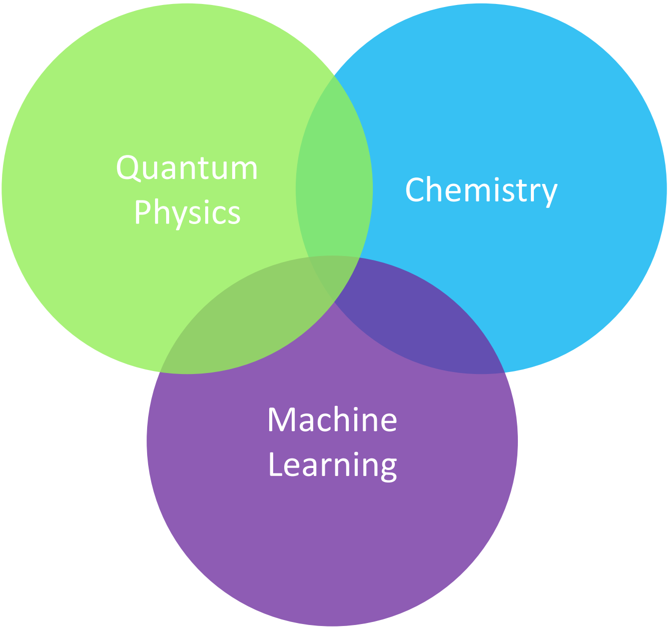
Chemistry 507B

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

I - Review of machine learning

The goal is to understand:

- ⇒ How ML works
- ⇒ What kind of problems ML solves
- ⇒ The underlying theory and how different methods are related
- ⇒ How ML can be used for Chemistry applications

This part will be in the form of lectures and discussions.

II - Practical exercises

The goal is to code using python packages:

⇒ ML models for several different applications
⇒ Each student will develop their own code

How is this going to work?

⇒ Each student will produce code for 5 different models
⇒ These models will cover different types of ML problems
⇒ Each of these models will be discussed in lectures
⇒ Data sets for different applications will be provided
⇒ Students can use their own data sets, if desired
⇒ Work on code from Sept 1 to Oct 18

Five classes starting on Oct 18 will be used for assessment of this work:

$\Rightarrow$ Each student will present and demonstrate their code for one randomly picked model

$\Rightarrow$ Each presentation will be 15 min, including questions

$\Rightarrow$ A random number generator will select the presenter

$\Rightarrow$ A random number generator will select the problem to be presented <u>on the day of the presentation</u>

$\Rightarrow$ Each student will be examined on the code and the theory behind the model presented

Each student will be expected to receive a mark of 30 for this examination unless their model fails to produce desired results or the student fails to answer the questions about the model.

III - Application of ML tools to a research problem

The goal is to apply ML tools to a research problem

How is this going to work?

⇒ Students will write a 2-page research proposal including:

- ∘ Statement of the research question
- ∘ Formulation of the hypothesis
- ∘ Description of the data to be analyzed
- ∘ Description of the ML method to be used
- ∘ Discussion of expected outcomes

⇒ This proposal must be submitted by October 31

⇒ The proposal will be marked and contribute up to 20 points to the final grade

What to include in the two-page proposal:

Title (should be similar to a title of a research article)

Section 1. Statement of the research question

Answer the following questions: what? why?

Section 2. Formulation of the hypothesis
Section 3. Description of the data to be analyzed
Section 4. Description of the ML method to be used
Section 5. Discussion of expected outcomes

During the final examination week (date to be determined):

⇒ Each student will present their project
⇒ Each presentation will be 15 min, including questions
⇒ A random number generator will select the presenters each day

Each student will be expected to receive a mark of 20 for this presentation unless their work is proven to have fundamental flaws that could have been avoided.

⇒ Each student will submit a 4-page final report describing the work done
⇒ The final report will be due on the day of the final presentations in December (TBD)
⇒ The report will be marked and contribute up to 30 points to the final grade

The final project report - please follow closely the following guidelines

Due date: the date of the final presentations in December

Page limit: 4 pages, including references

What to include in the report:

Title (should be similar to a title of a research article)

Abstract (should be similar to an abstract of a research article)

Section 1. Introduction – Answer the following questions: what? why? how?

Section 2. Hypothesis – state your hypothesis in a separate section so that it is absolutely clear.

Section 3. Describe your data – how is the data obtained? how expensive is it? Describe the accuracy of data and noise.

Section 4. Describe the machine learning methods used. Include a justification for the choice of the ML method used in this work.

Section 5. Results – present and discuss the results.

Section 6. Summary – briefly summarize what has been achieved.

Section 7. Further work – describe what needs to be done in order to make this work publishable.

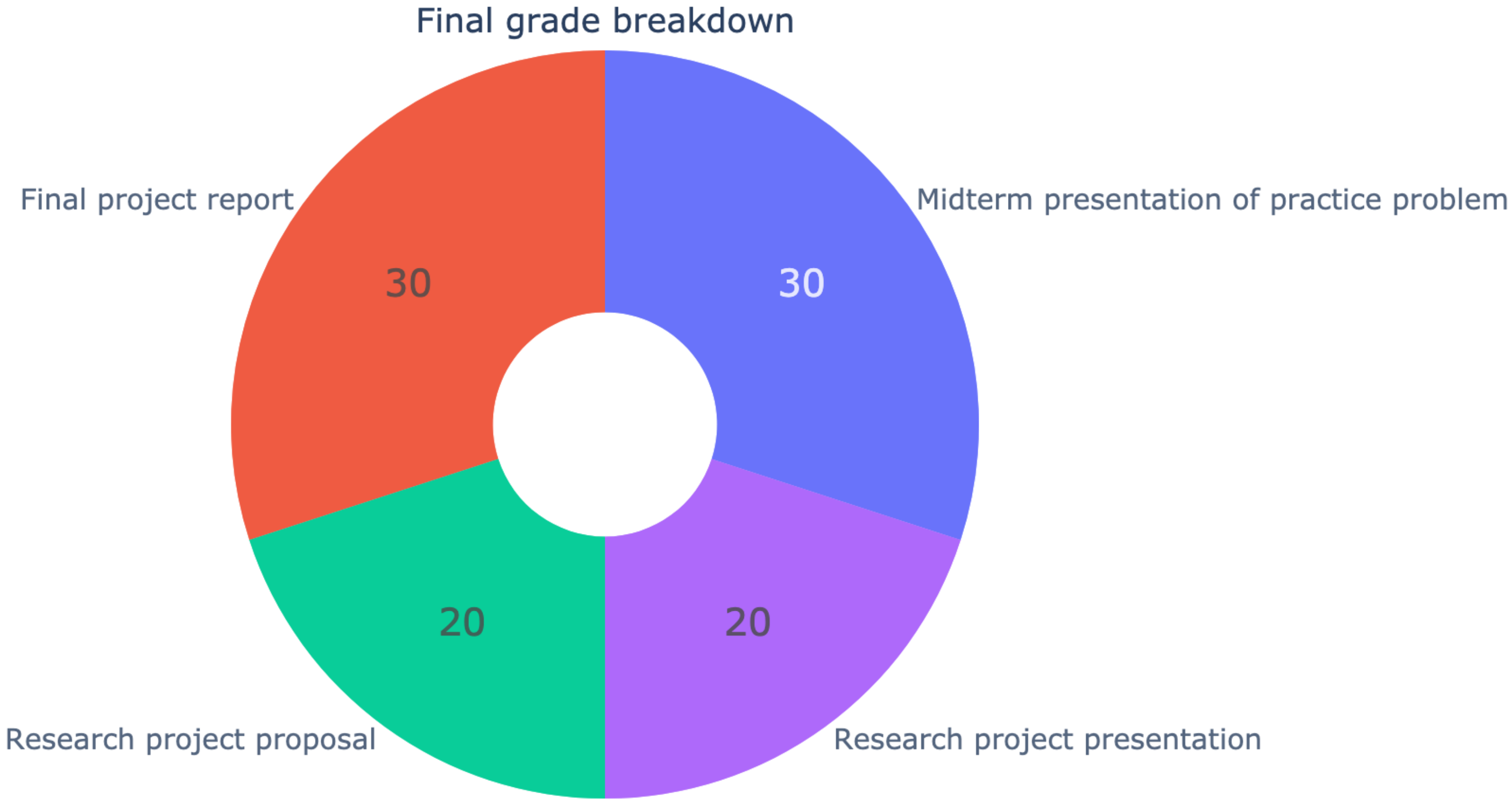
Section 8. References.

Summary of grading scheme:

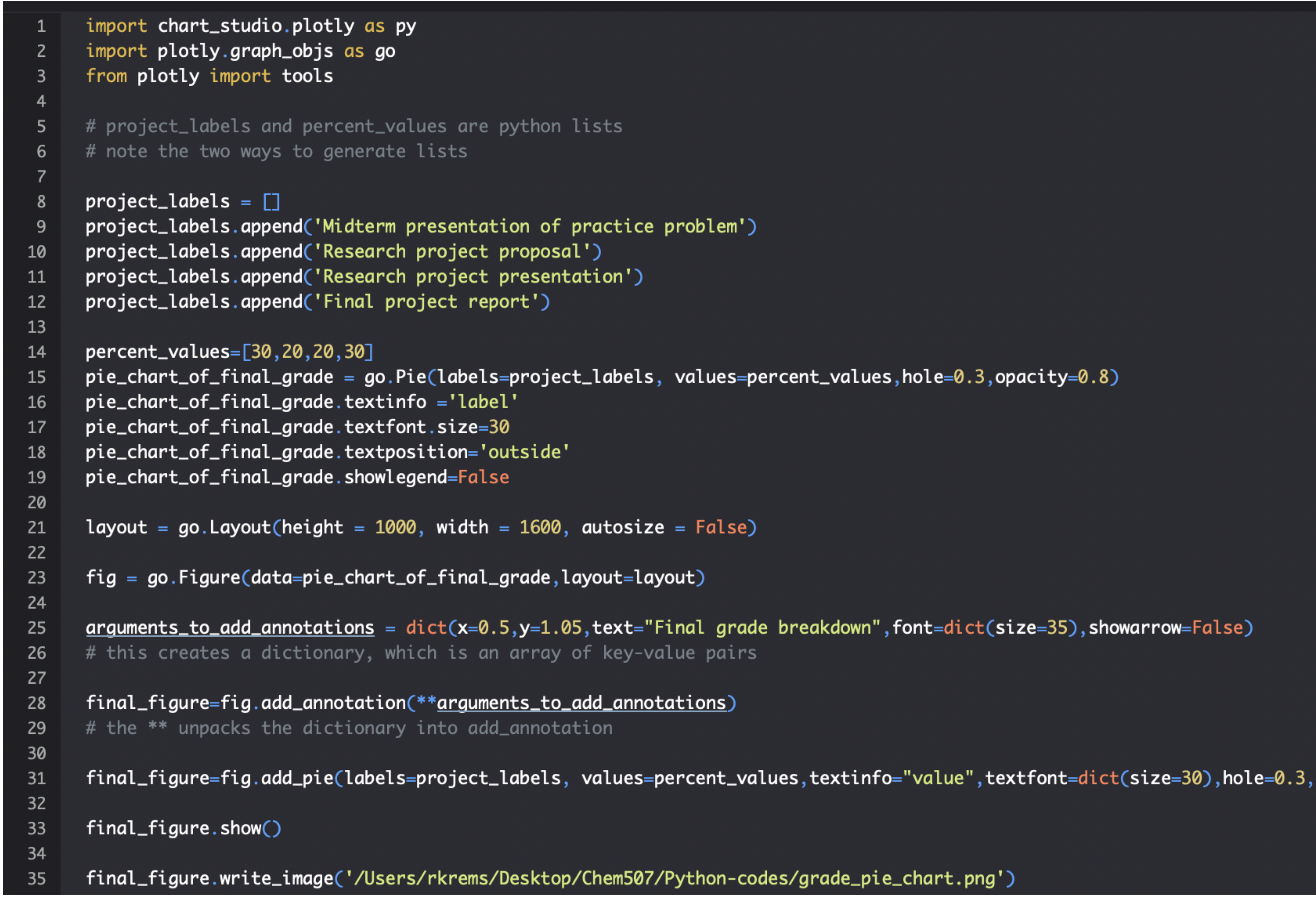
- Midterm presentation of the practice problem: 30
- Research project proposal: 20
- Research project presentation: 20
- Final project report: 30

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

Source file: pi-chart.py



```python
import chart_studio.plotly as py
import plotly.graph_objs as go
from plotly import tools

# project_labels and percent_values are python lists
# note the two ways to generate lists

project_labels = []
project_labels.append('Midterm presentation of practice problem')
project_labels.append('Research project proposal')
project_labels.append('Research project presentation')
project_labels.append('Final project report')

percent_values=[30,20,20,30]
pie_chart_of_final_grade = go.Pie(labels=project_labels, values=percent_values,hole=0.3,opacity=0.8)
pie_chart_of_final_grade.textinfo ='label'
pie_chart_of_final_grade.textfont.size=30
pie_chart_of_final_grade.textposition='outside'
pie_chart_of_final_grade.showlegend=False

layout = go.Layout(height = 1000, width = 1600, autosize = False)

fig = go.Figure(data=pie_chart_of_final_grade,layout=layout)

arguments_to_add_annotations = dict(x=0.5,y=1.05,text="Final grade breakdown",font=dict(size=35),showarrow=False)
# this creates a dictionary, which is an array of key-value pairs

final_figure=fig.add_annotation(**arguments_to_add_annotations)
# the ** unpacks the dictionary into add_annotation

final_figure=fig.add_pie(labels=project_labels, values=percent_values,textinfo="value",textfont=dict(size=30),hole=0.3,

final_figure.show()

final_figure.write_image('/Users/rkrems/Desktop/Chem507/Python-codes/grade_pie_chart.png')
```

Python exercise 1

Our department's website `http://chem.ubc.ca/faculty` categorizes faculty by area of chemistry.

Using the plotly example from previous page as a starting point, write code to build a pi-chart showing the distribution of faculty by research area.

I will include exercises – such as this one – throughout the course. These exercises will build on my codes (downloadable from Canvas) and will equip you with useful python skills required for more complex data analysis problems.

Summary of deadlines:

- Practice problem codes must be developed by: October 18
- Research project proposal (2 pages): October 31
- Final project presentation: December (TBD)
- Final project report (4 pages): December (TBD)

Python-related information:

If you are new to python, don't worry: python is easy to learn.

Start here: https://docs.python.org/3/tutorial/

We will be using TensorFlow to build ML python codes in this course:

https://www.tensorflow.org/

We will be using python 3. My version is python 3.7.7. TesnorFlow does not yet work with python version > 3.8. So make sure you have python $3.7 <$ version < 3.8.

List of discussion topics:

- Data Visualization

  learn to build powerful visualizations with Pandas + Seaborn

- 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

- Quantum Machine Learning

  Current state of the art in quantum computing, How quantum computers can be used to build kernels for machine learning models

Lectures/discussions

What we will aim to do:

⇒ Discuss theory behind ML

⇒ Discuss useful mathematics/statistics concepts

⇒ Discuss examples of ML applications in Chemistry

⇒ 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