CMPUT 296 - Basics of Machine Learning

Description

The field of machine learning involves the development of statistical algorithms that can learn from data, and make predictions on data. These algorithms and concepts are used in a range of computing disciplines, including artificial intelligence, robotics, computer vision, natural language processing, data mining, information retrieval, bioinformatics, etc. This course introduces the fundamental statistical, mathematical, and computational concepts in analyzing data. The goal for this introductory course is to provide a solid foundation in the mathematics of machine learning, in preparation for more advanced machine learning concepts. The course focuses on univariate models, to simplify some of the mathematics and emphasize some of the underlying concepts in machine learning, including how should one think about data; how can data be summarized; how models can be estimated from data; what sound estimation principles look like; how generalization is achieved; and how to evaluate the performance of learned models.

Overview

- Basic probability concepts, covering both discrete and continuous cases
- Basic optimization concepts, needed for estimating models
- Analyzing scalar data
- Analyzing paired data

Learning outcomes

By the end of the course, you should understand…

- The design process for solving a data analysis problem:
    - properties of the data
    - choosing a model
    - defining a computational problem (e.g. an optimization problem)
- Basic estimation algorithms, including maximum likelihood and linear regression and different optimization approaches for those problems
- Generalization, including the concept of over-fitting
- Evaluation of learned models

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

- Implementing basic estimation approaches (e.g., stochastic gradient descent for linear regression) in python
- Applying concepts from calculus and probability 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

- Univariate probability
    - discrete distribution
    - continuous distribution
    - random variable
    - cdf, pdf
    - expectation
        - linearity
        - conditional expectation
    - variance
- Histograms
    - median
    - quantiles
    - outliers
- Concentration phenomenon
    - basic central limit theorem
    - simple inequalities (Chebyshev, Hoeffding-Chernoff)
    - a foundation of generalization
- Estimation principles
    - maximum likelihood
    - maximum a posteriori
    - Bayesian approaches
    - regularization
- Optimizing functions
    - The role of derivatives and stationary points
    - First order and second order gradient descent
    - Batch and stochastic gradient descent
    - Step-size selection
- Regression
    - paired data (predicting target y from input x)
    - linear regression
    - polynomial regression
    - Bayesian regression
- Classification
    - logistic regression

- Regularization
    - The importance of constraining the function class
    - The impact on generalization error

Knowledge Prerequisites

In this course, we will cover some basics in probability and optimization that you will need for the course. However, you will be applying these concepts for machine learning, and so it is important that (a) you have been exposed to some of the concepts before, and (b) are at least enforcing some of the mathematical knowledge in parallel. You must have taken calculus before this course, and have some programming experience. Background in probability and a first course in programming is recommended to be taken before this course, but can be taken as a co-requisite. An excitement to understand the mathematics underlying machine learning is a must.

The course CMPUT 272 is included as a co-requisite, as that course helps you become more comfortable with mathematical formalization. This co-requisite is particularly pertinent to those in CS, where CMPUT 272 is a requirement. For other departments, other math classes might provide that background, and can be used in place of CMPUT 272.

Pre-requisites

- One of MATH 100, 114, 117, 134 or 146
- CMPUT 174 or 274

Co-requisite

- One of STAT 141, 151, 235 or 265 or SCI 151
- MATH 125/127 (Linear algebra)
- CMPUT 272

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

3 hour lab held weekly, in a classroom. This will be a question-and-answer period mainly, with some tutorials given by TAs. The first lab will be a python tutorial.

Late Policy

Any late work will not be accepted and will recieve 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.