Course Summary

Machine Learning is a first-class ticket to the most exciting careers in data analysis today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine learning brings together computer science and statistics to harness that predictive power. It’s a must-have skill for all aspiring data analysts and data scientists, or anyone else who wants to wrestle all that raw data into refined trends and predictions.

This is a class that will teach you the end-to-end process of investigating data through a machine learning lens. It will teach you how to extract and identify useful features that best represent your data, a few of the most important machine learning algorithms, and how to evaluate the performance of your machine learning algorithms.

Why Take This Course?

In this course, you’ll learn by doing! We’ll bring machine learning to life by showing you fascinating use cases and tackling interesting real-world problems like self-driving cars. For your final project you’ll mine the email inboxes and financial data of Enron to identify persons of interest in one of the greatest corporate fraud cases in American history.

Syllabus

Supervised

Linear Regression:

We will learn this approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X.

Logistic Regression:

We will study how binary classifier are handled using

Email Spamming.

Decision Trees:

In this course we will learn how Decision trees uses a**tree-**like graph or model of **decisions** and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

Random Forest

Unsupervised Learning

K-Mean Clustering

Feature Creation: Taking your human intuition about the world and turning it into data that a computer can use.

Feature Selection: Einstein said it best: make everything as simple as possible, and no simpler. In this case, that means identifying the most important features of your data.

Principal Component Analysis: A more sophisticated take on feature selection, and one of the crown jewels of unsupervised learning.

Feature Scaling: Simple tricks for making sure your data and your algorithm play nicely together.
Learning from Text: More information is in text than any other format, and there are some effective but simple tools for extracting that information.

Lessons 13-14: Validation and Evaluation

Training/testing data split: How do you know that what you're doing is working? You don't, unless you validate. The train-test split is simple to do, and the gold standard for understanding your results.

Cross-validation: Take the training/testing split and put it on steroids. Validate your machine learning results like a pro.

Assignments to be covered:

Email Spam

Fuel Efficiency

Twitter Sentiment analysis.

Project:

Churn Prediction:

We will work on Live data set and build a model using logistic regression.

STATISTICS COURSE FOR DATA SCIENCE

Topic: 1

Describe and explore data:

Mean,median,mode,SD,Variance,corelation,covariance,quartile
Displaying Data, Bar Chart, Contingency Table, Boxplot, Histogram

Topic: 2

Probability

Random varaibles:roll a dice example.Normal distribution and outliers.

Topic: 3

Basic concepts of probability.

Learning Objectives

Compute probability in a situation where there are equally-likely outcomes.
Apply concepts to cards and dice.
Compute the probability of two independent events both occurring.
Compute the probability of either of two independent events occurring.
Do problems that involve conditional probabilities

Topic 4:

Hypothesis:

Confidence intervals,degree of freedom,P-T-F test, Hypothesis Testing

Topic 5:
Correlation vs. Regression

This chapter will speak of both correlations and regressions both use similar mathematical procedures to provide a measure of relation;
The degree to which two continuous variables vary together ... or covary.The correlations term is used when 1) both variables are random variables, and 2) the end goal is simply to find a number that expresses the relation between the variables.The regression term is used when 1) one of the variables is a fixed variable, and 2) the end goal is use the measure of relation to predict values of the random variable based on values of the fixed variable

Topic 6:
Anova

This chapter will give deep understanding of analysis of variables:

Topic 7:
Time Series:

We will cover forecasting techniques.

Few terminologies we will understand after this course:

- ✓ Histograms and Cumulative Frequency
- ✓ Averages
- ✓ Measures of Dispersion
- ✓ Box and Whisker Diagrams
- ✓ Probability
- ✓ Linear Regression
- ✓ Skewness
- ✓ Discrete Random Variables
- ✓ Expectation and Variance
- ✓ Discrete Uniform Distribution
- ✓ Normal Distribution
- ✓ Binomial Distribution
- ✓ Continuous Random Variables
- ✓ Uniform Distribution
- ✓ Sampling
- ✓ Hypothesis Testing
- ✓ One and Two Tailed Tests

- ✓ Central Limit Theorem
- ✓ Confidence Intervals
- ✓ Permutations and Combinations
- ✓ Central tendancy
- ✓ Dispersion
- ✓ Probability
- ✓ Normal Distribution
- ✓ Students t test
- ✓ F test
- ✓ Correlation
- ✓ Linear Regression
- ✓ Analysis of Variance

After completing this course, the student should be able to:

1. Distinguish between quantitative and categorical data and know which graphical and tabular techniques to apply to each.
2. Produce and interpret graphical displays for simple data sets.
3. Calculate and interpret measures for the center and spread of a data set.
4. Identify how and when to use the Normal model.
5. Identify when correlation and regression analyses are appropriate.
6. Calculate and interpret correlation coefficient and regression line equations.
7. Discuss issues associated with collecting and interpreting data from sample surveys and polls.
8. Discuss the concept of a sampling distribution.
9. Describe what is central limit theorem, understand its relevance to statistical inference.
10. Calculate and interpret confidence intervals for estimating population proportions and means.
11. Formulate null and alternative hypotheses.
12. Use analysis of variance techniques to test the equality of two or more means.
13. Explain the meaning of P-values in hypothesis testing.
14. Identify when and how to use the t-distribution.
15. Determine appropriate sample sizes for estimating an unknown population proportion or mean.

Benifits:

These concepts will be applied in Machine Learning and will be a strong foundation for Machine learning and data analysis.

Assessment

Assignment 1	8%
Assignment 2	6%
Assignment 3	6%
Assignment 4	6%
Assignment 5	8%
Assignment 6	6%
Final Exam *	60%
Total	100%

* Mandatory