Some ideas for ML in DAT157

A draft proposal for the new machine learning component of DAT157. Under construction. Will be presented at the department seminar on November 21st. http://prosjekt.hib.no/ict/seminars

---

DAT157-ML: The fundamentals of machine learning
– Concepts, tools and techniques to build intelligent systems

Over the past few years the field of "artificial intelligence" has gone through a dramatic development. Computer's abilities to recognize objects in images have gone from being practically useless to reach an almost human level; from limited ability to understand and synthesize text and speech to wide use of personal digital assistants; from playing at an amateur level in chess, poker, Go and Dota to beating world champions; from driving assistance to self-driving cars.

What's behind these developments are breakthroughs in *machine learning*, a set of techniques enabling computers to uncover complicated patterns and connections in large data sets.

The course provides a project-based, hands-on tour through the fundamentals of machine learning, focusing on solving real, practical problems. It covers several useful methods and techniques from machine learning, with an emphasis on predictive analytics using Python and Scikit-learn.

The way the material is covered is inspired by the following principle:

1. What is the problem we want to solve?
2. Why do we want to solve it?
3. How is it solved in practice?
4. Theoretical explanations

> *We can be sure we understand something only when we can think and act flexibly with what we know*

By the end of the course you will have an understanding of the main algorithms in machine learning, and have experience with solving real-world problems using modern tools and frameworks from data analysis and machine learning.

Litterature and curriculum

Litterature

- Main textbook: A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2017
- Possibly supplemented with selected material from S. Raschka, Python Machine Learning, 2016
- Selected material from M. Nielsen, Neural Networks and Deep Learning (free online book)

Online supplemental material

The supporting course DAT157x contains various supplemental material in the form of readings, exercises and videos. Each week of DAT157 will consist of (i) a lecture, (ii) a lab, and (iii) an online module in DAT157x.

Course content
By the end of the course you will be familiar with the following techniques and tools (and more):

Techniques and concepts: Binary classification, cross-validation, confusion matrix, precision and recall, the ROC curve, multiclass classification, error analysis, multilabel classification, kNN, multioutput classification, linear regression, normal equation, gradient descent, polynomial regression, bias/variance tradeoff, regularization (ridge, lasso, elastic net, early stopping), logistic regression, softmax regression, linear and nonlinear SVM classification, kernels (polynomial, Gausssian RBF), SVM regression, decision trees, CART algorithm, Gini impurity, entropy, bagging, pasting, random patches and subspaces, random forests, feature importance, boosting, neural networks, backpropagation, a taste of deep learning.

Tools: Python, Numpy, Pandas, Jupyter Notebooks, Scikit-learn, GitHub, (cloud computing)

Recommended prerequisites

Proficiency in programming is a prerequisite. Knowledge of Python, Numpy and Pandas is a plus. In case these are unfamiliar to you, several online resources for acquiring the necessary background knowledge will be provided in DAT157x before the course begins (and will form the course's "Week 0").

Learning outcomes

Knowledge
- Knowledgeable about the fundamental concepts in machine learning
- Able to explain the basic algorithms in machine learning
- Able to explain what "learning" means in the context of machine learning
- Able to explain how machine learning can be used to solve practical problems from a wide variety of domains, and how it can be used to build "intelligent" applications

Skills
- Develop solutions that can solve concrete, practical problems using machine learning
- Can find and use modern, state-of-the-art software tools and frameworks for data analysis, visualization, reporting and code-sharing
- Design and develop "intelligent" applications using machine learning

General competence / soft skills
- Ability to formulate and complete a machine learning project
- Ability to present their work, both in writing and orally
- Ability to work in teams

Teaching methods

Lectures, labs, online supplemental material (DAT157x).

Obligatory projects

Project 1: End-to-end machine learning project. Graded as pass/fail.

Project 2: A self-selected machine learning project (subject to approval from the lecturer)

Evaluation

Oral exam.

A majority of the exam will be dedicated to a presentation of Project 2, and questioning related to the project work. A smaller part will consist of general questions about the material covered in the course.

Aids at the examination

No restrictions, except direct communication with people outside of the examination room. The student should bring their computer to present their project work.

Lectures

- The lectures will cover a lot of material in a relatively short time. It is necessary to work through the reading materials before each lecture, and helpful to also go through the weekly supplemental material in DAT157x. It's also important to realize that it's when you solve problems yourself that you really learn.

    A major illusion on which the school system rests is that most learning is the result of teaching.

Part 0: Some setup before the course starts, plus establishing background knowledge

Most of the programming will be done in Python. Knowledge of Python, Numpy and also Jupyter Notebooks will be crucial.

You'll need quite a lot of linear algebra, and also some calculus and probability theory. It's useful to refresh the basics of linear algebra before the course starts. One should at least be familiar with matrices and vectors, and common operations on these.

DAT157x Part 0 contains supplemental material, including the links below.

Todo in Part 0:

- Read through the sources below
- Install Python + necessary packages (using Anaconda)
- Join the course classroom on GitHub Classroom: https://classroom.github.com/classrooms/33482132-dat157-ml
- Create an account at our JupyterHub

Sources:

- DAT157x, part 0
- Refreshers of linear algebra:
  - https://betterexplained.com/articles/linear-algebra-guide/
  - https://machinelearningmastery.com/linear-algebra-machine-learning/
- A quick taste of Python: http://www.learnpython.org
- Python tutorials at https://www.python.org
- CS231n Python Numpy tutorial: http://cs231n.github.io/python-numpy-tutorial
- CS231n IPython tutorial: http://cs231n.github.io/ipython-tutorial
- DataCamp Jupyter Notebook Tutorial: The Definitive Guide: https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook#gs.iNj8ujk

Part 1: What is machine learning

Lectures: What is machine learning?

- An introduction to DAT157: motivation, content, plans.
- An introduction to machine learning: What is it? What is it used for? The history of machine learning, types of machine learning problems, and a birds-eye view of what machine learning is trying to achieve.
- Some fundamental concepts: supervised/unsupervised learning, training, training data, overfitting, underfitting, testing, validating.
- Presentation of Project 1.

Sources:

- Géron Ch. 1
- *Additional sources TBA (medicine, cyber security, physics, commerce, ...)*

Supplemental material:

- DAT157x, part 1

Lab: End-to-end machine learning project

This will form the basis of Project 1, which should be submitted during part 3.

Through this project you will become intimately familiar with crucial steps that are part of almost any machine learning project: creating a plan, obtaining data, visualizing and exploring, data cleaning and preparation, machine learning model selection, and model fine-tuning.

Sources:

- Géron Ch. 2
- Datasets:
  - Higgs Boson Machine Learning Challenge: https://www.kaggle.com/c/higgs-boson
  - Machine learning for cyber security: https://github.com/jivoi/awesome-ml-for-cybersecurity#-datasets
  - SSB: https://www.ssb.no/omssb/tjenester-og-verktoy/api
  - Kaggle: https://www.kaggle.com/datasets
  - UCI repo: https://archive.ics.uci.edu/ml/datasets.html
  - https://www.reddit.com/r/datasets/
  - Amazon AWS Public Datasets: https://aws.amazon.com/public-datasets/

Part 2: Classification

Lectures: Classification
Binary classification, cross-validation, confusion matrix, precision and recall, the ROC curve, multiclass classification, error analysis, multilabel classification, multioutput classification

Sources:

- Géron Ch. 3
- Raschka Ch. 3

Supplemental material:

- DAT157x, part 2

Lab: Continuing the end-to-end machine learning project started in part 1

Part 3: Linear and logistic regression, gradient descent and regularization

Lectures: Training models
Linear regression, gradient descent, polynomial regression, regularization (ridge, lasso, elastic net, early stopping), logistic regression, softmax regression.

Sources:

- Géron ch. 4

- *Additional sources TBA*

Supplemental material:
- DAT157x, part 3

Lab: Implementing gradient descent
In this lab you will implement various versions of gradient descent using Numpy. This is a fundamental algorithm in machine learning, particularly important when we come to neural networks.

Sources:
- *TBA*

Note: project 1 deadline this week
The first project should be submitted to the DAT157 GitHub Classroom during this week: **Project 1: End-to-end machine learning project**

Part 4: Support vector machines and decision trees

Lectures: SVM and decision trees

Linear and nonlinear SVM classification. Kernels (polynomial, Gausssian RBF), SVM regression, (decision function, quadratic programming, dualization, kernelized SVM, online SVMs).

Decision trees, CART algorithm, Gini impurity, entropy.

Sources:
- Géron ch. 5 and 6
- *TBA*

Supplemental material:
- DAT157x, part 4

Lab: Theoretical lab
In this lab we will focus on solving theoretical exercises, using both mathematics and computation.

Sources:
- *TBA*

Part 5: Ensemble learning and random forests

Lectures: Ensembling and random forests
Bagging, pasting, random patches and subspaces, random forests, feature importance, boosting (AdaBoost, Gradient Boosting, XGBoost)

Sources:
- Géron ch. 7
- *TBA*

Supplemental material:
- DAT157x, part 5

Lab: A machine learning project based on a dataset selected by the student
This lab will form the basis for Project 2.

Sources:
- Datasets:
  - Higgs Boson Machine Learning Challenge: https://www.kaggle.com/c/higgs-boson
  - Machine learning for cyber security: https://github.com/jivoi/awesome-ml-for-cybersecurity#-datasets
  - SSB: https://www.ssb.no/omssb/tjenester-og-verktoy/api
  - Kaggle: https://www.kaggle.com/datasets
  - UCI repo: https://archive.ics.uci.edu/ml/datasets.html
  - https://www.reddit.com/r/datasets/

Part 6: Neural networks

Lectures: Neural networks
Neural networks, backpropagation, and a taste of *deep learning*.

Sources:
- Nielsen ch. 1
- Nielsen ch. 2
- Raschka ch. 12
- 3Blue1Brown, *But what *is* a Neural Network? | Deep learning, chapter 1*: https://www.youtube.com/watch?v=aircAruvnKk
- 3Blue1Brown, *Gradient descent, how neural networks learn | Deep learning, chapter 2:* https://www.youtube.com/watch?v=IHZwWFHWa-w
- 3Blue1Brown, *What is backpropagation and what is it actually doing? | Deep learning, chapter 3*: https://www.youtube.com/watch?v=Ilg3gGewQ5U
- 3Blue1Brown, *Backpropagation calculus | Appendix to deep learning chapter 3:* https://www.youtube.com/watch?v=tIeHLnjs5U8
- Calculus on Computational Graphs: Backpropagation. http://colah.github.io/posts/2015-08-Backprop/

Supplemental material:

- DAT157x, part 6

Lab: Neural networks and backpropagation

Based on Nielsen's exercises in chapters 1 and 2: http://neuralnetworksanddeeplearning.com. Implementing neural networks and backpropagation essentially from skratch.

Source:

- https://github.com/mnielsen/neural-networks-and-deep-learning

---

Diverse notater

+ELMED219 BMED CompBioMed

Noe overlapp med

- MLM: https://akademix.no/courses/course-v1:AkademiX+MLM+2018/about
- ELMED219: https://akademix.no/courses/course-v1:UiB+ELMED219x+2018/about, www.uib.no/emne/ELMED219
- CBM101: https://nordbiomed.akademix.no/

- At the end of each lecture, the students will be asked to fill out a "one minute paper":
  What was the most important things I learned? What was the most confusing part? One question I still have after today's lecture.

  The first part of the following lecture will be dedicated to discussing the responses from the previous lecture.