Unit- Design and Analysis of Algorithms:

- Study of factors and responses related with experimentation,
- Hypothesis testing,
- performance analysis,
- Evaluation measures-bootstrapping & cross-validation, ROC curve.

EXPERIMENT

The term experiment is defined as the systematic procedure carried out under controlled conditions in order to discover an unknown effect, to test or establish a hypothesis, or to illustrate a known effect.

When analyzing a process, experiments are often used to evaluate which process inputs have a significant impact on the process output, and what the target level of those inputs should be to achieve a desired result (output).

Experiments can be designed in many different ways to collect this information. Design of Experiments (DOE) is also referred to as Designed Experiments or Experimental Design - all of the terms have the same meaning.

Experimental Design

We are concerned with the analysis of data generated from an experiment. It is wise to take time and effort to organize the experiment properly to ensure that the right type of data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible. This process is called *experimental design*.

The specific questions that the experiment is intended to answer must be clearly identified before carrying out the experiment. We should also attempt to identify known or expected sources of variability in the experimental units since one of the main aims of a designed experiment is to reduce the effect of these sources of variability on the answers to questions of interest. That is, we design the experiment in order to improve the precision of our answers.

Basic Concepts of experimentation/ Components

Usually the goal of a study is to find out the relationships between certain explanatory factors and the response variables. The design of a study thus consists of making decisions on the following:

- The set of explanatory factors.
- The set of response variables.
- The set of treatments.
- The set of experimental units.
- The method of randomization and blocking.

- Sample size and number of replications.
- The outcome measurements on the experimental units - the response variables.

Factors

Factors are explanatory variables to be studied in an investigation. Factors are inputs to the process. Factors can be classified as either controllable or uncontrollable variables

Examples:

1. In a study of the effects of colors and prices on sales of cars, the factors being studied are color (qualitative variable) and price (quantitative variable).

2. In an investigation of the effects of education on income, the factor being studied is education level (qualitative but ordinal).

Factor levels

Factor levels are the "values" of that factor in an experiment. For example, in the study involving color of cars, the factor car color could have four levels: red, black, blue and grey. In a design involving vaccination, the treatment could have two levels: vaccine and placebo.

Types of factors

- Experimental factors: levels of the factor are assigned at random to the experimental units.
- Observational factors: levels of the factor are characteristic of the experimental units and is not under the control of the investigators.
- There could be observational factors in an experimental study.

Treatments

- In a single factor study, a treatment corresponds to a factor level; thus the number of treatments equals the number of different factor levels of that factor.
- In a multi-factor study, a treatment corresponds to a *combination of factor levels across different factors;* thus the number of all possible treatments is the product of the number of factor levels of different factors.

Examples:

- In the study of effects of education on income, each education level is a treatment (high school, college, advanced degree, etc).
- In the study of effects of race and gender on income, each combination of race and gender is a treatment (Asian female; Hispanic male, etc).

Experimental units

- An experimental unit is the smallest unit of experimental material to which a treatment can be assigned.

Example: In a study of two retirement systems involving the 10 UC schools, we could ask if the basic unit should be an individual employee, a department, or a University.

Answer: The basic unit should be an entire University for practical feasibility.

- Representativeness: the experimental units should be representative of the population about which a conclusion is going to be drawn.

Sample size

Sample size is the number of experimental units in the study.

- In general, the larger the sample size, the better it is for statistical inference; however, the costlier is the study.

Response

Response is the output of the experiment. In the case of cake baking, the taste, consistency, and appearance of the cake are measurable outcomes potentially influenced by the factors and their respective levels. Experimenters often desire to avoid optimizing the process for one response at the expense of another. For this reason, important outcomes are measured and analyzed to determine the factors and their settings that will provide the best overall outcome for the critical-to-quality characteristics - both measurable variables and assessable attributes.

HYPOTHESIS TESTING

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter. The process of hypothesis testing is to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample

Hypothesis testing is an essential procedure in statistics. A **hypothesis test** evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. When **we** say that a finding is statistically significant, it's thanks to a **hypothesis test**.

Which are important parameter of hypothesis testing ?

Null hypothesis :- In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups

In other words it is a basic assumption or made based on domain or problem knowledge.

Example : a company production is = 50 unit/per day etc.

Alternative hypothesis :-

The alternative hypothesis is the hypothesis used in **hypothesis** testing that is contrary to the null hypothesis. It is usually taken to be that the observations are the result of a real effect (with some amount of chance variation superposed)

Example : a company production is !=50 unit/per day etc.

Level of significance: Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.

This is normally denoted with alpha(maths symbol) and generally it is 0.05 or 5% , which means your output should be 95% confident to give similar kind of result in each sample.

some of widely used hypothesis testing type :-

1. T Test (Student T test)
2. Z Test
3. ANOVA Test
4. Chi-Square Test

PERFORMANCE/ EVALUATION METRICS

Evaluating machine learning algorithm is an essential part of any project.

Classification Accuracy

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class.

Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

Lets assume we have a binary classification problem. We have some samples belonging to two classes : YES or NO. Also, we have our own classifier which predicts a class for a given input sample. On testing our model on 165 samples ,we get the following result.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

There are 4 important terms :

- **True Positives** : The cases in which we predicted YES and the actual output was also YES.
- **True Negatives** : The cases in which we predicted NO and the actual output was NO.
- **False Positives** : The cases in which we predicted YES and the actual output was NO.
- **False Negatives** : The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the **"main diagonal"** i.e

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$$

Area Under Curve (ROC)

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. *AUC* of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Before defining *AUC*, let us understand two basic terms :

- **True Positive Rate (Sensitivity)** : True Positive Rate is defined as *TP/ (FN+TP)*. True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

- **True Negative Rate (Specificity)** : True Negative Rate is defined as *TN / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are correctly considered as negative, with respect to all negative data points.

$$TrueNegativeRate = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

- **False Positive Rate** : False Positive Rate is defined as *FP / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$FalsePositiveRate = \frac{FalsePositive}{TrueNegative + FalsePositive}$$

F1 Score

F1 Score is used to measure a test's accuracy

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as :

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

F1 Score tries to find the balance between precision and recall.

- **Precision :** It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

- **Recall :** It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Mean Absolute Error

Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don't gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Mathematically, it is represented as :

$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j|$$

Mean Squared Error

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced than smaller error, hence the model can now focus more on the larger errors.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

BOOTSTRAP METHOD

The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples.

Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

The process for building one sample can be summarized as follows:

1. Choose the size of the sample.
2. While the size of the sample is less than the chosen size
    1. Randomly select an observation from the dataset
    2. Add it to the sample

The bootstrap method can be used to estimate a quantity of a population. This is done by repeatedly taking small samples, calculating the statistic, and taking the average of the calculated statistics. We can summarize this procedure as follows:

1. Choose a number of bootstrap samples to perform
2. Choose a sample size
3. For each bootstrap sample
    1. Draw a sample with replacement with the chosen size

2. Calculate the statistic on the sample
4. Calculate the mean of the calculated sample statistics.

The procedure can also be used to estimate the skill of a machine learning model.

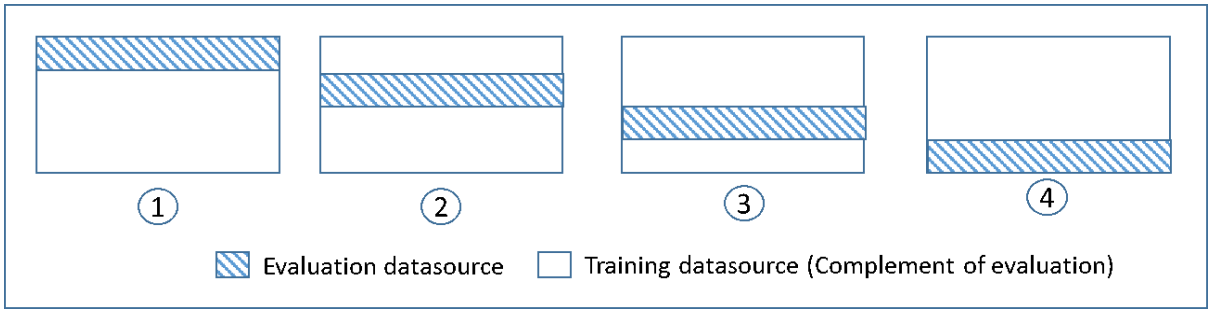
The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

CROSS-VALIDATION

Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern.

Can use the k-fold cross-validation method to perform cross-validation. In k-fold cross-validation, you split the input data into k subsets of data (also known as folds). You train an ML model on all but one (k-1) of the subsets, and then evaluate the model on the subset that was not used for training. This process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time.

The following diagram shows an example of the training subsets and complementary evaluation subsets generated for each of the four models that are created and trained during a 4-fold cross-validation. Model one uses the first 25 percent of data for evaluation, and the remaining 75 percent for training. Model two uses the second subset of 25 percent (25 percent to 50 percent) for evaluation, and the remaining three subsets of the data for training, and so on.



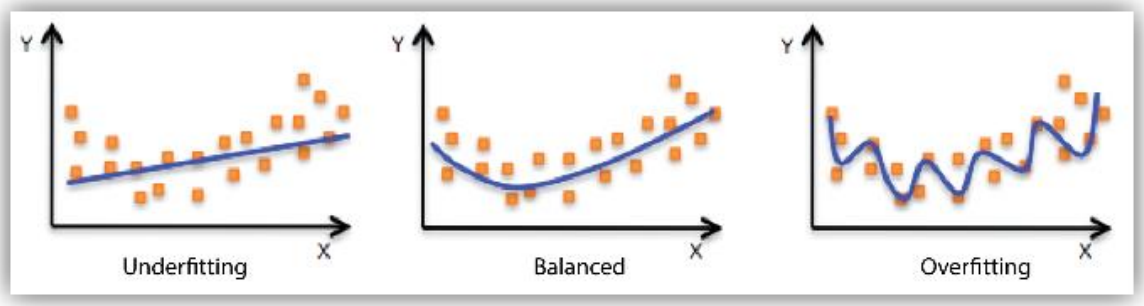
Each model is trained and evaluated using complementary datasources - the data in the evaluation datasource includes and is limited to all of the data that is not in the training datasource.

Performing a 4-fold cross-validation generates four models, four data sources to train the models, four data sources to evaluate the models, and four evaluations, one for each model.

ML generates a model performance metric for each evaluation. For example, in a 4-fold cross-validation for a binary classification problem, each of the evaluations reports an area under curve (AUC) metric. You can get the overall performance measure by computing the average of the four AUC metrics.

Model Fit: Underfitting vs. Overfitting

Understanding model fit is important for understanding the root cause for poor model accuracy. This understanding guides to take corrective steps. We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data.



The model is *underfitting* the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples (often called X) and the target values (often called Y). The model is *overfitting* your training data when you see that the model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.