2022-23

DIPLOMA IN APPLIED DATA SCIENCE

GURUKUL

EDUCATIONAL AND RESEARCH INSTITUTE, HARIDWAR

विधि:सम्पत्ति:अस्ति

GURUKUL
EDUCATIONAL AND
RESEARCH INSTITUTE

Subject and Syllabus

DIPLOMA IN APPLIED DATA SCIENCE

DURTATION:- 1 YEAR

UNIX

- History and Culture
- The Shell
- Streams
- Standard streams
- Pipes
- Philosophy
- In a nutshell
- More nuts and bolts

670 PROGRAMMING PREREQUISITES

671 VERSION CONTROL WITH GIT

- Background
- What is Git
- Setting Up
- Online Materials
- Basic Git Concepts
- Common Git Workflows
- Linear Move from Working to Remote
- Discarding changes in your working copy
- Erasing changes
- Remotes
- Merge conflicts

672 BUILDING A DATA CLEANING PIPELINE WITH PYTHON

- Simple Shell Scripts
- Template for a Python CLI Utility

673 NOTATION

- Notation for Structured Data

674 LINEAR REGRESSION

- Introduction
- Coefficient Estimation: Bayesian Formulation
- Generic setup
- Ideal Gaussian World
- Coefficient Estimation: Optimization Formulation
- The least squares problem and the singular value decomposition
- Overfitting examples
- L2 regularization
- Choosing the regularization parameter
- Numerical techniques
- Variable Scaling and Transformations
- Simple variable scaling
- Linear transformations of variables
- Nonlinear transformations and segmentation
- Error Metrics

675 LOGISTIC REGRESSION

- Formulation
- Presenter’s viewpoint
- Classical viewpoint
- Data generating viewpoint
- Determining the regression coefficient w
- Multinomial logistic regression
- Logistic regression for classification
- L1 regularization
- Numerical solution
- Gradient descent
- Newton’s method
- Solving the L1 regularized problem
- Common numerical issues
- Model evaluation

676 PROCESSING TEXT

- A Quick Introduction
- Regular Expressions
- Basic Concepts
- Unix Command line and regular expressions
- Finite State Automata and PCRE
- Backreference
- Python RE Module
- The Python NLTK Library
- The NLTK Corpus and Some Fun things to do

677 CLASSIFICATION

- Classification
- A Quick Introduction
- Naive Bayes
- Smoothing
- Measuring Accuracy
- Error metrics and ROC Curves
- Other classifiers
- Decision Trees
- Random Forest
- Out-of-bag classification
- Maximum Entropy

678 HIGHER PERFORMANCE PYTHON

- Memory hierarchy
- Parallelism
- Practical performance in Python
- Profiling
- Standard Python rules of thumb
- For loops versus BLAS
- Multiprocessing Pools
- Multiprocessing example: Stream processing text files
- Numba
- Cython