Introduction to Machine Learning

Syllabus

Introduction to Machine Learning

Time and Place

Online Nov 17, 2020– Jan 11, 2021

Description

This course is part of the Key Capabilities for Data Science program and will introduce you to key concepts in machine learning.

During the course, you will work with powerful Python libraries made for data-science, including Pandas for processing tabular data, Altair for data visualization and NumPy for working with numerical data types. In this course, we will be focusing our attention on a machine learning library called Scikit Learn.

This course covers the data science perspective on the introductory concepts in machine learning, with a focus on making predictions. It covers how to build different models such as K-NN, decision trees and linear classifiers as well as important concepts such as data splitting and fundamental rules and laws. In addition, this course will teach you how to evaluate models properly and question their validity all while streamlining the process with pipelines.

Prerequisites

- Programming in Python for Data Science

Content

We are using an interactive platform that contains a series of slide decks accompanied by multiple-choice questions and interactive coding exercises to apply what you learned during the module: https://ml-learn.mds.ubc.ca/. We recommend using Chrome, Firefox or Safari to complete the content as other browsers are incompatible with the content platform.

Course Software Platforms

Students will write code and perform their analysis using the Python Programming Language. Assignments as well as the final project analysis will be done using Jupyter Notebooks. Students will access the worksheets and tutorials in Jupyter Notebooks through Canvas. Students will require a laptop, Chromebook or tablet to complete this course.

Learning Outcomes

By the end of the course, students will be able to:

- Describe supervised learning and identify what kind of tasks it is suitable for.
- Explain common machine learning concepts such as classification and regression, training and testing, overfitting, parameters and hyperparameters, and the golden rule.
- Identify when and why to apply data pre-processing techniques such as scaling and one-hot encoding.

- Describe at a high level how common machine learning algorithms work, including decision trees, and k-nearest neighbours.
- Use Python and the scikit-learn package to develop an end-to-end supervised machine learning pipeline.

Course Facilitator

Position	Name	Email	Office Hours	Office Location
Course Facilitator	Socorro Dominguez	sedv8808@gmail.com	TBD	Zoom

Course Structure

There are 8 modules in total for this course. Each module will be accompanied by an auto-graded assignment to be submitted. The students will be responsible to complete the material of 2-3 modules every 2 weeks (3 modules for weeks 1-2, 2 modules for weeks 2-4, 3 modules for weeks 4-6). In weeks 4 and 7, students will need to take an online 45-minute open-book quiz testing, material from Modules 1-4 and 5-8 respectively. This quiz can be taken at any point during the week. At the end of the 8th module, students will have an individual project to complete where they are required to use the skills they learned in the course and produce an analysis regarding an assigned dataset.

Assignments

Each week students will submit 2 assignments for a grade. **Assignment due dates are posted on Canvas.** To open the assignment, click the link (e.g. `assignment_01`) from Canvas. To submit your assignment, just make sure your work is saved (File -> Save and Checkpoint to be sure) **on our server** (i.e., using the link from Canvas) before the deadline. Our server will automatically snapshot at the due date/time. We have supplied informative videos on how to do this in assignment 1 and referenced the links in subsequent assignments.

Course breakdown

Deliverable	Percent Grade
Assignments	56 (7% each)
2 Quizzes	32 (16% each)
Final project	12

Schedule

Module	Title	Description
Module 1	Machine Learning Terminology	Different branches of machine learning and the steps needed to build a model by constructing baseline models.
Module 2	Decision Trees	The structure of decision trees and the process it takes to make predictions. Introduce the concept of generalization.
Module 3	Splitting, Cross-Validation and the Fundamental Tradeoff	Why and how we split data and how cross-validation works on training data. The fundamental tradeoff and the golden rule.
Module 4	Similarity-Based Approaches to Supervised Learning	Similarity-based models specifically k-Nearest Neighbours (also known as k-NNs) and Support Vector Machines (SVMs with an RBF kernel).

Module	Title	Description
Module 5	Preprocessing Numerical Features, Pipelines and Hyperparameter Optimization	Preprocessing numeric data. Preparing data before model building through imputation and scaling, building pipelines and automated hyperparameter optimization.
Module 6	Preprocessing Categorical Variables and Sklearn's ColumnTransformer	Preprocessing categorical data. Preparing data before model building through one-hot encoding and implementing `ColumnTransformer` from the sklearn library for more complex pipelines.
Module 7	Assessment and Measurements	Appropriately assessing your model. How to evaluate and calculate your model using an assortment of different measurements.
Module 8	Linear Models	Linear models such as Logistic regression how to interpret them and their advantages and limitations.

Submission Schedule

Assignments are due on Sunday at 6 pm. Office hours will be held twice a week to support and answer questions regarding the concepts learned in the lecture. Quizzes will be opened for up to 7 days, please check Canvas for quiz submission dates.

Week	Modules to Complete	Submissions Due
Week 1 -2	Module 1 Module 2 Module 3	assignment_01, assignment_02, assignment_03
Week 3-4	Module4Module 5	assignment_04, assignment_05, quiz_01
Week 5-6	Module 6Module 7Module 8	assignment_06, assignment_07, assignment_08
Week 7	NA	quiz_02, final_project

Policies

Late Assignments

Students should submit all assignments by the due date. If for any reason a student is unable to do so, they can submit the assignments before the course end date. This is the final allowed submission date and is non-negotiable.

Late Quizzes

We allow 2 weeks to complete the first quiz and 1 week to complete the second. if you've submitted a quiz late or did not submit a quiz at all, please contact the course facilitator and we will try to make accommodations.

Autograder Policy

Many of the questions in assignments are graded automatically by the software. The grading computer has exactly the same hardware setup as the server that students work on. Students should make sure their assignments are *reproducible*, and run from beginning to end on the auto-grading computer. In particular, **please ensure that any data that needs to be downloaded is done so by the assignment notebook with the correct filename to the correct folder**.

Re-grading

If you have concerns about the way your work was graded, please contact the course facilitator.

Device/Browser

Students are responsible for using a device and browser compatible with all functionality of Canvas. Chrome or Firefox browsers are recommended; Safari has had issues with Canvas quizzes in the past.

Academic Integrity

The academic enterprise is founded on honesty, civility, and integrity. As members of this enterprise, all students are expected to know, understand, and follow the codes of conduct regarding academic integrity. At the most basic level, this means submitting only original work done by you and acknowledging all sources of information or ideas and attributing them to others as required. This also means you should not cheat, copy, or mislead others about what is your work. Violations of academic integrity (i.e., misconduct) lead to the breakdown of the academic enterprise, and therefore serious consequences arise and harsh sanctions are imposed. For example, incidences of plagiarism or cheating may result in a mark of zero on the assignment or exam and more serious consequences may apply if the matter is referred to the President's Advisory Committee on Student Discipline. Careful records are kept in order to monitor and prevent recurrences.

A more detailed description of academic integrity, including the University's policies and procedures, may be found in the Academic Calendar at http://calendar.ubc.ca/vancouver/index.cfm?tree=3,54,111,0.

Plagiarism

Students must correctly cite any code or text that has been authored by someone else or by the student themselves for other assignments. Cases of plagiarism may include, but are not limited to:

- the reproduction (copying and pasting) of code or text with none or minimal reformatting (e.g., changing the name of the variables)
- the translation of an algorithm or a script from a language to another
- the generation of code by automatic code-generation software

An "adequate acknowledgment" requires a detailed identification of the (parts of the) code or text reused and a full citation of the original source code that has been reused.

The above attribution policy applies only to assignments. **No code or text may be copied (with or without attribution) from any source during a quiz or exam. Answers must always be in your own words. At a minimum, copying will result in a grade of 0 for the related question.**

Repeated plagiarism of any form could result in larger penalties

Attribution

Parts of this syllabus (particularly the policies) have been copied and derived from the UBC MDS Policies as well as the Syllabus from UBC's DSCI 100 course.