Introduction to Machine Learning (Jan-April, 2024)

Instructor: Rajdeep Banerjee, PhD

Teaching Assistant: Nimish D, PhD

Syllabus:

1. Introduction to probability and information theory (week 1, theory: 2 Hr, hands-on: 1 Hr)
    1.1. Definitions of probability
    1.2. Random variables
    1.3. Probability distributions – marginal, conditional, chain rule
    1.4. Expectations, variance, covariance
    1.5. Discrete vs. continuous probability distributions
    1.6. Fundamentals of sampling distributions – central limit theorem
    1.7. Bayes' rule
    1.8. Information theory – entropy, mutual information, KL-divergence
    1.9. Basics of some probabilistic concepts used in machine learning (conceptual introduction)
        1.9.1. Naïve Bayes
        1.9.2. Structured probabilistic models (N-gram language models) (Lab)

2. Exploratory data analysis with Python (week 2, hands-on: 3 Hr)
    2.1. Python basics
        2.1.1. Data structures, loops, functions
        2.1.2. Pandas, NumPy
        2.1.3. OOP
    2.2. Data pre-processing
        2.2.1. Handling missing values
        2.2.2. Handling categorical variables
        2.2.3. Feature engineering
        2.2.4. Feature selection
        2.2.5. Scaling
    2.3. Univariate and bivariate analysis

3. Introduction to machine learning (week 3, theory: 2 Hr, hands-on: 1 Hr)
    3.1. Statistical learning
        3.1.1. Supervised vs. unsupervised
        3.1.2. Regression vs. classification
    3.2. Accuracy vs. interpretability
    3.3. Hyperparameters, train-validation-test splits
    3.4. Bias-variance trade-off
    3.5. Maximum likelihood estimation (Lab)

4. Regression (week 4, theory: 2 Hr, hands-on: 1 Hr)
    4.1. Linear regression
        4.1.1. Estimating coefficients
        4.1.2. Estimating errors
    4.2. Gradient descent
    4.3. Shrinkage methods (regularization)
        4.3.1. Ridge regression
        4.3.2. Lasso regression
    4.4. Lab session

5. Classification (week 5, theory: 2 Hr, hands-on: 1 Hr)
    5.1. Logistic regression
        5.1.1. Estimating probabilities
        5.1.2. Cost function
        5.1.3. Choosing the probability threshold
    5.2. K-nearest neighbours
    5.3. Naïve Bayes (revisit)
    5.4. Linear discriminant analysis
    5.5. Comparing different classification models (Lab)

6. Support vector machines (week 6, theory: 2 Hr, hands-on: 1 Hr)
    6.1. Maximal margin classifier
    6.2. Support vector classifier
        6.2.1. Linear- Soft-margin
        6.2.2. Non-linear – the kernel trick
    6.3. Support vector regressor
    6.4. Relation to logistic regression
    6.5. Lab session

7. Unsupervised learning – dimensionality reduction and clustering (week 7, theory: 2 Hr, hands-on: 1 Hr)
    7.1. The curse of dimensionality
    7.2. Principal component analysis
    7.3. K-means clustering
    7.4. Hierarchical clustering
    7.5. Gaussian mixtures (if time permits)
    7.6. Lab session

8. Decision trees (week 8, theory: 2 Hr, hands-on: 1 Hr)
    8.1. Regression trees
    8.2. Classification trees
    8.3. The CART algorithm
    8.4. Computational complexity
    8.5. Gini or entropy?
    8.6. Lab session

9. Ensemble learning – bagging (week 9, theory: 2 Hr, hands-on: 1 Hr)
    - 9.1. Bagging and pasting
    - 9.2. Out of bag evaluation
    - 9.3. Random forests – why?
    - 9.4. Feature importance
    - 9.5. Lab session

10. Ensemble learning – boosting methods (week 10, theory: 2 Hr, hands-on: 1 Hr)
    - 10.1. Boosting
    - 10.2. AdaBoost
    - 10.3. Gradient boost
    - 10.4. XGBoost (introduction)
    - 10.5. Comparison of performance (Lab)

11. Model analysis and data selection (week 11, theory: 2 Hr, hands-on: 1 Hr)
    - 11.1. Cross-validation
        - 11.1.1. Leave-out
        - 11.1.2. k-fold
    - 11.2. Bootstrap
    - 11.3. Introduction to data-centric AI
        - 11.3.1. Detecting label issues
        - 11.3.2. Data selection for retraining
    - 11.4. Lab session

12. Introduction to neural networks (week 12, theory: 2 Hr, hands-on: 1 Hr)
    - 12.1. The perceptron
    - 12.2. Building a neural network from scratch
        - 12.2.1. The feed-forward neural network
        - 12.2.2. Back-propagation
    - 12.3. Stop overfitting
        - 12.3.1. Drop-outs
        - 12.3.2. Early-stop
        - 12.3.3. Batch normalization
    - 12.4. The PyTorch and Tensorflow frameworks (Lab)

13. Doubt clearing session and project ideas discussion (week 13, hands-on: 3 Hr)
    - 13.1. Doubt-clearing session
    - 13.2. Open-source materials datasets
    - 13.3. Materials feature generation tools
    - 13.4. Project/paper topic discussion and selection
    - 13.5. Create groups

14. Project/paper presentation, Q&A (week 14)