Foundations of Data Science

The target groups for this course are the MSc students (with major or minor in data science, or studying computer science) and PhD students across the university.

Learning Outcomes

This course introduces gradually a wealth of supervised and unsupervised learning methods and models.

> Students will learn the algorithms which underpin popular machine learning techniques, as well as developing an understanding of the theoretical relationships between these algorithms.
>
> During the seven exercise sessions, the students will apply the concepts taught in the lectures to further strengthen their understanding of these concepts. The last exercise session is a revision class in preparation for the written exam.
>
> The three practical tasks will concern the application of machine learning to a range of real-world problems.

Course Topics

1. **Introduction to Data Science**

   What is Data Science?; Origins of Data Science; Full Scope of Data Science; Scope of This Course

   Data Science vs: Computer Science, Statistics, Machine Learning

   The Future of Data Analysis (Tukey); The Two Cultures (Breiman)

   The Big Data Buzz

2. **Introduction to Machine Learning**

What is Machine Learning? Programming vs Learning; Evolution of Machine Learning; Machine Learning in Action

An Early Automatic Classification Example and The Perceptron Algorithm

Models and Methods Covered in This Lecture

Applications: House Price Prediction; Object Detection and Classification

Learning Flavours: Supervised (Regression, Classification); Unsupervised (Dimensionality Reduction, Clustering); Active; Semi-Supervised; Collaborative Filtering; Reinforcement Learning

3. **Mathematics for Machine Learning**

   Linear Algebra: Vectors (Vector Norms, Inner Product Spaces); Matrices (Operations); Eigenvectors and Eigenvalues; Positive (Semi-)Definiteness

   Calculus: Continuous and Differentiable Functions of One or Multiple Variables; Finding Extrema (First Derivative Test, Second Derivative Test, Critical Points); Partial Derivatives, Gradients, Hessian, Jacobian; Matrix Calculus, Chain Rule in Higher Dimensions; Optimality with Side Conditions, Lagrange Multipliers

   Probability Theory: Probability Space; Conditional Probability, Bayes Rule; Random Variables; Joint Probability Distributions; Expectation, Variance, Standard Deviation, Covariance; Discrete and Continuous Probability Disctributions (Bernoulli, Binomial, Multivariate Normal, Laplace)

4. **Linear Regression**

   Linear Regression by Example

   Definition, Bias, Noise, One-hot Encoding, Learning vs Testing

   Least Squares Objective, Gradient, Computing the Paramters

   Finding Optimal Solution using Matrix Calculus, Differentiating Matrix Expressions, Deriving the Least Squares Estimates

   Computational Complexity of Parameter Estimation

   Least Squares Estimate in the Presence of Outlliers

5. **Maximum Likelihood**

   Probabilistic vs Optimisation Views in Machine Learning

   Maximum Likelihood Principle, Examples

   Probabilistic Formulation of the Linear Model via Maximum Likelihood, Gaussian Noise Model

   Maximum Likelihood Estimator vs Least Squares Estimator, (Log, Negative Log) Likelihood of Linear Regression

   Outliers, Maximum Likelihood for Laplace Noise Model

6. **Basis Expansion, Learning Curves, Overfitting, Validation**

    Basis Expansion: Polynomial, Radial Basis, using Kernels, Kernel Trick

    How to Choose Hyperparameters for RBF Kernel?

    Generalisation Error, Bias-Variance Tradeoff, Learning Curves

    Overfitting: How Does it Occur? How to Avoid it?

    Validation Error, Training and Validation Curves, Overfitting on the Validation Dataset (Kaggle Learderboard)

    k-fold Cross-Validation, Grid Search

7. **Regularisation**

    Estimate for Ridge Linear Regression, Lagrangian (Constrained Optimisation) Formulation

    LASSO: Least Absolute Shrinkage and Selection Operator

    Effect of Ridge and Lasso Hyperparameter on Weights

8. **Feature Selection**

    Goal, Premise, and Motivation

    Feature Selection to Reduce Overfitting

    Feature Selection Methods: Wrapper methods (Forward Stepwise Selection), Filter methods (Mutual Information, Pearson Correlation Coefficient), Embedded methods (LASSO, Elastic Net Regularisation)

9. **Convex Optimisation**

    Convex Sets, Examples, Proving Common Cases of Convex Sets (PSD Cone, Norm Balls, Polyhedra)

    Convex Functions, Examples

    Convex Optimisation Problems: Classes (Linear Programming, Quadratically Constrained Quadratic Programming), Local vs Global Optima, Proof of Local=Global Theorem

    Examples: Linear Model with Absolute Loss, Minimising the Lasso Objective, Linear Regression with Gaussian Noise Model

10. **First-Order and Second-Order Optimisation**

    Calculus Background: Gradient Vectors, Contour Curves, Direction of Steepest Increase, Sub-gradient, Hessian

    Gradient Descent: Algorithm, Geometric Interpretation, Choosing Step Size (Backtracking Line Search), Convergence Test, Stochastic vs (Mini-)Batch, Sub-gradient Descent

Constrained Convex Optimisation: Projected Ggradient Descent

Newton's Method: second-order Taylor Function Approximation, Geometric Interpretation, Computation and Convergence

11. **Generative Models for Classification**

    Discriminative vs Generative Models

    Supervised Learning: Regression vs Classification

    Generative Classification Model: Definition, Prediction

    Maximum Log-Likelihood Estimator for Class Probability Distribution

    Naïve Bayes Classifier: Training and Predicting with Missing Data

    Gaussian Discriminant Analysis: Maximum Likelihood Estimator, Quadratic/Linear Discriminant Analysis, Two-Class Linear Discriminant Analysis, Decision Boundaries, Sigmoid and Softmax Functions

12. **Logistic Regression**

    Models for Binary Classification

    Logistic Regression: Definition, Prediction, Contour Lines Represent Class Label Probabilities, Negative Log-Likelihood vs Cross-Entropy, Maximum Likelihood Estimate, Newton Method for Optimising the Negative Log-Likelihood, Iteratively Re-Weighted Least Squares

13. **Multiclass Classification**

    One-vs-One, One-vs-Rest, Error Correcting Approach

    Softmax, Multiclass Logistic Regression

14. **Measuring Performance for Classification**

    Confusion Matrix, Sensitivity, Recall, Specificity, Precision, Accuracy; Examples

    ROC (Receiver Operating Characteristic) Curve, Confusion Matrices for Different Decision Boundaries, Area under the ROC Curve

    Precision-Recall Curve

15. **Support Vector Machines**

    Maximum Margin Principle, Support Vectors, Formulation as Convex Optimisation Problem in the Linearly (Non-)Separable Case

    Hinge Loss Optimisation, Hinge vs Logistic

    Primal vs Dual Formulation, Constrained Optimisation with Inequalities, Karush-Kuhn-Tucker Conditions, When to Prefer the Dual Formulation

Kernel Methods: Mercer Kernels in SVM Dual Formulation, Kernel Engineering, Examples with Polynomial, RBF, and String Kernels

16. **Neural Networks**

    Multi-layer Perceptrons: Example, Matrix Notation, Multi-layer Perceptron vs Logistic Regression

    The Backpropagation Algorithm: Example, Forward and Backward Equations, Computational Aspects

    Training Neural Networks: Difficulties (Saturation, Vanishing Gradient, Overfitting), Known Hacks (Early Stopping, Adding Data, Dropout), Rectified Linear Unit, Dying ReLU, Leaky ReLU, Initialising Weights and Biases, Examples

    Convolutional Neural Networks: Convolution, Pattern-Detecting Filters, Convolutional Layers, Pooling, Popular Convolutional Neural Networks, Training

17. **Clustering**

    Clustering Objective

    k-Means Clustering: Algorithm, Convergence, Choosing k,

    Transforming input formats: Euclidean Space, Dissimilarity Matrix, Singular Value Decomposition, Multidimensional Scaling

    Hierarchical Clustering: Linkage Algorithms

    Spectral Clustering

18. **Principal Component Analysis**

    Dimensionality Reduction

    Maximum Variance View vs Best Reconstruction View of Principal Component Analysis

    Finding Principal Components using Singular Value Decomposition and Iterative Method

    Applications: Reconstruction of an Image using PCA, Eigenfaces, Latent Semantic Analysis

---

Practicals

The practical tasks require implementation using jupyter notebooks, Python, Scikit-learn, and TensorFlow.

1. Implementation of Linear Regression (Ridge, Lasso)
2. Comparison of Generative and Discriminative Models
3. Classification of Handwritten Digits using the MNIST dataset and TensorFlow 2.0

Exercise Sessions

1. Mathematical Basics
2. Linear Regression; Perceptron
3. Maximum Likelihood; Regularisation
4. Optimisation; Generative Models
5. Logistic Regression; Support Vector Machines; Kernel Methods
6. Neural Nets; Clustering
7. Revision Class

Prerequisites

Introductory courses on: (1) Calculus, (2) Linear Algebra, (3) Probability Theory, (4) Design and Analysis of Algorithms. The course is not recommended for students without the necessary mathematical background. The students who would like to recall necessary background should consult the following resources:

Very brief overview on › Mathematics for Machine Learning

› Multivariate Calculus

› Linear Algebra

› Linear Algebra and its Applications

Familiarity with Python and jupyter notebook is of advantage as these languages will be used in the practicals.

Recommended Resources

Most material covered in the course can be found in the following books (available online for free, search for them).

C. M. Bishop. Pattern Recognition and Machine Learning. Springer 2006.

I. Goodfellow, Y. Bengio, A. Courville. Deep Learning, MIT Press 2016.

K. P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press 2012.

In addition, students may find the following books useful as supplementary reading.

T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning. Springer 2011. (Available for download on the authors' web-page)

M. Nielsen. Neural Networks and Deep Learning.

Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.

S. Boyd, L. Vandenberghe. Convex Optimization, Cambridge University Press 2004.