E301 - Statistics and Data Analysis for Cancer Engineers

Dates: Nov 10 – Dec 10

Instructor: Aleksei Pakharev, PhD, pakhara@mskcc.org

TA: Joshua Lau, lauj2@mskcc.org

Prerequisites: Probability I, Python

Grading policy: Pass/Fail, details below

Course Summary

This course aims to give the students an overview of the fundamental concepts of data analysis and statistics – how to extract meaningful signal from real-world data, quantify its uncertainty, and use the findings to make further predictions. We will first look at the classical statistical ideas and terminology, then focus on specific data analysis tools and techniques applied now. The ultimate goal of the course is to give the students hands-on experience of making meaningful data analysis back-to-back.

Learning Objectives

After completing the course, you should be able to:

- Investigate data using available software to your advantage
- Identify robust conclusions and argue for them
- Know how to study new statistical concepts and data analysis models

Class Structure

Each class will consist of a programming session in a Jupyter Python notebook guided by the lecturer. It is of utmost importance that you bring a capable personal device to each class.

Before the Course

Please make sure you have a personal device with functional Python environment and working Jupyter. If you don't have it installed yet, my suggestion is to install conda-forge first.

a. If your device is running Windows:
    1. Submit a ticket to the Research Computing (RC) team to request admin privileges to install conda-forge.
    2. Download the installer: latest conda-forge release.
    3. Follow the official instructions: conda-forge/miniforge README.md.
b. If your device is running MacOS:
    1. Setup Homebrew: official homebrew page.
    2. Follow the official instructions: conda-forge/miniforge README.md.
c. If your device is running any other Unix-like distribution:
    1. Follow the official instructions: conda-forge/miniforge README.md.

After installing conda-forge, create an environment following the instructions: getting started with conda. In this environment, install JupyterLab following Option 1 of this guide: How to use Jupyter notebooks in a conda environment?. If at any point of the process you need help with the installation, send me an email.

Finally, if you still have time, please brush up the basics of probability (think Bayes formula).

Materials

There is no assigned textbook for the course. All lecture slides will be available online. If you want to read about the topics we'll cover from a different angle or in more depth, you may find the following resources useful:

1. For statistics questions: Introduction to Probability and Statistics at MIT
2. For various model questions: Detailed explanations of scikit-learn methods

Assignments

There will be a single data analysis assignment (due Dec 10 at 1:30pm) to help you solidify your knowledge of data analysis methods. You will have the opportunity to ask for help and feedback on the project throughout the course duration.

Grading

Class attendance is expected. The pass/fail grade will be given based on the data analysis project.

Course Topics

Week 1 (Nov 10 and Nov 12): Essential Statistics

- Distributions and samples
- Estimators
- Frequentist and Bayesian perspectives
- Statistical tests and p-value
- Bootstrapping

Week 2 (Nov 17 and Nov 19): Regression and Classification Models

- Maximum Likelihood Estimation (MLE)
- Linear regression
- Logistic regression
- Support Vector Machine (SVM)

Week 3 (Nov 24): Nearest Neighbors and Clustering Methods

- Distance in high-dimensional spaces
- k nearest neighbors graph (kNN)
- k-means and mixture model clustering
- HDBSCAN

Week 4 (Dec 1 and Dec 3): Dimensionality Reduction Methods

- Principal Component Analysis (PCA)
- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Diffusion Component Analysis (DCA)
- Uniform Manifold Approximation and Projection (UMAP)

Week 5 (Dec 8 and Dec 10): Real-world Data Practice