CMPUT 367 - Intermediate Machine Learning

Calendar Description

This second course in machine learning focuses on higher-dimensional data and a broader class of nonlinear function approximation approaches. Topics include: optimization approaches (constrained optimization, hessians, matrix solutions), kernel machines, neural networks, dimensionality reduction, latent variables, feature selection, more advanced methods for assessing generalization (cross-validation, bootstrapping), introduction to non-iid data and missing data. Prerequisites: CMPUT 204 and 267; one of MATH 115, 118, 135, 145, or 155.

Longer Description

Machine Learning is all about analyzing high-dimensional data. The goal for this second course in machine learning is to expand on the foundations from the first course. We will revisit several of the concepts--including how models can be estimated from data; sound estimation principles; generalization; and evaluating models--but with the additional nuances from handling high-dimensional inputs. Topics include: optimization approaches (constrained optimization, hessians, matrix solutions), kernel machines, neural networks, dimensionality reduction, latent variables, feature selection, more advanced methods for assessing generalization (cross-validation, bootstrapping), introduction to non-iid data and missing data.

Overview

- Multivariate probability, covering both discrete and continuous cases
- Analyzing high dimensional data
- Understanding the perils of high dimensional spaces
- Introduction to nonlinear models and representations of data
- Introduction to non-IID data and missing data

Learning outcomes

By the end of the course, you should understand…

- The design process for solving a real data analysis problem:
    - identifying key data issues (high-dimensionality, dependence, missing data)
    - identifying an appropriate model and optimization problem
    - Identifying an appropriate algorithm and understanding its assumptions
- Representative instances of different learning algorithms
    - maximum likelihood for a broader range of problem settings

    - Data re-representation approaches, including dimensionality reduction approaches, kernel machines and neural networks
- Generalization, and how it relates to the complexity of the hypothesis class
- Evaluation of learned models
    - including resampling approaches to use data efficiently
    - the role of cross-validation to select hyperparameters
    - evaluating generative models, in addition to predictive models

By the end of the course, you will have improved your skills in…

- Implementing more advanced estimation approaches (e.g., optimization algorithms for neural networks) in python
- Applying mathematical concepts to solve real data problems
- Problem solving, by facing open-ended data analysis problems and needing to both formulate the problem and identifying appropriate algorithms to solve the problem

Technical topics

- Multivariate Probability basics
    - discrete multivariate distribution
    - continuous multivariate distributions
    - covariance matrices and multivariate Gaussians
    - Bayesian networks
    - entropy and KL-divergences
- Estimation for general distributions
    - generalized linear models (review linear regression, logistic regression)
    - mixture models
    - expectation-maximization
- Matrices
    - eigenvalues and singular value decomposition
    - orthogonality
    - matrix norms
- Dimensionality reduction
    - principal components
    - canonical correlation analysis
    - random projections, Johnson-Lindenstrauss
- Kernel representations
    - re-representing data using similarities to prototypes
    - high-dimensional phenomena and the curse of dimensionality
    - kernel similarities (RBFs, matching kernel, graph Laplacians)
    - nonlinear dimensionality reduction with kernels (e.g, Isomap)
- Nonlinear prediction models for regression and classification

    - neural networks
    - kernel machines
    - KNNs (k-nearest neighbors), using kernels
- Optimizing an n-D function
    - gradients and Hessians
    - convexity, positive definiteness
    - quasi-second-order algorithms (and relation to stepsize selection)
    - constrained optimization
- Imputation of missing variables
    - using a learned distribution
    - matrix completion
- Assessing generalization and evaluating models
    - cross-validation and bootstrapping
    - assessing generative models
- Regularization and overfitting
    - l1 for feature selection (sparsity)
    - early stopping and validation sets
- Generalization theory basics
    - rademacher complexity
    - uniform convergence
- Introduction to temporal data
    - introduction to time series data
    - Markov chains
    - prediction using histories

Knowledge Prerequisites

This course follows CMPUT 267, and relies on the understanding of the basic concepts in ML taught in that course. We will review many of these concepts, but now in more advanced settings (e.g., maximum likelihood for mixture models). The course relies on more knowledge in calculus and linear algebra than was needed for CMPUT 267. The numerical methods course (CMPUT 340) is a complementary and useful course for CMPUT 367, and so is a recommended co-requisite. An excitement to understand the mathematics underlying machine learning is a must.

Pre-requisites

- One of MATH 115, 118, 145 or 155 (Calculus II)
- MATH 125 or 127 (Linear algebra)
- CMPUT 204 (Algorithms)
- CMPUT 267 (Basics of ML)

More syllabus details:

Evaluation:

Quiz: 5%
Midterm: 20%
Final: 35%
Assignments (3): 30%
Thought Questions: 10%

Marks will be converted to Letter Grades at the end of the course, based on relative performance. There are no set boundaries, because each year we modify exams and there is some variability in performance. Set boundaries would penalize students in a year where we inadvertently made a question too difficult. A good indicator for final performance is performance on the exams, which are a large percentage of the grade. If you fail all three exams (less than 50% on all three), then you will likely get an F in the course.

Textbook/Materials

The notes are written specifically for this course, and provided on the website. These are designed to be short, so that you can read every chapter. I recommend avoiding printing these notes, since later parts of the notes are likely to be modified (even if only a little bit), since these notes are still being improved.

You are expected to read the corresponding sections about a class's topic from notes before class as each class will discuss each topic in more detail and address questions about the material.

Lab requirements

There is no formal lab. Instead, we may hold extra sessions to go over any requested background topics. TAs will hold weekly office hours.

Late Policy

Any late work will not be accepted and will receive 0 marks.

Academic Honesty

All assignments and exams are individual, except when collaboration is explicitly allowed. All the sources used for problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see https://www.deanofstudents.ualberta.ca/en/AcademicIntegrity.aspx.