BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Introduction to Machine learning

Course Code: 23ST01VAC

Unit-II

Resampling Methods

Dr. T. Jai Sankar	Ms. E. Devi
Associate Professor and Head	Guest Faculty
Department of Statistics	Department of Statistics

UNIT II

RESAMPLING METHODS:

Resampling method is a tool consisting in repeatedly drawing samples from a dataset and calculating statistics and matrices on each of those samples in order to obtain further information about performance of a model in the machine learning.

Two common methods of Resampling are:

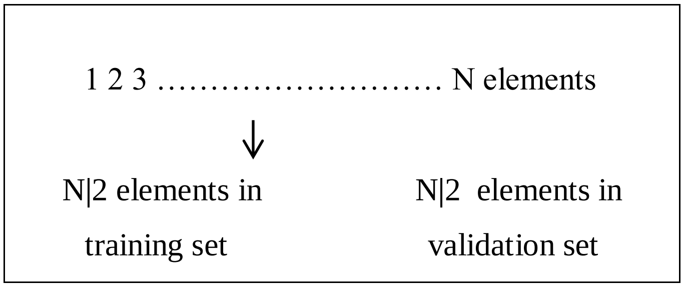
- CROSS VALIDATION
- BOOT STRAPPING

CROSS VALIDATION:

Cross Validation is used to estimate the test error associated with a model to evaluate its performance.

<u>Definition</u> – Cross Validation is a statistical method of evaluating and comparing learning algorithm by dividing dataset into two segments: One used to learn or train model and the other used to validate the model.

<u>Validation set approach</u> – This is the most basic approach. It simply involves randomly dividing the dataset into two parts. First a training set and second a validation set or hold-out set. The model is fit on the training set and the fitted model is used to make predictions on the validation set.



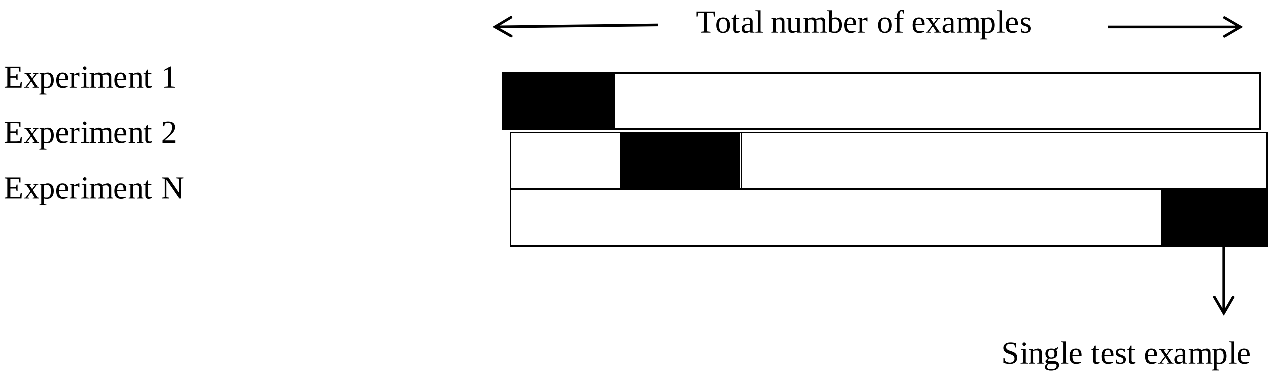
1 2 3 ……………………… N elements

N|2 elements in training set

N|2 elements in validation set

<u>LEAVE – ONE – OUT – CROSS – VALIDATION –</u>

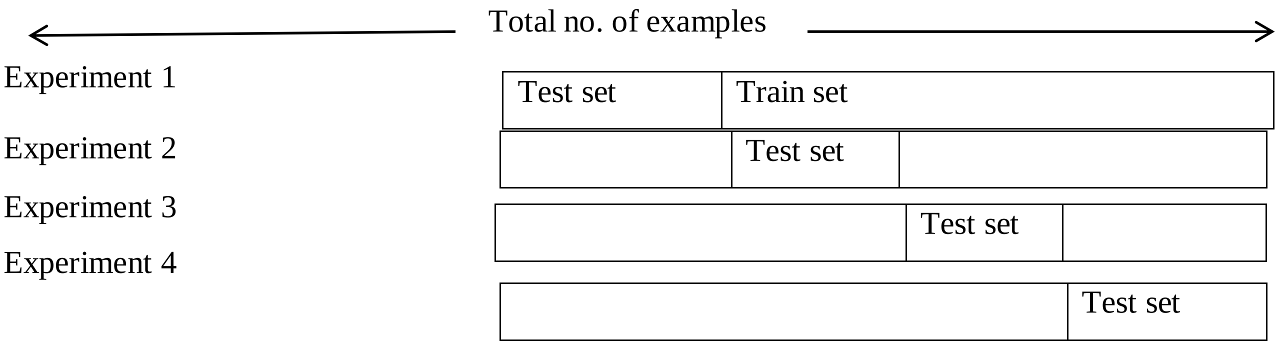
Leave – one – out – cross – validation is a better option than the validation set approach. Instead of splitting the entire dataset into two halves only one observation is used for validation and the rest is used to fit the model.



K – FOLD CROSS – VALIDATION –

This approach involves randomly dividing the set of observations into K folds of nearly equal size. The first fold is treated as a validation set and the model is fit on the remaining folds. The procedure is then repeated up took times, where a different group each time is treated as the validation set.

k folds cross validation approach



BOOT STRAPPING:

a. Boot Strapping is a resampling technique that involves repeated drawing samples from our source data with replacement. It is used to estimate a population parameter.

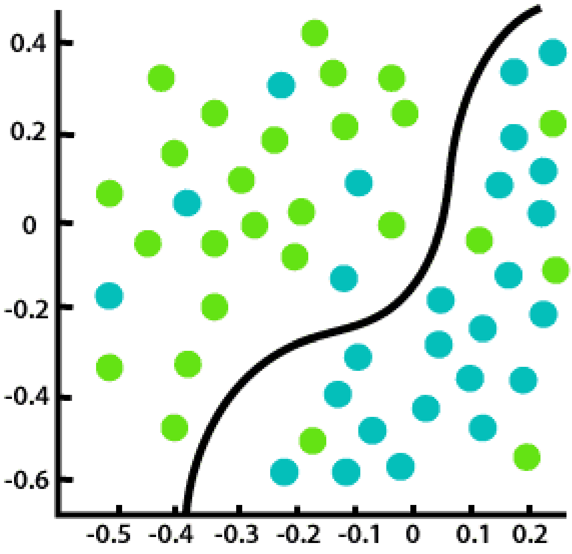
b. Boot Strapping is based on the law of large numbers, which says that with enough data the empirical distribution will be good approximation of the true distributions.

c. Using boot strapping, we can generate a distribution of estimate rather than single point estimation. The distribution gives us information about certainty or the lack of it.

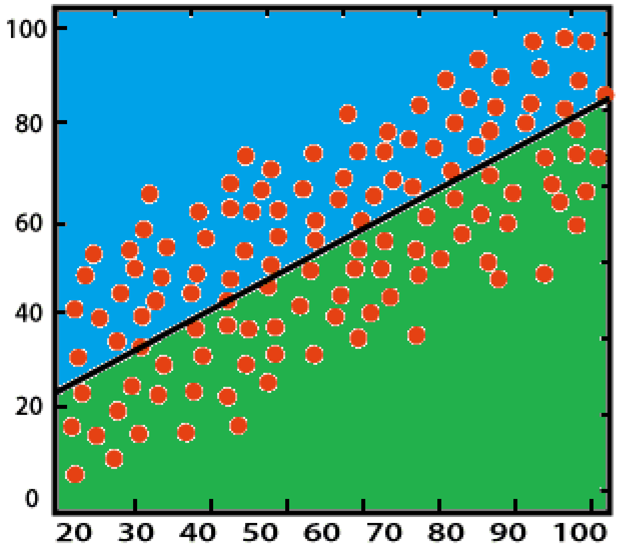
d. Boot Strapping allows us to estimate uncertainty, allowing computation of confidence intervals.

CLASSIFICATION & REGRESSION (ALGORITHM):

Regression and Classification algorithms are supervised learning algorithms. Both the algorithms are used for prediction in Machine Learning and work with the labeled datasets. The main difference between Regression and Classification algorithms is that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and classification algorithms are used to predict or classify the discrete values such as male or female, true or false, spam or not spam, etc.



Classification



Regression

CLASSIFICATION

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In classification, a computer program is trained on the training dataset and based on that training, it categories the data into different classes.

The task of the classification algorithm is to find the mapping function to map the input (x) to the discrete output (y).

Example: The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

Classification Algorithm can be further divided into the following parts:

- LOGISTIC REGRESSION
- K – NEAREST NEIGHBOURS
- SUPPORT VECTOR MACHINES
- KERNEL SVM
- NAÏVE BAYES
- DECISION TREE CLASSIFICATION
- RANDOM FOREST CLASSIFICATION

REGRESSION

Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of Market Trends, prediction of House prices, etc.

The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

Types of Regression Algorithm:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

Difference between Regression and Classification:

REGRESSION ALGORITHM	CLASSIFICATION ALGORITHM
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.

NAÏVE BAYES:

The Bayesian classifier is an algorithm for classifying multiclass datasets. This is based on the Baye’s theorem in probability theory.

Assumption:

The naïve bayes algorithm is based on the following assumptions:

1. All the features are independent and are unrelated to each other. Presence or absences of a feature does not influence the presence or absence of any other feature.

2. The data has class – conditional independence, which means the events are independent as long as they are conditioned on the same class value.

These assumptions are true in many real world problems. It is because of these assumptions; the algorithm is called a naïve algorithm.

Algorithm – let f_1 denote an arbitrary value of $F_{1,}$ f_2 of F_2 and so on. Let the set of class labels be $\{c_1, c_2 \dots c_p\}$. Let there be given a test instance having feature vector

$$X = (x_1, x_2, \dots., x_n)$$

We required to determine the most appropriate class label that should be assigned to the test instance.

STEP 1: Compute the probabilities $P(c_k)$ for $k = 1, 2, \dots p$.

STEP 2: From a table showing the conditional probabilities

$$P(f_1|c_k), P(f_2|c_k), \dots P(f_n|c_k)$$

for all values of $f_1, f_2, \dots f_n$ and for $k = 1, 2, \dots p$.

STEP 3: Compute the products

$$q_k = P(x_1|c_k) P(x_2|c_k) \dots. P(x_n|c_k) P(c_k) \quad ; \text{for } k = 1, 2, \dots p.$$

STEP 4: find j such $q_j = \max(q_1, q_2 \dots. q_p)$

STEP 5: assign the class label c_j to the test instance x.

PROBLEM:

Consider a training dataset consisting of the FAUNA of the world. Each unit has three features named “swim”, “fly” and “crawl”. Let the possible values of these features be as follows;

Swim	Fast, slow, no
Fly	long, short, rarely, no
Crawl	yes, no

For simplicity, each unit is classified as “Animal”, “bird” or “fish”. Let the training dataset be as shown in table. Use naïve Bayes algorithm to classify a particular species if its features are (slow, rarely, no)?

S.NO	SWIM	FLY	CRAWL	CLASS
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish
11	No	Long	No	Bird
12	Fast	No	No	Bird

Table: sample dataset for NAÏVE BAYES algorithm.

Solution:

In this example, the given features are;

F_1 = "swim", F_2 = "fly", F_3 = "crawl"

The class labels are; c_1 = "Animal", c_2 = "Bird", c_3 = "Fish"

The test instance is (slow, rarely, no) and so we have;

x_1 = "slow", x_2 = "rarely", x_3 = "no"

We construct the frequency;

Class	Swim(F_1)			Features Fly(F_2)				Crawl(F_3)		Total
	Fast	Slow	No	Long	Short	Rarely	No	Yes	No	
Animal[c_1]	2	2	1	0	0	1	4	2	3	5
Bird[c_2]	1	0	3	1	2	0	1	1	3	4
Fish[c_3]	1	2	0	0	0	0	3	0	3	3
Total	4	4	4	1	2	1	8	4	8	12

STEP 1: We compute following probabilities;

$$P(c_1) = \frac{\text{No.of records with class label "Animal"}}{\text{Total no.of examples}} = 5/12$$

$$P\ (c_2) = \frac{\text{N0.of records with class label "Bird"}}{\text{Total no.of examples}} = 4/12$$

$$P\ (c_3) = \frac{\text{No.of recoeds with class label "Fish"}}{\text{Total no.of examples}} = 3/12$$

STEP 2: We construct the following table of conditional probabilities;

Class	**Features**								
	Swim[F_1] **f_1**			**Fly[F_2]** **f_2**				**Crawl[F_3]** **f_3**	
	Fast	**Slow**	**No**	**Long**	**Short**	**Rarely**	**No**	**Yes**	**No**
Animal[c_1]	2/5	2/5	1/5	0/5	0/5	1/5	4/5	2/5	3/5
Bird[c_2]	1/4	0/4	3/4	1/4	2/4	0/4	1/4	0/4	4/4
Fish[c_3]	1/3	2/3	0/3	0/3	0/3	0/3	3/3	0/3	3/3

The conditional probabilities are calculated as follows;

$$P\ ((F_1 = \text{slow})/c_1) = \frac{\text{No.of records with F1=slow and class label c1}}{\text{No.of records with class label c1}} = 2/5$$

STEP 3: We now calculate the following numbers;

$q_1 = P\ (x_1|c_1)\ P\ (x_2|c_1)\ P\ (x_3|c_1)\ P\ (c_1) = (2/5)\ (1/5)\ (3/5)\ (5/12) = 0.02$

$q_2 = P\ (x_1|c_2)\ P\ (x_2|c_2)\ P\ (x_3|c_2)\ P\ (c_2) = (0/4)\ (0/4)\ (3/4)\ (4/12) = 0$

$q_3 = P\ (x_1|c_3)\ P\ (x_2|c_3)\ P\ (x_3|c_3)\ P\ (c_3) = (2/3)\ (0/3)\ (3/3)\ (3/12) = 0$

STEP 4: now; max $\{q1, q_2, q_3\} = 0.05$

STEP 5: the maximum is $q_1$and it corresponds to the class label; c_1 = “Animal”.

So we assign the class label “Animal” to the test instance “(slow, rarely, no)”.

ZERO FREQUENCY PROBLEMS:

If an instance in the test dataset has a category that was not present during training, then it will assign it “zero probability and won’t be able to make a prediction”.

MISSING VALUES AND NUMERIC ATTRIBUTES:

The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms that do not support missing values.

Types of Missing Data:

- Missing completely at random (**MCAR**)
- Missing at random (**MAR**)
- Not missing at random (**NMAR**)

MCAR – These are data that are missing completely at random. The missingness is important from the data. There is visible pattern to this type of data missingness.

MAR – These types of data are missing at random but not completely missing. The missingness was caused by the observed data.

NMAR – These are the data that are not missing at random and are also known as ignorable data. The missingness of the missing data is determined by the variable of interest.

MULTIPLE REGRESSIONS:

Let there be more than one independent variable, say x_1, x_2, … $x_{n,}$ and let the relation between y and the independent variables be modeled as

$$\mathbf{y = \alpha_0 + \alpha_1 x_1 + \ldots.. + \alpha_n}$$

then it is case of multiple linear regression or multiple regression.

We assume that there are N independent variables x_1, x_2, … $x_{n.}$ Let the dependent variable be y. let there also be n observed values of these variables:

Variables (features)	Values (examples)			
	Example 1	Example 2	………	Example n
x_1	x_{11}	x_{12}	………	x_{1n}
x_2	x_{21}	x_{22}	……….	x_{2n}
….	….	….	……….	….
x_N	x_{N1}	x_{N2}	……….	x_{Nn}
y(outcome)	y_1	y_2	……….	y_n

The multiple linear regression model defines the relationship between the N independent variables and the dependent variable by an equation of the following form:

$$y = \beta_0 + \beta_1 x_1 + \ldots. + \beta_N x_N$$

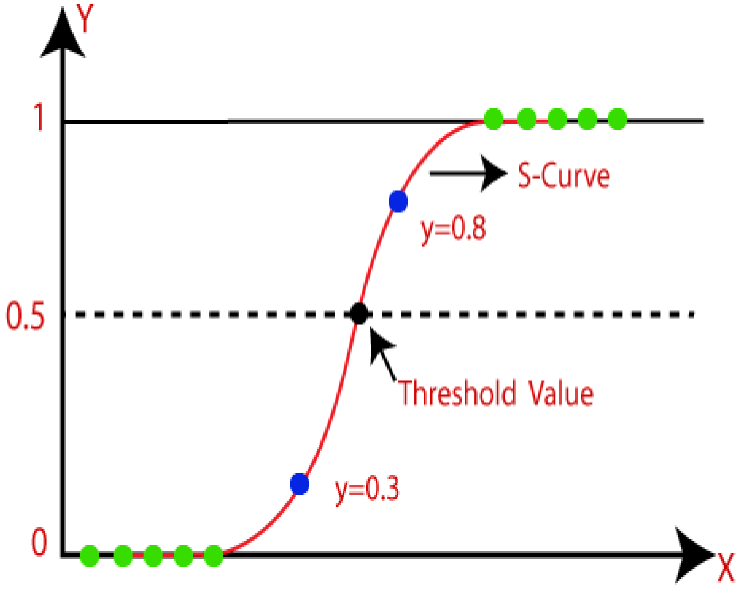
then regression coefficients are given by

$$\mathbf{B = (xTx)^{-1} xTy}$$

LOGISTIC REGRESSION:

Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature. Even though the output is a binary variable, probability function which may take any value from 0 to 1.

Logistic regression is used for solving the classification problem. Logistic regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.



LOGISTIC FUNCTION OR SIGMOID FUNCTION:

- The sigmoid function is a mathematical function used to map the predicted values to probabilities
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which can't go beyond this limit, so it forms a curve like the "s" form. The s-form curve s called sigmoid function or the logistic function.

Types of logistic regression –

Logistic regression can be classified into 3 types:

1. **Binomial –** In binomial logistic regression, there can be only two possible types of the dependent variables such as 0 or 1, pass or fail etc.
2. **Multinomial –** In multinomial logistic regression, there can be 3 or more possible unordered types of the dependent variable such as “cat”, “dogs” or “sheep”.
3. **Ordinal –** In ordinal logistic regression, there can be 3 or more possible ordered types of dependent variables such as “low”, “medium” or “high”.

K – NEAREST – NEIGHBOURS:

- K-nearest neighbors is one of the simplest machine learning algorithms based on supervised learning technique.
- K-NN algorithm can be used for regression as well as for classification.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

ADVANTAGES OF K-NN ALGORITHM:

- It is simple to implement and easy to understand.
- It can be more effective if the training data is large.
- It can be used for both classification and regression problems.
- It can naturally handle multi-class cases.
- It can perform well with enough representative data.

DISADVANTAGES OF K-NN:

- Associated computation cost is high as it stores all the training data.
- Requires high memory storage.
- Need to determine the value of k.
- Prediction is below if the value of N is high.

APPLICATION OF K-NN:

- Banking system
- Calculating credit rating
- Politics

TYPES OF NAÏVE BAYES ALGORITHM:

1. **Gaussian Naïve Bayes** – when characteristics values are continuous in nature then an assumption is made that the values linked with each class are dispersed according to Gaussian that is normal distribution.
2. **Multinomial Naïve Bayes** – multinomial naïve bayes is favored to use on data that is multinomial distributed. It is widely used in text classification in NLP. Each event in text classification constitutes the presence of a word in a document.
3. **Bernoulli Naïve Bayes –** when data is dispersed according to the multivariate Bernoulli distribution then Bernoulli naïve bayes is used. That means there exist multiple features but each one is assumed to contain a binary value. So it requires features to be binary-valued.

ADVANTAGES OF NAÏVE BAYES CLASSIFICATION:

- It is a highly extensive algorithm.
- It can be used for both binaries as well as multiclass classification.
- It can be easily trained on small datasets and can be used for large volumes of data as well.

DISADVANTAGES OF NAÏVE BAYES CLASSIFICATION:

The main disadvantage of it is considering all the variables independent that contributes to the probability.

ADVANTAGES OF BOOT STRAPPING:

- Avoids the cost of taking new samples.
- Checking parametric assumptions.
- Used when parametric assumption cannot be made (or) are very complicated.
- Estimation of variance in quantiles.

DISADVANTAGES OF BOOT STRAPPING:

- Relies on a representative sample.
- Variability due to finite replications.