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Three methods of cross validation are discussed here.
 1. **The Holdout Method**
 2. ***k*-fold Cross Validation**
 3. **Bootstrap Sampling**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **holdout method** is the simplest kind of cross validation.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In **holdout method**
- **Step-1:** the data set is split into two – a **training data set** and a **test data set**.

```
X_Train,X_Test,y_train,y_test=train_test_split(X,y,test_size=.20,random_state=42)
```

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Step-2:** The Machine Learning Algorithm **fits a function** using the **training data set** only.

```
model=SVC(kernel='poly',degree=8)  
model.fit(X_Train,y_train)
```

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Step-3:** Then, the **function is used to predict** the output values for unseen data in the **test data set**.

```
y_pred=model.predict(X_Test)
```

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Step-4:** The errors the function makes are used to evaluate the model.

```
pd.concat([X_Test,y_test,pd.Series(y_pred,name='Predicted',index=X_Test.index)],
          ignore_index=False,axis=1)
```

	sepalength	sepalwidth	petallength	petalwidth	species	Predicted
73	6.1	2.8	4.7	1.2	Iris-versicolor	Iris-virginica
18	5.7	3.8	1.7	0.3	Iris-setosa	Iris-setosa
118	7.7	2.6	6.9	2.3	Iris-virginica	Iris-virginica
78	6.0	2.9	4.5	1.5	Iris-versicolor	Iris-versicolor
76	6.0	2.9	4.9	1.4	Iris-versicolor	Iris-versicolor

```
accSVC=metrics.accuracy_score(y_test,y_pred)
```

```
print('Accuracy Score of Support Vector Classifier: {0:.2f}%'.format(accSVC*100))
```

Accuracy Score of Support Vector Classifier: 96.67%

LEARN . GROW . EXCEL

GOPEEKRISHNAN R

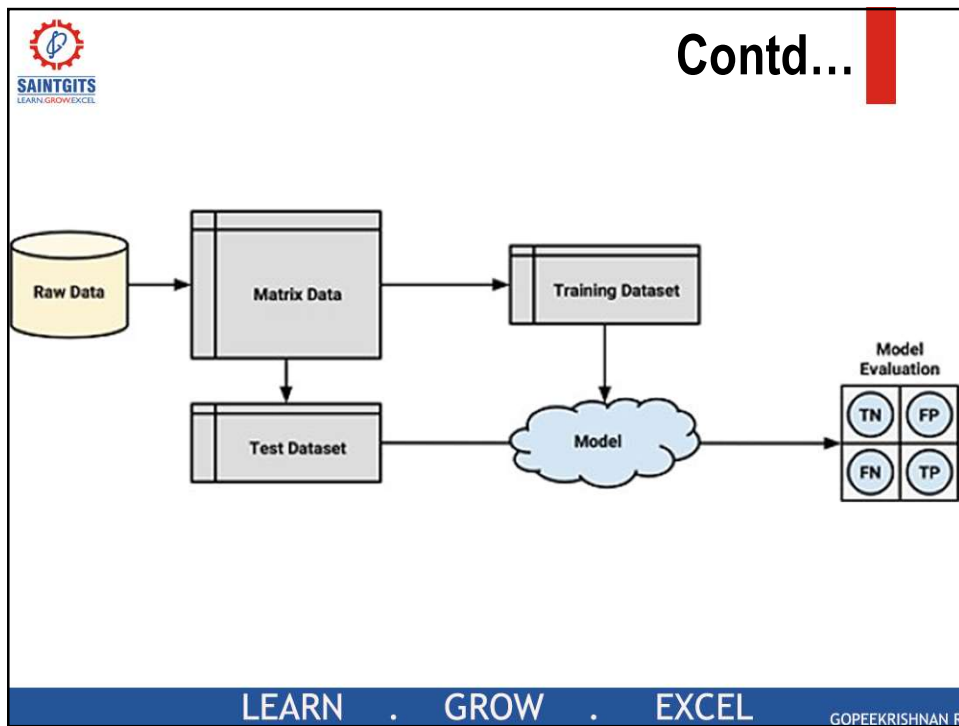


1. The Holdout Method

- Dividing the given data (sample) into training data set and test data set is known as the **holdout method**.
- This is illustrated below.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Typically, about **two – third** ($\frac{2}{3} = 67\%$) is used for training and, **one – third** ($\frac{1}{3} = 33\%$) of the data is **held out** for testing.
- **The larger** the training data set, **the better** the classifier.
- **The larger** the test data set, **the more accurate** the performance measure estimates become.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- But this proportion **can vary** depending on the amount of available data.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Suppose** that, **we built several models** on the training data.
- Then, **we select** the one with the highest accuracy on the test data (this means the testing has to be performed one by one on to each model).
- To **avoid repeated testing**, we can adopt another holdout method...that requires the data set to be split into three.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- That is, **in addition to** the **training data set** and **test data set**, we can use **validation data sets** also.
- Thus we have the following data sets.
 - **Training Data Set**
 - **Test Data Set**
 - **Validation Data Set.**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R

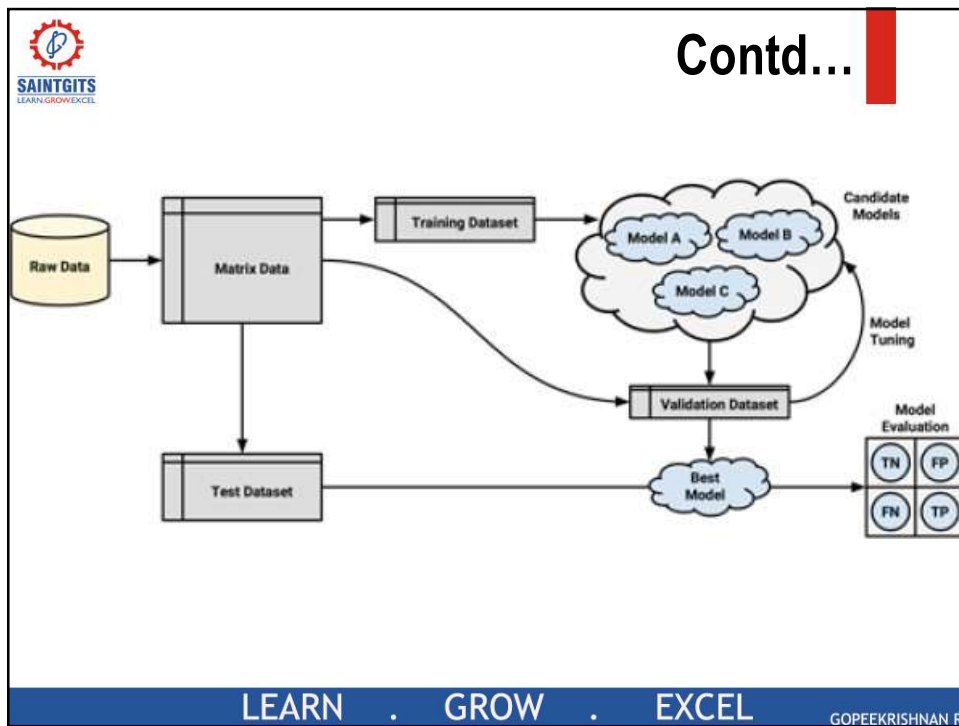


Contd...

- The **validation data set would be used for** iterating and refining the model or models chosen.
- Thus, **we use the test dataset only once** as a final step to report an estimated error rate for future predictions.
- A typical split between training, test and validation would be 50%, 25% and 25% respectively.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- But there are **certain issues** that may crop up when we split the data set like this.
- Some of the **reported problems** are
 1. *Uneven distribution of different classes of data among splits.*
 2. *Huge reservation of data to both test data set and validation data set.*

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **In this first problem** with holdout sampling is that, each partition (like training data / testing data / validation data) may have a larger or smaller proportion of some classes.
- **Example:** Training data may have a smaller number of examples which belong to "**Class-C**", while the remaining are belonging to "**Class-A**", in a classification problem involving three classes viz. **Class-A**, **Class-B** and **Class-C**.
- In certain cases, this can lead a class to be omitted from the training data set.
 - In the above example, **Class-B** is completely omitted from the training data set.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- This is **a significant problem**, because the model will not be able to learn from this.
 - Especially, for the above example, the trained model **has not learnt anything** about examples belonging to **Class-B**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- To avoid this from occurring, a technique **called stratified random sampling** can be used.
- **Stratified random sampling guarantees that** the random partitions **have nearly the same proportion of each class** as the full dataset, even when some classes are small.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **This means**
 - Training Data Set has nearly the **same proportion** (**not same number**) of **Class – A** examples as is with Testing Data Set and Validation Data Set.
 - Training Data Set has nearly the **same proportion** of **Class – B** examples as is with Testing Data Set and Validation Data Set.
 - Training Data Set has nearly the **same proportion** of **Class – C** examples as is with Testing Data Set and Validation Data Set.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In the **second problem** with the holdout method, substantial portions of the data must be reserved for test data set and validation data set. ($25 + 25 = 50\%$).
- Since these data **cannot be used** to train the model **until** its performance has been measured (*with the already chosen training data set*), the performance estimates are likely to be **too low...**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- One technique called **repeated holdout** is sometimes used to solve such problems.
- Here, it **uses the average result** from several random holdout samples to evaluate the model's performance.

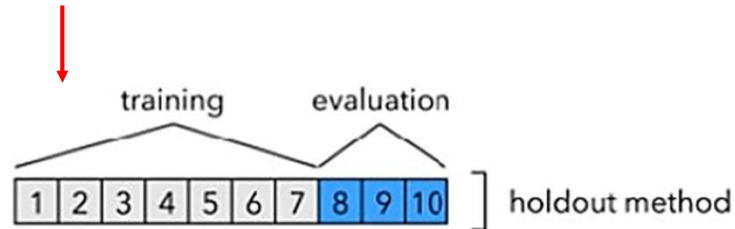
LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **simple holdout** can be depicted as shown here.



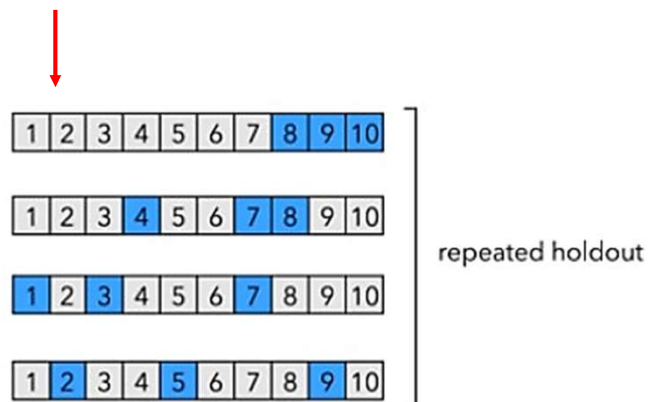
LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **repeated holdout** can be depicted as shown here.



LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- It is evident that, in **repeated holdout**, the same record could be used more than once for testing.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



2. k – fold Cross Validation (k -fold CV)

- **Repeated holdout** is the basis for k – fold cross validation.
- The **problem with the repeated holdout**: the same record could be used more than once for testing.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In k -fold cross validation, the given data set X (i.e., complete data set) is divided randomly into k equal sized parts, $X_i, i = 1, 2, 3, \dots, k$.
- To generate each pair of training data set and validation data set
 - we keep the **one of the k parts** as the validation data set V_i and
 - combine the **remaining $(k - 1)$ parts** to form the training set T_i

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Doing this process k times, each time **leaving out** another one of the k parts out for validation, we get the pairs (V_i, T_i)
- That is,

$$V_1 = X_1, T_1 = X_2 \cup X_3 \cup X_4 \dots \cup X_k$$

$$V_2 = X_2, T_2 = X_1 \cup X_3 \cup X_4 \dots \cup X_k$$

$$V_3 = X_3, T_3 = X_1 \cup X_2 \cup X_4 \dots \cup X_k$$

...

$$V_k = X_k, T_k = X_1 \cup X_2 \cup X_3 \dots \cup X_{k-1}$$

- The usual k value is 10 or 30 thus forming **10-fold cross validation** or **30-fold cross validation**.

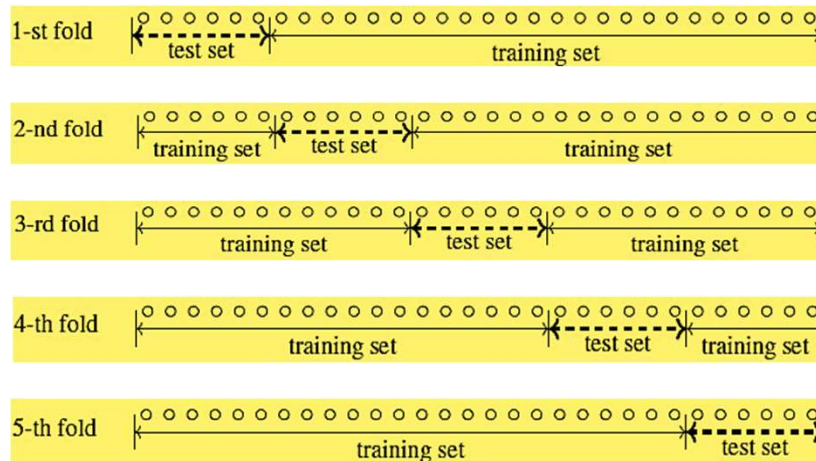
LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Example:** An iteration of a 5-fold cross validation is illustrated here.



LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- After the process of training and evaluating the model for k times, the **average performance** across all the folds is reported.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



3. Bootstrap Sampling

- This is a **less frequently used alternative** of k -fold cross validation.
- It is also called **bootstrap** or **bootstrapping**.
- Generally speaking, **this refers to statistical methods applied to "small" random samples to estimate the properties of a large population.**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In this method, **not the complete examples are used** for both training and testing.
 - That is, **only a random sample of the population** (of examples) **will be chosen** for both training and testing.
- **Then onto this small samples, some statistical measures are applied to estimate the performance measures.**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The “above procedure” **is repeated** for several such random training *and* test samples.
- The results from various random samples **are then averaged** to obtain **a final estimate** of future performance.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- But, **one drawback** of this method is that...
 - several examples **can / may be selected** **multiple times** for both training and testing.
 - this is known as **sampling with replacement**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In this sampling with replacement,
 - it is estimated that, **the probability** of any given instance **is included in the training data set** is **63.2%**.
 - also, it is estimated that, **the probability** of any given instance **is included in the test data set** is **36.8%**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **In contrast** to the k -fold cross validation, where 90% of the examples were used for training, the bootstrap sample is less representative of the full data set.
 - i.e., the training **is got to only 63.2%** of the entire data set.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Therefore, in terms of the performance estimates of bootstrap sampling, the model's performance will be **much lower than what would be obtained** when the model is trained on the full data set.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R




Contd...

- A **special case** of bootstrapping known as the **0.632 bootstrap** takes care of this low performance.
- **How it is taken care?**
 - It is by calculating the final performance measure as a function of performance on both the training data and the test data.
 - The final error rate is then estimated as

$$\text{error} = 0.632 \times \text{errortest} + 0.368 \times \text{errortrain}$$

LEARN . GROW . EXCEL


GOPEEKRISHNAN R



SAINTGITS
LEARN · GROW · EXCEL

3. Improving Model Performance

LEARN · GROW · EXCEL GOPEEKRISHNAN R



SAINTGITS
LEARN · GROW · EXCEL

Contd...

- We know, there are several algorithms for learning the same task.
- No single learning algorithm ever produces the most accurate output.
- So, here, we are trying to **combine multiple learners** to attain higher accuracy values.

LEARN · GROW · EXCEL GOPEEKRISHNAN R



Contd...

- When many learning algorithms are combined, the individual algorithms are called the **base learners** of the collection.
- When we use multiple base learners...
 - we want each of them to be reasonably accurate.
 - this means we do not require them to be very accurate individually.
 - the base learners **are chosen** out of their **simplicity only**, not on their accuracy.
 - what we care for, is the final accuracy when the base learners are combined.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The technique of combining and managing the predictions of multiple base learners (*aka. **weak learners***) into a powerful learner is known as **meta learning**.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

• How we are selecting the base learners?

1. We can use different learning algorithms.
 - like Naïve Bayes algorithm, ANNs, SVMs etc.
2. We can use the same algorithm with different hyperparameters.
 - for example, in the case of ANNs, the hyperparameters include number of input nodes, number of hidden layers, number of nodes in each hidden layer etc.
3. We can use different representations of the input object.
 - for instance, in speech recognition applications, to recognize uttered words, words may be in acoustic form or words can be understood from the video frames of the lip movement of the speaker etc.
4. We can use different training data sets to train different base learners.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Ensembles

- The word **ensemble** literally means “*a group of things or people acting or taken together as a whole, especially a group of musicians who regularly play together.*”



LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In machine learning, **ensemble / ensembling** is the art of combining diverse set of learners (individual models) together to improve predictive power of the (combined) model.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The members of the ensemble **might be predicting**
 - real-valued numbers
 - class labels
 - posterior probabilities
 - rankings
 - clusterings *or*
 - any other quantity.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The way we combine all the predictions (of the individual learners) together is termed as **ensemble learning**.
- This is not an easy task.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R

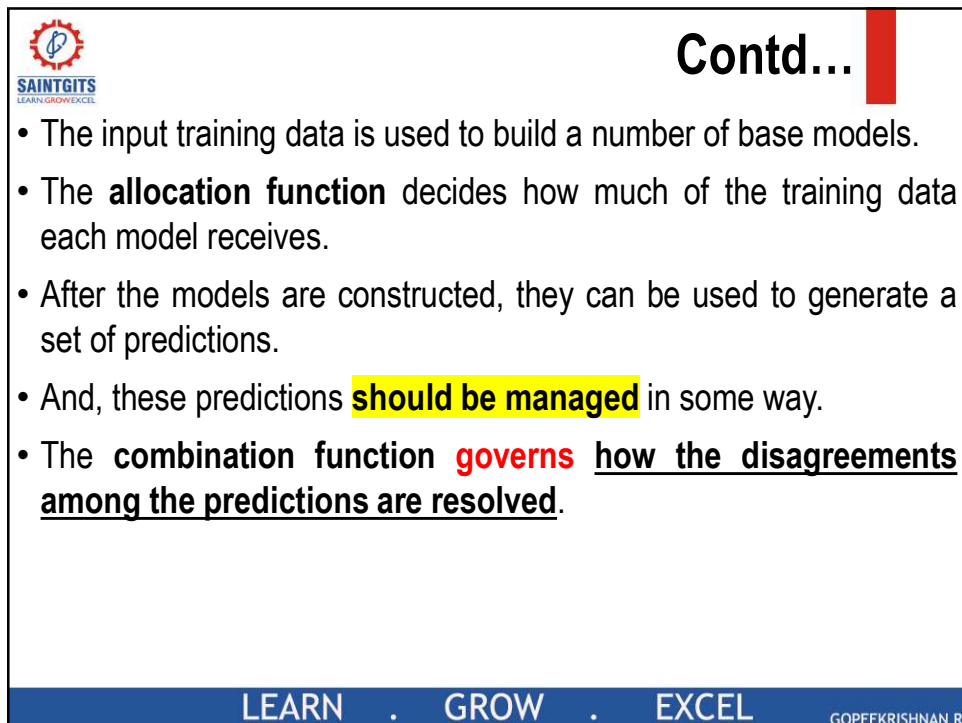
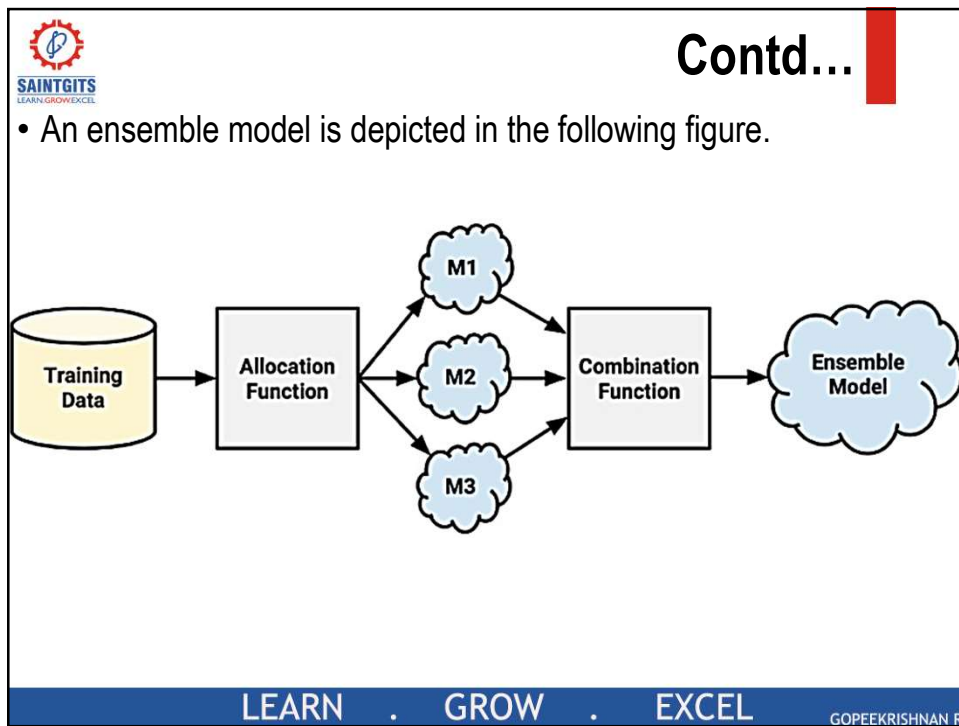


Contd...

- In machine learning, an ensemble learning method **consists of** the following two steps:
 1. **Create different models**, for solving a particular problem using a given data.
 2. **Combine the models created**, to produce improved results (ensembling + ensemble modeling).

LEARN . GROW . EXCEL

GOPEEKRISHNAN R





Contd...

- The **different models** may be chosen in many different ways:
 - The models may be created using appropriate **different** algorithms like k-NN algorithm, Naive-Bayes algorithm, Decision tree algorithm, etc.
 - The models may be created by using the same algorithm (for example, k-NN algorithm only) but using **different** splits of the same dataset into training data and test data.
 - The models may be created by assigning **different** initial values to the parameters in the algorithm as in ANN algorithms.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The **predictions** from the models created in the ensemble modeling **are combined** in several ways as described below.
 - **Majority Voting**
 - **Weighted Voting**
 - **Simple Averaging**
 - **Weighted Averaging**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

• Majority Voting:

- Every model makes a prediction (vote) for each example, and the final output prediction is the one that receives more than half of the votes.
- **Example:** a classifier that classifies test examples into Class A / B / C. Each test example is classified into its correct class.
 - The **final output prediction from this classifier is B** when
 - 52% of the examples belong to class B,
 - 31% of the examples belong to class C,
 - while only 17% of the examples belong to class A.
 - This means **the classifier has voted for class B examples, the most.**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

• Weighted Voting:

- In majority voting, each model has the same right to vote.
- Here, **we increase the right to vote** of one or more models **by increasing their importance.**
- Importance of the models is increased by giving suitable weights to them.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Simple Averaging:**

- In simple averaging method, for every example of the test data set, the average predictions are calculated.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Weighted Averaging:**

- Weighted averaging is a slightly modified version of simple averaging.
- Here, the prediction of each model is multiplied by appropriate weights and, then their average is calculated.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- There are many ensemble based learning methods.
- Important three methods are discussed here.
- They are
 1. **Bagging**
 2. **Boosting**
 3. **Random Forests**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Bagging

- Bagging is an ensemble learning method.
- Bagging stands for bootstrap aggregating.
- The **learning method** (combining of the base learners) used here is **voting**.
- The base learners **are made different** by **training each model with different training data sets**.
 - Different training data sets are given to the base learners using bootstrapping.
- This means, the **same learning algorithm will be used** for training different training data sets.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- The model's predictions are combined using
 - **voting** (*for classification*) and,
 - **averaging** (*for numeric prediction*).

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Bagging is usually used with **unstable learners**.
- The unstable learners are those learners that tend to change (classification / prediction) **substantially**, when the input data changes only slightly.
- For this reason, bagging **is often used with decision tree learning**, **which have the tendency to vary dramatically**, given minor changes in the input data.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Boosting

- Boosting is an ensemble learning method.
- Here, **we** actively **try to generate complementary base-learners** by training the next learner on the mistakes done by the previous learners.

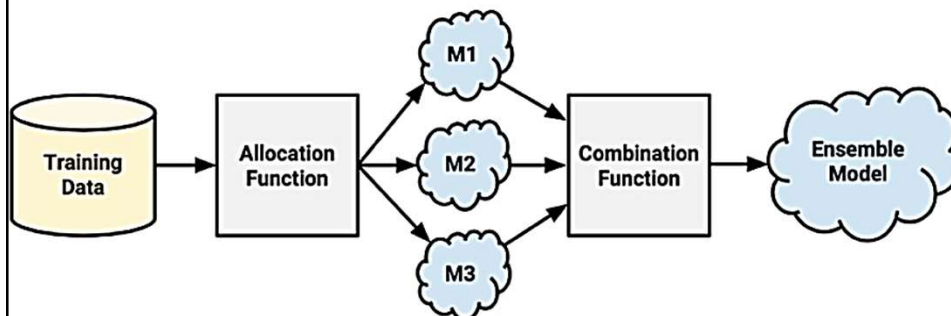
LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- If a data point is **incorrectly** classified / predicted by the first model (M1), and, probably by all the models that follow (M2 and M3, here) in an ensemble, **will combining** the classifications / predictions **give us better results?**
- Such situations are taken care of by **boosting**.



LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- In machine learning, the term **boosting** **refers to** a family of algorithms which converts weak learner to strong learners.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Working:**

1. The first base learner takes all the available data and assign equal weights to all examples, and starts classification / prediction.
2. If there is any classification / prediction **error** caused by the first learning algorithm, then **we give higher attention to the examples having prediction error**. Then, we apply the next base learning algorithm.
3. Iterate step 2 till higher accuracy is reached or the limit of the base learning algorithms is reached.
4. Finally, **it combines the outputs from weak learners** and creates a strong learner which eventually improves the prediction power of the model.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- There are many boosting algorithms which use other types of engine such as:
 1. **AdaBoost (Adaptive Boosting)**
 2. **Gradient Tree Boosting**
 3. **XGBoost**

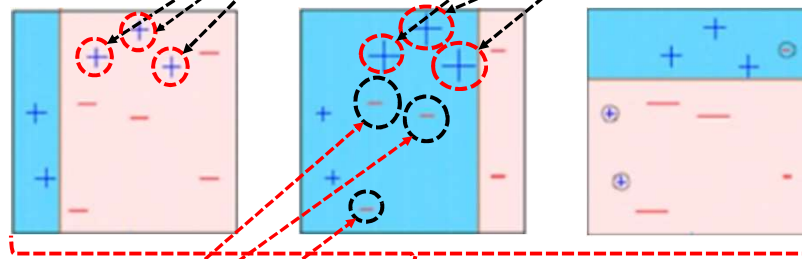
LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Illustration using AdaBoost:



Misclassifications

Combining outputs
from weak
base learners

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Random Forest

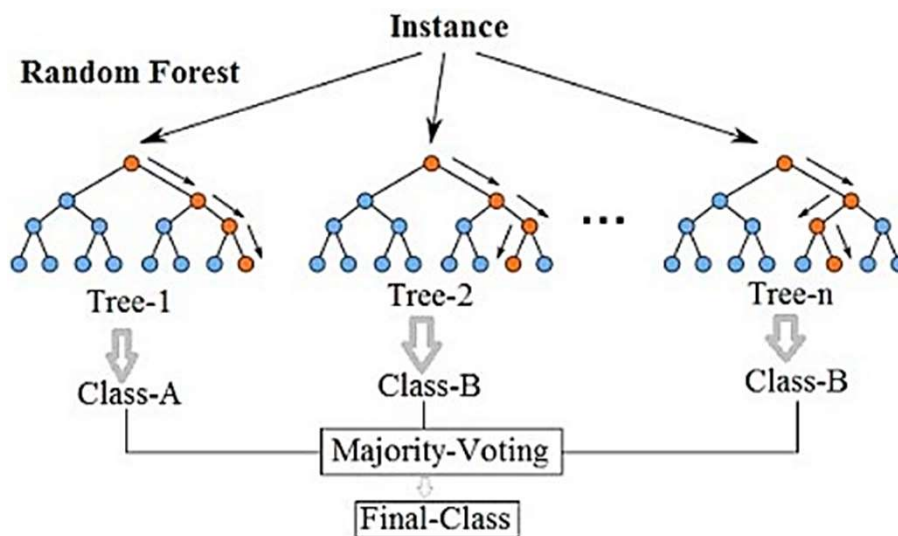
- A random forest is an ensemble learning method.
- Here, multiple decision trees are constructed.
- That, is random forests focusses only on decision tree base learners.
- After the ensemble of trees (i.e., the forest) is generated, the model uses a vote to combine the trees' predictions.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...



LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- Follows is **an outline** of the **random forest learning algorithm**.

1. Training Phase

- The random forest algorithm generates many decision trees. **Each tree** is generated as follows:
 - From the original data set, take N training examples at random – with replacement. This is the training set.
 - Assume, there are M features available in the data set. Out of M , a number m is specified such that **at each node, m features are selected at random** and, the best among these m is used to split the node. The value of m is held constant during the generation of the various trees in the forest.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

2. Testing Phase

- To classify an unseen test example from an input vector, put the input vector down each of the trees in the forest.
- Each tree gives a classification, and we say the tree “votes” for that class.
- The class for which majority of votes is obtained, is considered as the class of the test example.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



Contd...

- **Self-Study:**

- Strengths and Weaknesses of Random Forests.

LEARN . GROW . EXCEL

GOPEEKRISHNAN R



References

1. **Machine Learning with R, Second Edition, Brett Lantz, PACKT Publishing.**

LEARN . GROW . EXCEL

GOPEEKRISHNAN R