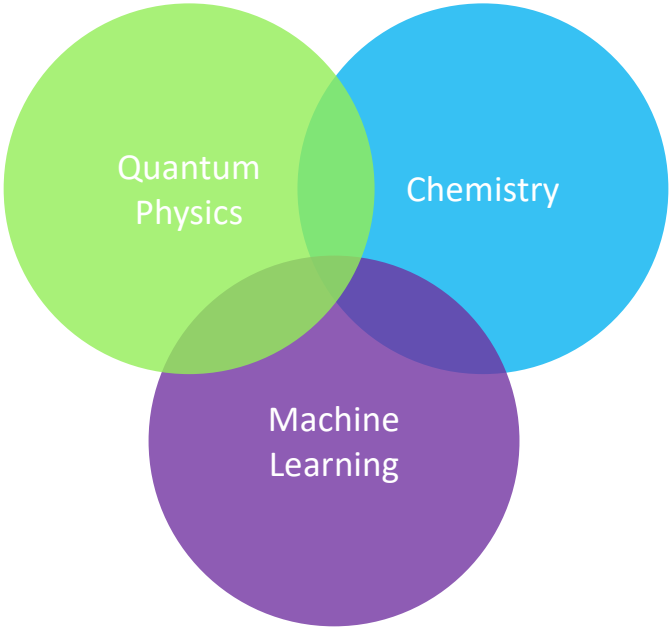
Chemistry 507B

Machine Learning for Chemistry

Professor Roman Krems



https://www.chem.ubc.ca/roman-krems

Course outline

Major goals of this course:

⇒ Learn how machine learning (ML) works
⇒ Learn how ML can help solve Chemistry problems
⇒ Get hands-on experience with ML analysis
⇒ Apply ML to a research problem

This course will include three components:

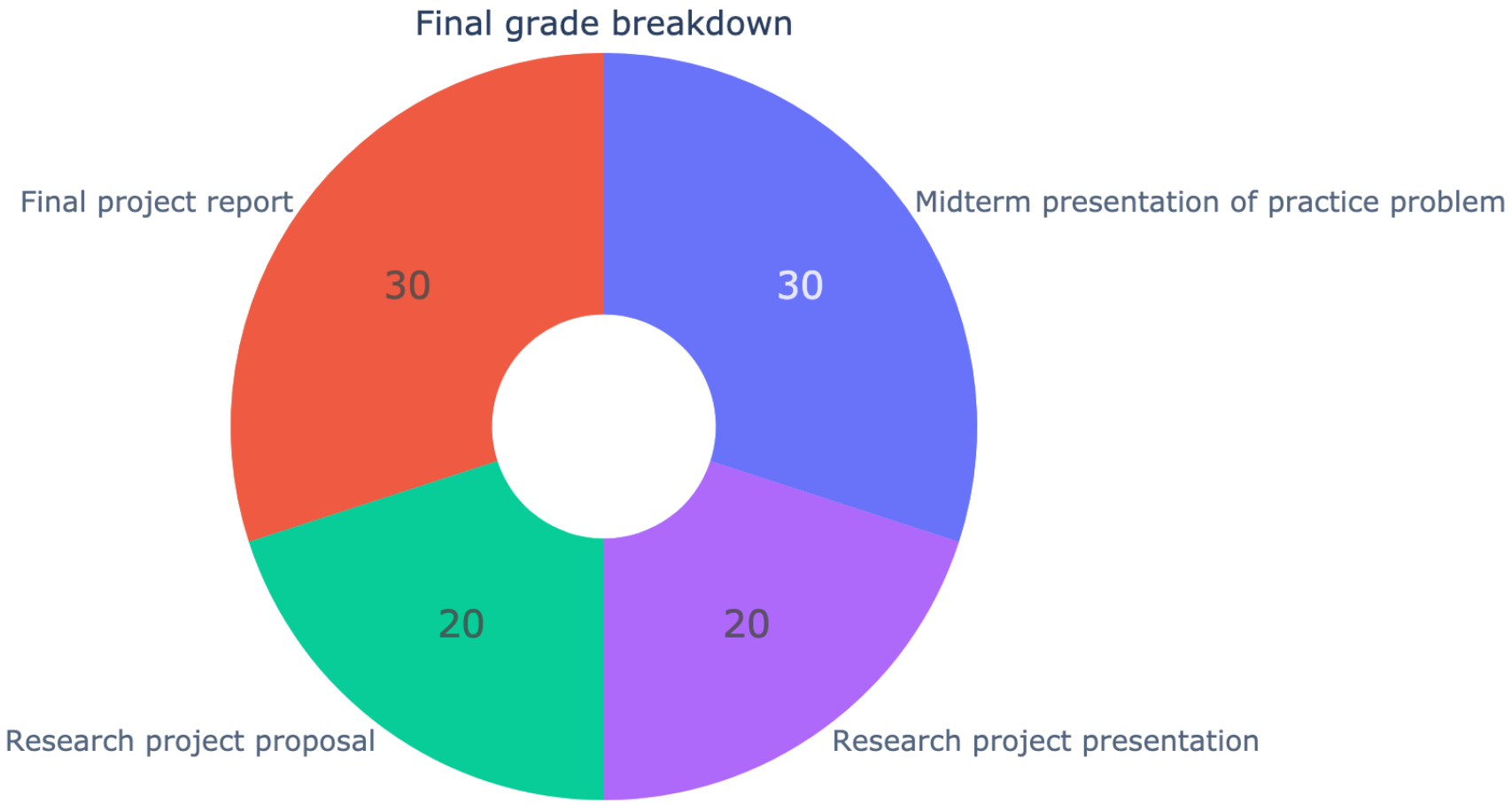
I - Theoretical
⇒ Learn basic theory of ML
⇒ Discuss applications of ML to Chemistry
II - Practical
⇒ Learn to code ML models using python packages
III - Application of ML tools to a research problem
⇒ Use tools learned for a problem in your lab

Summary of grading scheme:

- Midterm presentation of the practice problem: 30
- Research project proposal: 20
- Research project presentation: 20
- Final project report: 30

Important note

To receive full mark for the final report, students need to prove that the work presented may lead to a peer-reviewed publication. In other words, the ultimate goal is to obtain a research insight or develop a new research methodology by combining the tools learned in this course with the students' own research.


Final grade breakdown
Final project report
30
Midterm presentation of practice problem
30
20
20
Research project proposal
Research project presentation

List of discussion topics:

- General overview of ML problems

  supervised learning, unsupervised learning, reinforcement learning, optimization

  regression, classification, clustering, interpolation, generalization, dimensionality reduction

- General review of how ML works

  feature space, feature-engineering, non-linear transformations, high-dimensional feature spaces, feature mapping

  regression, underfitting and overfitting, the bias-variance trade-off, regularization, reproducing kernel Hilbert space, kernel trick, neural networks

- Linear regression as a neural network problem

- Bayes's theorem, Likelihood, Marginal Likelihood

- Linear classification (binary)

  Logistic regression, Perceptron Learning Algorithm, Linear Discriminant Analysis

- Dimensionality reduction

  Principle Component Analysis, Clustering

- Regularization

  Bias-variance trade off, Regularization, Ridge Regression

- Reproducing Kernel Hilbert Space

  Hilbert Space of Functions, Reproducing Kernels, RKHS, Kernel Ridge Regression

- Nonlinear Classification

  Support Vector Machines, Hinge Loss

- Bayesian Machine Learning

  Bayesian Neural Networks, Gaussian Processes, Gaussian Process Regression, Bayesian Optimization

- Machine Learning for Chemistry

  Feature Spaces for Chemistry applications, Chemistry Optimization, Molecular Descriptors, Self-driving Labs, Inverse Problems, Interpolation in Chemical Space, Δ-Learning

- How to build better kernels

  Model Selection Metrics, Marginal Likelihood, Bayesian Information Criterion, Extrapolation with Machine Learning, Bayesian Optimization of Time-consuming Experiments

- Multiclass classification

  Reduction to Binary Classification, Linear Discriminant Analysis, Naive Bayes, Classification with Neural Networks

- Neural Networks

  Neuron Activation Functions, Dense Neural Networks, Recurrent Neural Networks

- Convolutional Neural Networks

  Convolution, Translational Equivariance, Translational Invariance, Pooling, Cats vs dogs

- Data Visualization

Lectures/discussions

What we will aim to do:

$\Rightarrow$ Discuss theory behind ML

$\Rightarrow$ Discuss useful mathematics/statistics concepts

$\Rightarrow$ Discuss examples of ML applications in Chemistry

$\Rightarrow$ Discuss examples of python code, if necessary

All python codes used in lectures will be posted on Canvas.

Download them and play with them before each class.

List of 5 applications for the coding excercises:

- Principle component analysis
- Kernel ridge regression
- Support vector classification
- Regression with a Deep Neural Network
- Gaussian process regression (interpolation)

List of optional but recommended applications for coding excercises:

- Bayesian optimization
- Convolutional Neural Networks
- Recurrent Neural Networks
- Gaussian process regression (extrapolation)
- Multi-output regression