 **BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**COURSE HANDOUT**

**Part A: Content Design**

|  |  |
| --- | --- |
| **Course Title** | Classification |
| **Course No(s)** |  |
| **Credit Units** |  |
| **Course Author** | N.L.Bhanu Murthy |
| **Version No** |  |
| **Date** | 21/08/2018 |

**Course Description**

|  |
| --- |
| Classification is a type of supervised learning techniques wherein the target attribute takes discrete values. This course emphasizes the three types of techniques to solve classification problems – discriminant function, generative and probabilistic discriminative approaches. This course lays down a strong foundation on algorithmic perspective of popular classification algorithms - k-NN, Naïve Bayes, Decision Tree, Logistic Regression and SVM. The implementation details of these models along with tuning of parameters will be illustrated. The ensemble methods, bagging, boosting, Random Forest and eXtreme Gradient Boosting will be taught. The interpretability/explicability of the models will also be discussed. |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | Objective |
| **CO1** | Provide deeper understanding of three types of techniques to solve classification problems |
| **CO2** | Provide comprehensive algorithmic perspective of popular classification algorithms |
| **CO3** | Provide hands-on to solve real life classification problems |
| **CO4** | Provide the skill to interpret the predicted model |
| **CO5** | Provide the competence to build ensemble classifiers using well known techniques |

**Text Book(s)**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| T1 | Christopher Bishop: Pattern Recognition and Machine Learning, Springer International Edition |
| T2 | Tom M. Mitchell: Machine Learning, The McGraw-Hill Companies, Inc.. |

**Reference Book(s) & other resources**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| R1 | An Introduction to Data Mining – Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar – 2005 |

**Content Structure**

<List down the modular content structure of the course either in the tabular form given below or as bullets>

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Title of the Module** | **Duration in mins** | **Reference** |
|  |  |  |  |
| M1 | Overview of the Classification Module   1. Introduction to Classification 2. Types of classification algorithms - Discriminant Functions, Probabilistic Generative models and Probabilistic Discriminative models, Tree based models 3. Classification Algorithms covered in the course and type of these algorithms 4. Applications of classification and case study of the course | 10  25  15  15 | Class Notes  T1 – Ch. 4 |
| M2 | Nearest-neighbor Methods   1. kNN Classifier 2. Measures of prediction accuracies of classifiers – precision, recall, AUC of ROC etc. 3. Finding optimal k 4. Python Implementation of kNN | 20  20  15  10 | T2 – Ch. 8 |
| M3 | Naïve Bayes Classifier   1. Probability Foundations – Discrete & Continuous Random Variables, Conditional Independence, Bayes Theorem (1) 2. Probability Foundations – Discrete & Continuous Random Variables, Conditional Independence, Bayes Theorem (2) 3. Naïve Bayes Classifier – Derivation 4. An illustrative example 5. Python implementation of Naïve Bayes Classifier 6. Naïve Bayes Classifier is a generative model 7. Advantages of Naïve Bayes Classifier and when to use Naïve Bayes Classifier? 8. Interpretability of Naïve Bayes Classifier | 15  15  25  20  10  10  15  10 | T2 – Ch. 6 |
| M4 | Logistic Regression   1. Significance of Sigmoid function and finding its derivative 2. Statistics Foundations – Maximum likelihood estimation 3. Cross entropy error function for logistic regression and its optimal solution 4. Logistic Regression is probabilistic discriminative model and an illustrative example 5. Implementation of logistic Regression using Python 6. Decision boundary of logistic regression 7. Overfitting of logistic regression and counter measures 8. Interpretability of logistic regression | 15  20  25  20  10  10  15  15 | T1 – 4.3. |
| M5 | Decision Tree   1. Decision Tree Representation 2. Entropy and Information Gain for an attribute 3. Search in Hypothesis space, ID3 Algorithm for decision tree learning 4. Implementation of Decision Tree using Python 5. Prefer short hypothesis to longer ones, Occam’s razor 6. Overfitting in Decision Tree 7. Reduced Error Pruning and Rule post pruning 8. Alternative measures for selecting attributes 9. Interpretability of Decision Tree | 10  15  20  10  15  15  20  15  10 | T2 – Ch. 3 |
| M6 | Optimization Foundations for Support Vector Machines   1. Constrained and Unconstrained Optimization 2. Primal and Dual of an optimization problem 3. Quadratic Programming 4. KKT conditions 5. Lagrange Multiplier | 15  20  15  25  20 | Class Notes |
| M7 | Support Vector Machines   1. Understanding the spirit and significance of maximum margin classifier 2. Posing an optimization problem for SVM in non-overlapping class scenario 3. Converting the constrained optimization problem into unconstrained using Legrange multipliers 4. Dual of the optimization problem 5. Appreciation of sparse kernel machine and support vectors in the solution of the optimization problem 6. Implementation of SVM in python | 15  20  20  15  25  10 | T1 – 6.1, 6.2 and 7.1 |
| M8 | Support Vector Machines in overlapping class distributions & Kernels   1. Issues of overlapping class distribution for SVM 2. Posing an optimization problem for SVM in overlapping class scenario 3. Solving the optimization problem using Legrange multipliers, dual representations 4. Kernel Trick and Mercer’s theorem 5. Techniques for constructing Kernels and advantages of Kernels in SVM 6. Implementation of SVM using different kernels | 15  20  25  25  25  15 | T1 – 6.1, 6.2 and 7.1 |
| M9 | Ensemble Methods   1. Rational for Ensemble Method 2. Methods for constructing an Ensemble Classifier 3. Bagging, Boosting, AdaBoost 4. Random Forest 5. eXtreme Gradient Boosting (XGBoost) 6. Python Implementation of Random Forest and XGBoost 7. Class Imbalance Problem & approaches to solve it | 10  20  25  20  25  15  20 | R1 - 5.6 and 5.7 |
|  |  |  |  |

**Weekly coverage of the course**

|  |  |
| --- | --- |
| **Week** | **Content / Assignments / Exercises** |
| Week1 | Video Content: M1, M2  Evaluative Quiz: Nil  Exercises (90):   1. List the limitations of some prediction measures (like Accuracy). Provide the scenarios highlighting the limitations and suggest measures countering these shortfalls. 2. Provide a case study where in prediction accuracies differs based on the similarity measures and thus appreciate the importance of similarity measures suitable for a given problem.   Assignment (210):   1. For the case study-2 put up in video M1.4, implement k-NN classifier using Python. Find out the optimal k for the case study and put up appropriate reasoning/arguments for choosing the optimal k. |
| Week 2 | Video Content: M3  Evaluative Quiz (210): Q1  Exercises (90): Problems on conditional independence, Bayes Theorem  Assignment (180): For the case study-2 put up in video M1.4, implement Naïve Bayes Classifier using Python. Check whether the attributes/feature of the case study are conditional independent. |
| Week 3 | Video Content: M4  Evaluative Quiz: Nil  Exercises: Nil  Assignments (300): For the case study-2 put up in video M1.4,  Implement logistic regression using Python. Find out the decision boundary. Discuss the interpretability of the model. Compare performance of the logistics regression with and worth out regularization. |
| Week 4 | Video Content:M5  Evaluative Quiz: Nil  Exercises: Nil  Assignments (300):  Assignment: For the case study-2 put up in video M1.4,  1. Implement Decision Tree Classifier using Python.  2. Implement Reduced Error Pruning and Rule post pruning to combat overfitting.   1. Build decision trees using alternate measures of selecting attributes and compare performances |
| Week 5 | Video Content: M6  Evaluative Quiz (300): Q2 (M4 and M5)  Exercises: Nil  Assignment: Nil |
| Week 6 | Video Content: M7  Evaluative Quiz: Nil  Exercises (120):   1. Given a constrained optimization problem, write its equivalent unconstrained optimization problem using Legrange multiplier 2. Write dual of the above unconstrained optimization problem and KKT conditions   Assignment (180): For the case study-2 put up in video M1.4,   1. Implement SVM using Python 2. Identify the support vectors for the problem stated in case study |
| Week 7 | Video Content: M8  Evaluative Quiz (180): Q3  Exercises: Nil  Assignment (120): For the case study-2 put up in video M1.4,   1. Implement SVM with four appropriate kernels and compare performance with and without kernels. |
| Week 8 | Video Content: M9  Evaluative Quiz: Nil  Exercises: Nil  Assignment: For the case study-2 put up in video M1.4,  1. Implement Random Forest in Python by varying parameter and compare its performance with typical decision tree  2. Implement eXtreme Gradient Boosting (XGBoost) in Python and its performance with that of Random Forest  3. Implement an ensemble classifier of all classifiers that are built during Week 1 to Week 7 and compare results. |
|  |  |

**Evaluation**

|  |  |  |
| --- | --- | --- |
| Evaluation Component | Marks | Type |
| Comprehensive Examination | 40% | Closed |
| Quizzes (3) | 24% | Open |
| Assignments (6) and  Exercises (3) | 36% | Open |

**Learning Outcomes:**

|  |  |
| --- | --- |
| No | Learning Outcomes |
| LO1 | Ability to build appropriate classifier for a given real life business problem |
| L02 | Demonstrate the capability to understand classification algorithms deeply and fine tuning the parameters therein to enhance performance of the classifier |
| LO3 | Ability to build ensemble classifier using well known techniques |
| LO4 | Ability to interpret the regression model |