

INTRODUCTION TO DATA ANALYTICS

Class # 22

Sensitivity Analysis

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

TOPICS COVERED IN THIS PRESENTATION

- Introduction
- Estimation Strategies
- Accuracy Estimation
- Error Estimation
- Statistical Estimation
- Performance Estimation
- ROC Curve

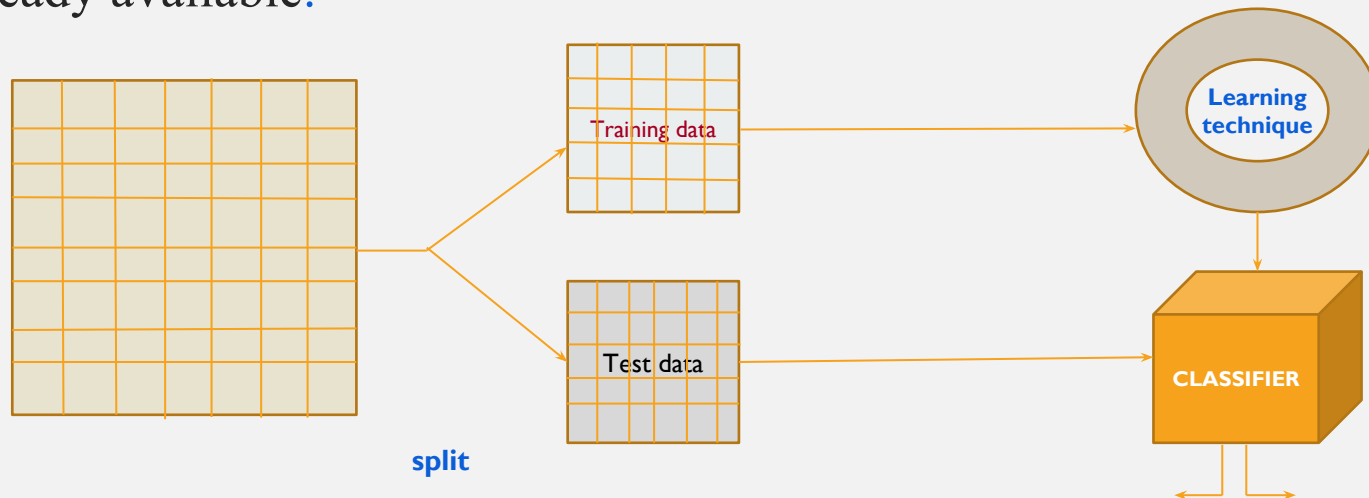
INTRODUCTION

- A classifier is used to predict an outcome of a test data
 - Such a prediction is useful in many applications
 - Business forecasting, cause-and-effect analysis, etc.
 - A number of classifiers have been evolved to support the activities.
 - Each has their own merits and demerits
- There is a need to estimate the accuracy and performance of the classifier with respect to few controlling parameters in data sensitivity
- As a task of sensitivity analysis, we have to focus on
 - Estimation strategy
 - Metrics for measuring accuracy
 - Metrics for measuring performance

Estimation Strategy

PLANNING FOR ESTIMATION

- Using some “**training data**”, building a classifier based on certain principle is called “**learning a classifier**”.
- After building a classifier and before using it for classification of unseen instance, we have to validate it using some “**test data**”.
- Usually training data and test data are outsourced from a large pool of data already available.



ESTIMATION STRATEGIES

- Accuracy and performance measurement should follow a strategy. As the topic is important, many strategies have been advocated so far. Most widely used strategies are
 - Holdout method
 - Random subsampling
 - Cross-validation
 - Bootstrap approach

HOLDOUT METHOD

- This is a basic concept of estimating a prediction.
- Given a dataset, it is partitioned into two disjoint sets called training set and testing set.
- Classifier is learned based on the training set and get evaluated with testing set.
- Proportion of training and testing sets is at the discretion of analyst; typically 1:1 or 2:1, and there is a trade-off between these sizes of these two sets.
- If the training set is too large, then model may be good enough, but estimation may be less reliable due to small testing set and vice-versa.

RANDOM SUBSAMPLING

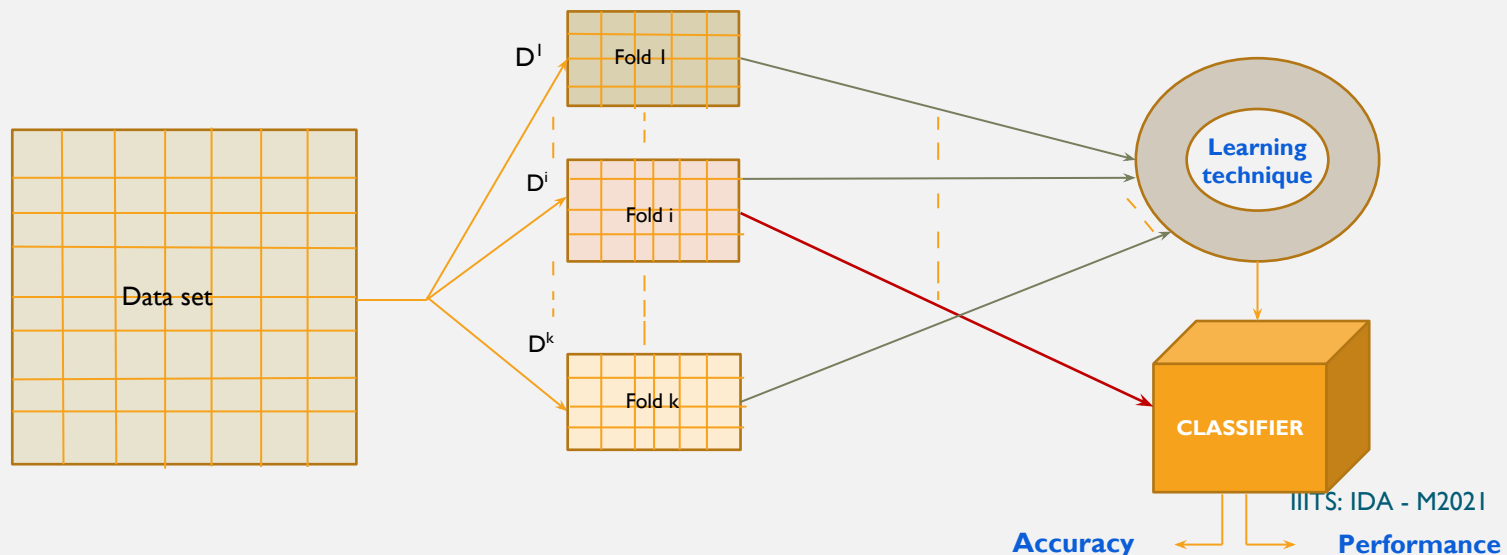
- It is a **variation** of Holdout method to overcome the **drawback of over-presenting a class** in one set thus under-presenting it in the other set and vice-versa.
- In this method, **Holdout method is repeated k times**, and in each time, two **disjoint sets are chosen at random** with **a predefined sizes**.
- Overall estimation is taken **as the average of estimations** obtained from each iteration.

CROSS-VALIDATION

- The main drawback of Random subsampling is, it does not have control over the number of times each tuple is used for training and testing.
- Cross-validation is proposed to overcome this problem.
- There are two variations in the cross-validation method.
 - k-fold cross-validation
 - N-fold cross-validation

K-FOLD CROSS-VALIDATION

- Dataset consisting of N tuples is divided into k (usually, 5 or 10) equal, mutually exclusive parts or folds (D_1, D_2, \dots, D_k), and if N is not divisible by k , then the last part will have fewer tuples than other $(k-1)$ parts.
- A series of k runs is carried out with this decomposition, and in i^{th} iteration D_i is used as test data and other folds as training data
 - Thus, **each tuple is used same number of times for training and once for testing.**
- Overall estimate is taken as the average of estimates obtained from each iteration.



N-FOLD CROSS-VALIDATION

- In k -fold cross-validation method, $\frac{k-1}{N}$ part of the given data is used in training with k -tests.
- N -fold cross-validation is an **extreme case** of k -fold cross validation, often known as **“Leave-one-out” cross-validation**.
- Here, dataset is divided into as many folds as there are instances; thus, all most each tuple forming a training set, building N classifiers.
- In this method, therefore, N classifiers are built from $N-1$ instances, and each tuple is used to classify a single test instances.
- Test sets are mutually exclusive and effectively cover the entire set (in sequence). This is as if **trained by entire data as well as tested by entire data** set.
- Overall estimation is then averaged out of the results of N classifiers.

N-FOLD CROSS-VALIDATION : ISSUE

- So far the estimation of accuracy and performance of a classifier model is concerned, the *N*-fold cross-validation is comparable to the others we have just discussed.
- The drawback of *N*-fold cross validation strategy is that it is computationally expensive, as here we have to repeat the run *N* times; this is particularly true when data set is large.
- In practice, the method is extremely beneficial with very small data set only, where as much data as possible to need to be used to train a classifier.

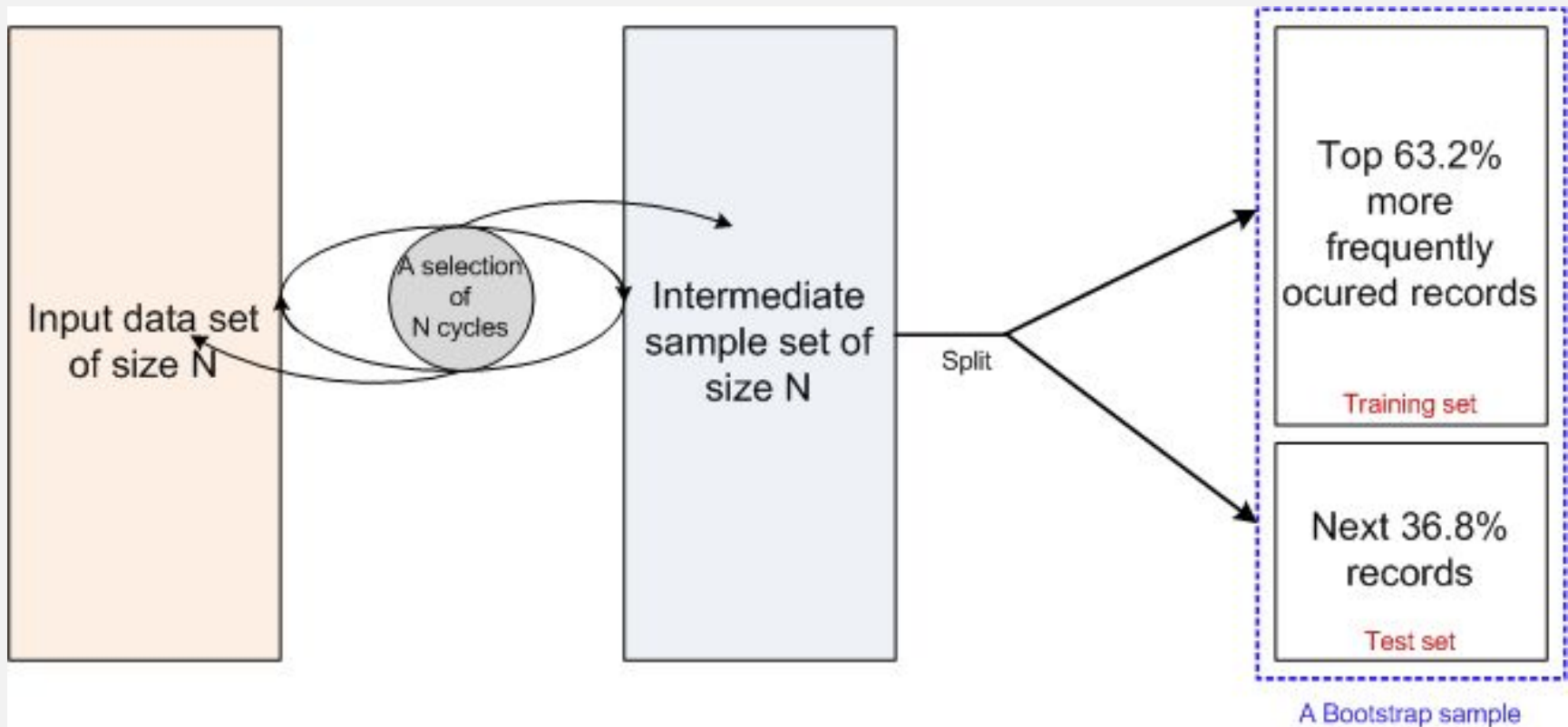
BOOTSTRAP METHOD

- The Bootstrap method is a variation of **repeated version of Random sampling method**.
- The method suggests the **sampling of training records with replacement**.
 - Each time a record is selected for training set, is put back into the original pool of records, so that it is equally likely to be redrawn in the next run.
 - In other words, the Bootstrap method samples the given data set **uniformly with replacement**.
- The rational of having this strategy is that let some records be occur **more than once** in the samples of both training as well as testing.
 - **What is the probability that a record will be selected more than once?**

BOOTSTRAP METHOD

- Suppose, we have given a data set of N records. The data set is sampled N times with replacement, resulting in a bootstrap sample (i.e., training set) of I samples.
 - Note that the entire runs are called a bootstrap sample in this method.
- There are certain chance (i.e., probability) that a particular tuple occurs **one or more** times in the training set
 - If they do not appear in the training set, then they will end up in the test set.
 - Each tuple has a probability of being selected $\frac{1}{N}$ (and the probability of not being selected is $\left(1 - \frac{1}{N}\right)$).
 - We have to select N times, so the probability that a record will not be chosen during the whole run is $\left(1 - \frac{1}{N}\right)^N$
 - Thus, the probability that a record is chosen by a bootstrap sample is $1 - \left(1 - \frac{1}{N}\right)^N$
 - For a large value of N , it can be proved that $\left(1 - \frac{1}{N}\right)^N \approx e^{-1}$
 - **Thus, the probability that a record chosen in a bootstrap sample is $1 - e^{-1} = 0.632$**

BOOTSTRAP METHOD : IMPLICATION



- This is why, the Bootstrap method is also known as 0.632 bootstrap method

Accuracy Estimation

ACCURACY ESTIMATION

- We have learned how a classifier system can be tested. Next, we are to learn the metrics with which a classifier should be estimated.
- There are mainly two things to be measured for a given classifier
 - Accuracy
 - Performance
- **Accuracy estimation**
 - If N is the number of instances with which a classifier is tested and p is the number of correctly classified instances, the accuracy can be denoted as

$$\epsilon = \frac{p}{N}$$

- Also, we can say the **error rate** (i.e., misclassification rate) denoted by $\bar{\epsilon}$ is denoted by
$$\bar{\epsilon} = 1 - \epsilon$$

ACCURACY : TRUE AND PREDICTIVE

- Now, this accuracy may be **true (or absolute) accuracy** or **predicted (or optimistic) accuracy**.
- **True accuracy** of a classifier is the accuracy when the classifier is tested with **all possible unseen instances** in the given classification space.
 - However, the number of possible unseen instances is potentially very large (if it is not infinite)
 - For example, classifying a hand-written character
 - Hence, measuring the true accuracy beyond the dispute is impractical.
- **Predictive accuracy** of a classifier is an **accuracy estimation for a given test data** (which are mutually exclusive with training data).
 - If the predictive accuracy for test set is ϵ and if we test the classifier with a different test set it is very likely that a different accuracy would be obtained.
 - The predictive accuracy when estimated with a given test set it should be acceptable without any objection

PREDICTIVE ACCURACY

Example 21.1 : Universality of predictive accuracy

- Consider a classifier model M^D developed with a training set D using an algorithm M .
- Two predictive accuracies when M^D is estimated with two different training sets T_1 and T_2 are

$$(M^D)_{T_1} = 95\%$$

$$(M^D)_{T_2} = 70\%$$

- Further, assume the size of T_1 and T_2 are

$$|T_1| = 100 \text{ records}$$

$$|T_2| = 5000 \text{ records.}$$

- Based on the above mentioned estimations, neither estimation is acceptable beyond doubt.

PREDICTIVE ACCURACY

- With the above-mentioned issue in mind, researchers have proposed two heuristic measures
 - Error estimation using **Loss Functions**
 - Statistical Estimation using **Confidence Level**
- In the next few slides, we will discuss about the two estimations

Error Estimation using Loss Functions

- Let T be a matrix comprising with N test tuples

$$\begin{bmatrix} X_1 & y_1 \\ X_2 & y_2 \\ \vdots & \vdots \\ X_N & y_N \end{bmatrix}_{N \times (n+1)}$$

where X_i ($i = 1, 2, \dots, N$) is the n -dimensional test tuples with associated outcome y_i .

- Suppose, corresponding to (X_i, y_i) , classifier produces the result (X_i, y'_i)
- Also, assume that $(y_i - y'_i)$ denotes a difference between y_i and y'_i (following certain difference (or similarity), (e.g., $(y_i - y'_i) = 0$, if there is a match else 1)
- The two loss functions measure the error between y_i (the actual value) and y'_i (the predicted value) are

Absolute error: $|y_i - y'_i|$

Squared error: $|y_i - y'_i|^2$

Error Estimation using Loss Functions

- Based on the two loss functions, the test error (rate) also called **generalization error**, is defined as the average loss over the test set T. The following two measures for test errors are

Mean Absolute Error (MAE):
$$\frac{\sum_{i=1}^N |y_i - y_i'|}{N}$$

Mean Squared Error (MSE):
$$\frac{\sum_{i=1}^N (y_i - y_i')^2}{N}$$

- Note that, MSE aggregates the presence of outlier.
- In addition to the above, a relative error measurement is also known. In this measure, the error is measured relative to the mean value \tilde{y} calculated as the mean of y_i ($i = 1, 2, \dots, N$) of the training data say D. Two measures are

Relative Absolute Error (RAE):
$$\frac{\sum_{i=1}^N |y_i - y_i'|}{\sum_{i=1}^N |y_i - \tilde{y}|}$$

Relative Squared Error (RSE):
$$\frac{\sum_{i=1}^N (y_i - y_i')^2}{\sum_{i=1}^N (y_i - \tilde{y})^2}$$

Statistical Estimation using Confidence Level

- In fact, if we know the value of predictive accuracy, say ϵ , then we can guess the true accuracy within a certain range given a **confidence level**.
- **Confidence level**: The concept of “confidence level ” can be better understood with the following two experiments, related to tossing a coin.
- **Experiment 1**: When a coin is tossed, there is a probability that the head will occur. We have to experiment the value for this probability value. A simple experiment is that the coin is tossed many times and both numbers of heads and tails are recorded.

N=10		N=50		N=100		N=250		N=500		N=1000	
H	T	H	T	H	T	H	T	H	T	H	T
3	7	29	21	54	46	135	115	241	259	490	510
0.30	0.70	0.58	0.42	0.54	0.46	0.54	0.46	0.48	0.42	0.49	0.51

- Thus, we can say that $p \rightarrow 0.5$ after a large number of trials in each experiment.

Statistical Estimation using Confidence Level

- **Experiment 2:** A similar experiment but with different counting is conducted to learn the probability that a coin is flipped its head 20 times out of 50 trials. This experiment is popularly known as Bernoulli's trials. It can be stated as follows.

$$P(X = v) = \binom{N}{v} p^v (1 - p)^{N-v}$$

- where N = Number of trials
- v = Number of outcomes that an event occurs.
- p = Probability that the event occur
- Thus, if $p = 0.5$, then $P(X = 20) = \binom{50}{20} 0.5^{20} \times 0.5^{30} = 0.0419$
- **Note:**
 - Also, we may note the following
 - Mean = $N \times p = 50 \times 0.5 = 25$ and Variance = $p \times (1-p) \times N = 50 \times 0.5 \times 0.5 = 12.5$

Statistical Estimation using Confidence Level

- The task of predicting the class labels of test records can also be considered as a binomial experiment, which can be understood as follows. Let us consider the following.
 - N = Number of records in the test set.
 - n = Number of records predicted correctly by the classifier.
 - $\epsilon = n/N$, the observed accuracy (it is also called the empirical accuracy).
 - $\tilde{\epsilon}$ = the true accuracy.
- Let τ_{α}^L and τ_{α}^U denotes the lower and upper bound of a confidence level α . Then the confidence interval for α is given by

$$P\left(\tau_{\alpha}^L \leq \frac{\epsilon - \tilde{\epsilon}}{\sqrt{\epsilon(1-\epsilon)/N}} \leq \tau_{\alpha}^U\right) = \alpha$$

- If τ_{α} is the mean of τ_{α}^L and τ_{α}^U , then we can write

$$\tilde{\epsilon} = \epsilon \pm \tau_{\alpha} \times \sqrt{\epsilon(1-\epsilon)/N}$$

Statistical Estimation using Confidence Level

$$\tilde{\epsilon} = \epsilon \pm \tau_{\alpha} \times \sqrt{\epsilon (1 - \epsilon) / N}$$

- A table of τ_{α} with different values of α can be obtained from any book on statistics. A small part of the same is given below.

	0.5	0.7	0.8	0.9	0.95	0.98	0.99
	0.67	1.04	1.28	1.65	1.96	2.33	2.58

- Thus, given a confidence level α , we shall be able to know the value of τ_{α} and hence the true accuracy ($\tilde{\epsilon}$), if we have the value of the observed accuracy (ϵ).
- Thus, knowing a test data set of size N , it is possible to estimate the true accuracy!

Statistical Estimation using Confidence Level

Example 21.2: True accuracy from observed accuracy

A classifier is tested with a test set of size 100. Classifier predicts 80 test tuples correctly. We are to calculate the following.

- a) Observed accuracy
- b) Mean error rate
- c) Standard error
- d) True accuracy with confidence level 0.95.

Solution:

- a) The observed accuracy(ϵ) = $80/100 = 0.80$ So error (p) = 0.2
- b) Mean error rate = $p \times N = 0.2 \times 100 = 20$
- c) Standard error rate (σ) = $\sqrt{\epsilon(1-\epsilon)/N} = \sqrt{\frac{0.8 \times 0.2}{100}} = 0.04$
- d) $\tilde{\epsilon} = \epsilon \pm \tau_{\alpha} \times \sqrt{\epsilon(1-\epsilon)/N} = 0.8 \pm 0.04 \times 1.96 = 0.7216$ with $\tau_{\alpha} = 1.96$ and $\alpha = 0.95$.

Statistical Estimation using Confidence Level

Note:

- Suppose, a classifier is tested k times with k different test sets. If ϵ_i denotes the predicted accuracy when tested with test set N_i in the i -th run ($1 \leq i \leq k$), then the overall predicted accuracy is

$$\epsilon = \sum_{i=1}^k \frac{\epsilon_i \times N_i}{\sum N_i}$$

Thus, ϵ is the weighted average of ϵ_i values. The standard error and true accuracy at a confidence α are

$$\text{Standard error} = \sqrt{\epsilon(1-\epsilon) / \sum_{i=1}^k N_i}$$

$$\text{True accuracy} = \epsilon \pm \sqrt{\frac{\epsilon(1-\epsilon)}{\sum_{i=1}^k N_i}} \times \tau_\alpha$$

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?