



Data Science, EICT Course Curriculum

COURSE DESIGN

High-quality videos, slides, hands-on examples, quizzes, automated assessments, case studies, and real-world projects.

COURSE MATERIAL

Lifetime access to cutting-edge self-paced learning content.

LAB

330 days Lab

SUPPORT

Email support to answer your queries and we've also launched Discussions - a Q&A site for Artificial Intelligence, Machine Learning, Deep Learning, Big Data & Data Science professionals.

CERTIFICATE

Earn a certificate in Data Science Specialization from E&ICT, IIT Roorkee

SELF PACED LEARNING

Lifetime access to the recordings of a previous batch of instructor-led training, assessments, quizzes, and projects



Data Science Specialization

Course Syllabus Python for Machine Learning

Introduction to Linux

Introduction to Python

Hands-on Using Jupyter on Cloudxlab

Overview of Linear Algebra

Introduction to Numpy & Pandas Project - Analyze Your Mailbox

Big Data Introduction

- What Is Big Data?
- Why Now?
- Big Data Use Cases
- Various Solutions
- Overview of Hadoop Ecosystem
- Spark Ecosystem Walkthrough
- Quiz

Foundation & Environment

- Understanding the Cloudxlab
- Cloudxlab Hands-on
- Hadoop & Spark Hands-on
- Quiz and Assessment
- Basics of Linux - Quick Hands-on
- Understanding Regular Expressions

- Quiz and Assessment
- Setting up Vm (optional)

Data Formats & Management

- Inputformat and Inputsplit
- Json
- Xml
- Avro
- How to Store Many Small Files - Sequencefile?
- Parquet
- Protocol Buffers
- Comparing Compressions

Scala Basics

- Introduction to Scala?
- Accessing Scala Using Cloudxlab
- Getting Started: Interactive, Compilation, Sbt
- Types, Variables & Values
- Functions
- Collections
- Classes
- Parameters
- More Features
- Quiz and Assessment

Spark Basics

- What Is Apache Spark?
- Why Spark?
- Using the Spark Shell on Cloudxlab

- Example 1 - Performing Word Count
- Understanding Spark Cluster Modes on Yarn
- Rdds (resilient Distributed Datasets)
- General Rdd Operations: Transformations & Actions
- Rdd Lineage
- Rdd Persistence Overview
- Distributed Persistence

Writing and Deploying Spark Applications

- Creating the Sparkcontext
- Building a Spark Application (scala, Java, Python)
- the Spark Application Web Ui
- Configuring Spark Properties
- Running Spark on Cluster
- Rdd Partitions
- Executing Parallel Operations
- Stages and Tasks
- Project: Churning the Logs of Nasa Kennedy Space Center Www Server

Common Patterns in Spark Data Processing

- Common Spark Use Cases
- Example 1 - Data Cleaning (movielens)
- Example 2 - Understanding Spark Streaming
- Understanding Kafka
- Example 3 - Spark Streaming from Kafka
- Iterative Algorithms in Spark
- Project: Real-time Analytics of Orders in an E-commerce Company

Dataframes and Spark Sql

- Spark Sql and the Sql Context
- Creating Dataframes
- Transforming and Querying Dataframes
- Saving Dataframes
- Dataframes and Rdds
- Comparing Spark Sql, Impala, and Hive-on-spark

Machine Learning With Spark

- Graphx: Graph Processing and Analysis
- Understanding Machine Learning
- Mllib Example: K-means
- Sparkr Example

Introduction to Statistics

- Statistical Inference
- Types of Variables
- Probability Distribution
- Normality
- Measures of Central Tendencies
- Normal Distribution

Machine Learning Applications & Landscape

- Introduction to Machine Learning,
- Machine Learning Application

Introduction to Ai

- Different Types of Machine Learning - Supervised, Unsupervised

- Reinforcement

Building End-to-end Machine Learning Project

- Machine Learning Projects Checklist,
- Frame the Problem and Look at the Big Picture
- Get the Data
- Explore the Data to Gain Insights
- Prepare the Data for Machine Learning Algorithms
- Explore Many Different Models and Short-list the Best Ones
- Fine-tune Model
- Present the Solution
- Launch, Monitor and Maintain the System

Classifications

- Training a Binary Classification,
- Performance Measures
- Confusion Matrix
- Precision and Recall
- Precision/recall Tradeoff
- the Roc Curve
- Multiclass Classification
- Multilabel Classification
- Multi-output Classification

Training Models

- Linear Regression
- Gradient Descent
- Polynomial Regression

- Learning Curves
- Regularized Linear Models
- Logistic Regression

Support Vector Machines

- Linear Svm Classification
- Nonlinear Svm Classification
- Svm Regression

Decision Trees

- Training and Visualizing a Decision Tree
- Making Predictions
- Estimating Class Probabilities
- the Cart Training Algorithm
- Gini Impurity or Entropy
- Regularization Hyperparameters
- Regression
- Instability

Ensemble Learning and Random Forests

- Voting Classifiers
- Bagging and Pasting
- Random Patches and Random Subspaces
- Random Forests
- Boosting
- Stacking

Dimensionality Reduction

- the Curse of Dimensionality

- Main Approaches for Dimensionality Reduction
- Pca
- Kernel Pca
- Lle
- Other Dimensionality Reduction Techniques

Projects

1. Predict the Median Housing Prices in California
2. Classify Handwritten Digits in Mnist Dataset
3. Noise Removal from the Images
4. Predict the Class of Flower in Iris Dataset
5. Predict Which Passengers Survived in the Titanic Shipwreck
6. Predict Bikes Rental Demand
7. Build a Spam Classifier

Introduction to Deep Learning

- Deep Learning Applications,
- Artificial Neural Network,
- Tensorflow Demo,
- Deep Learning Frameworks

Tensorflow

- Installation,
- Creating Your First Graph and Running it in a Session,
- Managing Graphs,
- Lifecycle of a Node Value,
- Linear Regression With Tensorflow,
- Implementing Gradient Descent,

- Feeding Data to the Training Algorithm,
- Saving and Restoring Models,
- Visualizing the Graph and Training Curves Using Tensorboard,
- Name Scopes, Modularity,
- Sharing Variables

Introduction to Artificial Neural Networks

- from Biological to Artificial Neurons,
- Training an Mlp With Tensorflow's High-level Api,
- Training a Dnn Using Plain Tensorflow,
- Fine-tuning Neural Network Hyperparameters

Training Deep Neural Nets

- Vanishing / Exploding Gradients Problems,
- Reusing Pretrained Layers,
- Faster Optimizers,
- Avoiding Overfitting Through Regularization,
- Practical Guidelines

Convolutional Neural Networks

- the Architecture of the Visual Cortex,
- Convolutional Layer,
- Pooling Layer,
- Cnn Architectures

Recurrent Neural Networks

- Recurrent Neurons,
- Basic Rnns in Tensorflow,
- Training Rnns,

- Deep Rnns,
- Lstm Cell,
- Gru Cell,
- Natural Language Processing

Autoencoders

- Efficient Data Representations,
- Performing Pca With an Undercomplete Linear Autoencoder,
- Stacked Autoencoders,
- Unsupervised Pre-training Using Stacked Autoencoders,
- Denoising Autoencoders,
- Sparse Autoencoders,
- Variational Autoencoders

Reinforcement Learning

- Learning to Optimize Rewards,
- Policy Search,
- Introduction to Openai Gym,
- Neural Network Policies,
- Evaluating Actions: the Credit Assignment Problem,
- Policy Gradients,
- Markov Decision Processes,
- Temporal Difference Learning and Q-learning,
- Learning to Play Ms. Pac-man Using Deep Q-learning

Projects

1. Analyze Emails

Churn the mail activity from various individuals in an open source project development team.

2. Predict the median housing prices in California

We start Machine Learning course with this end-to-end project. Learn various data manipulation, visualization and cleaning techniques using various libraries of Python like Pandas, Scikit-Learn and Matplotlib.

3. Classify handwritten digits in MNIST dataset

The MNIST dataset is considered as "Hello World!" of Machine Learning. Write your first classification logic. Starting with Binary Classification learn Multiclass, Multilabel, Multi-output classification and different error analysis techniques.

4. Noise removal from the images

Build a model that takes a noisy image as an input and outputs the clean image.

5. Predict the class of flower in IRIS dataset

IRIS dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The three classes in this dataset are Setosa, Versicolor, and Verginica. Learn Decision Trees, CART algorithm and Ensemble method. Then use Random Forest classifier to make predictions.

6. Predict which passengers survived in the Titanic shipwreck

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. In this project, you build a model to predict which passengers survived the tragedy.

7. Predict bikes rental demand

Build a model to predict the bikes demand given the past data.

8. Build a spam classifier

Build a model to classify email as spam or ham. First, download examples of spam and ham from Apache SpamAssassin's public datasets and then train a model to classify email.

9. Build cats classifier using neural network

In this project, you will build a basic neural network to classify if a given image is of cat or not.

10. Classify large images using Inception v3

Download images of various animals and then download the latest pretrained Inception v3 model. Run the model to classify downloaded images and display the top five predictions for each image, along with the estimated probability.

11. Classify clothes using TensorFlow

Build a model to classify clothes into various categories in Fashion MNIST dataset.

12. Predict the hourly rain gauge total

This is a time series prediction task: you are given snapshots of polarimetric radar values and asked to predict the hourly rain gauge total.

13. Sentiment analysis

Sentiment analysis of "Iron Man 3" movie using Hive and visualizing the sentiment data using BI tools such as Tableau

14. Process the NSE

Process the NSE (National Stock Exchange) data using Hive for various insights

15. MovieLens Project

Analyze MovieLens data using Hive

16. Spark MLlib

Generate movie recommendations using Spark MLlib

17. Spark GraphX

Derive the importance of various handles at Twitter using Spark GraphX

18. Churn the logs

Churn the logs of NASA Kennedy Space Center WWW server using Spark to find out useful business and devops metrics

19. Spark application

Write end-to-end Spark application starting from writing code on your local machine to deploying to the cluster

20. Analytics Dashboard

Real-time analytics dashboard for an e-commerce company using Apache Spark, Kafka, Spark Streaming, Node.js, Socket.IO and Highcharts