Lectures: Mon & Wed at 9:00am-10:20am in ANB 105
Instructor: Konstantin (Kostia) Zuev
Office: Annenberg 114
Email: kostia@caltech.edu (please include “158” in the subject line)
Office Hour: Wed 1pm-2pm, or by appointment (please, send an email to schedule)
Head TA: Harsh Gandhi (hgandhi@caltech.edu)
TAs and OHs: https://piazza.com/caltech/winter2024/ids158/staff
You are welcome to attend as many office hours as you like.

Course Description
The main goal of the course is to provide an introduction to the central concepts and core methods of statistical learning, an interdisciplinary field at the intersection of statistics, machine learning, information and data sciences. The course focuses on the mathematics and statistics of methods developed for learning from data. Students will learn what methods for statistical learning exist, how and why they work (not just what tasks they solve and in what built-in functions they are implemented), and when they are expected to perform poorly. The course is oriented for upper level undergraduate students in IDS, ACM, and CS and graduate students from other disciplines who have sufficient background in probability, statistics, and linear algebra. Topics covered include statistical decision theory, regression and classification problems, classical linear regression, subset selection, shrinkage methods, ridge regression and lasso, cross-validation, logistic regression, linear and quadratic discriminant analysis, support-vector machines, tree-based methods, bagging, and random forests.

Prerequisites
A solid understanding of linear algebra, probability, and statistics is required. Familiarity with MATLAB is desired.

- ACM 104 “Applied Linear Algebra”
- ACM 116 “Introduction to Probability Models”
- IDS 157 “Statistical Inference”

Textbooks
- Comprehensive Lecture Notes (will be posted on Piazza after each lecture)
- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning
- G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning
  https://www.statlearning.com/

Course Plan
The following is a tentative outline of the topics to be covered this term.

- Statistical Learning, Statistical Decision Theory, Nearest-Neighbor Methods, the Curse of Dimensionality
- The Bayes Classifier, the Bias-Variance Trade-Off, Linear Algebra of Linear Regression
- Geometry of Ordinary Least Squares, the Gauss-Markov Theorem, Hypotheses Testing, Multiple Outputs
- Model Selection and Model Assessment, the Validation Set Approach, Cross-Validation, Subset Selection
- Shrinkage Methods, Ridge Regression, LASSO
- Logistic Regression, Linear and Quadratic Discriminant Analysis
- Maximal Margin Classifier, Soft Margin Classifier, Support Vector Machine, Mercer’s Theorem
- Tree-Based Methods, Regression Trees, Classification Trees
- Bagging, Out-of-Bag Error Estimation, Random Forests
Grading
Your final grade will be based on your total score. Your total score is the sum of the scores you will get for Problem Sets. You can increase your total score by up to 5% if you participate actively in Piazza discussions in the Q&A section. Every answer submitted before TAs or instructor answer, which is later endorsed as “good answer” by TAs or instructor, gets 1% of the total score. There are no fixed thresholds for grades, but if your total score is 90% (80%, 70%, 60%), then you are guaranteed at least “A” (“B”, “C”, “D”).

Problem Sets
There will be five Problem Sets. Each problem set will contribute 20% toward your total score. Problems (and solutions) will be posted on Piazza. For assignment and due dates see “Important Dates” below. Late submissions will not be accepted. Submitting wrong files or files in a wrong format is considered as a late submission. Extensions may be granted for academic, personal, or medical reasons. For extensions, please email the Head TA.

Collaboration Policy
A detailed collaboration policy is given on the course website at:
In general, collaboration is encouraged. The goal is to help each other and learn together! If you get stuck with a homework problem, I encourage you to discuss it with other students (offline or online on Piazza). But remember that you will have to prepare and submit your solution by yourself.

Important Dates (All times are Pacific Times)

<table>
<thead>
<tr>
<th>Available</th>
<th>Due</th>
<th>Based on (roughly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Set 1</td>
<td>1pm Wed, Jan 10</td>
<td>9pm Wed, Jan 24</td>
</tr>
<tr>
<td>Problem Set 2</td>
<td>1pm Wed, Jan 24</td>
<td>9pm Mon, Feb 5</td>
</tr>
<tr>
<td>Problem Set 3</td>
<td>1pm Mon, Feb 5</td>
<td>9pm Wed, Feb 14</td>
</tr>
<tr>
<td>Problem Set 4</td>
<td>1pm Wed, Feb 14</td>
<td>9pm Wed, Feb 28</td>
</tr>
<tr>
<td>Problem Set 5</td>
<td>1pm Wed, Feb 28</td>
<td>9pm Fri, March 8</td>
</tr>
<tr>
<td>Head TA Lecture</td>
<td>9:00am Wed, March 6</td>
<td>NA</td>
</tr>
</tbody>
</table>

Websites
- Course website:
  http://www.its.caltech.edu/~zuev/teaching/2024Winter/IDS158.html
- Lecture notes, further reading materials, problem sets, data sets, solutions, announcements, and class discussions will be managed via Piazza, which is designed such that you can get a quick help from your classmates, TA(s), and instructor. Instead of emailing questions to the teaching staff, I encourage you to post your questions on Piazza because a) you will get the answers faster and b) your classmates may also benefit from seeing the answers to your questions.
  https://piazza.com/caltech/winter2024/ids158/home
- Problem sets and exams will be graded via Gradescope.
  - If you are a registered student, you will be enrolled on Gradescope by the end of the 1st week of classes, and you will receive a notification from Gradescope about your enrollment.
    - Please make sure that the email that you use on Gradescope is your official Caltech email.
  - If you are a registered student, but have not been enrolled on Gradescope by the end of the 1st week of classes, please email the Head TA as soon as possible and ask to enroll you to Gradescope. Your absence on Gradescope means that, according to my records, you are not registered.
  - If you want just to audit the course, it is fine, you will have access to Piazza and all course materials there (please email me and I will enroll you on Piazza), but you will not have access to Gradescope and your submissions will not be graded. If you audit the course this year, you should not register for the course in the future.

To submit your solution via Gradescope, your need to create a single PDF (not images) that contains the whole solution (for example, by scanning your solution), and then upload it to Gradescope. Here are some useful links:
- Scanning on a mobile device: https://help.gradescope.com/article/0chl25eed3
- Submitting an assignment: https://help.gradescope.com/article/cbpppzio9

Should you have any questions regarding Gradescope, please ask on Piazza: we will have many experts there.
Suggested Study Process

To get the most out of IDS 158, here is my suggestion on the study process:

- Attend Lectures, focus on understanding the big picture of what is going on.
- Review Lecture Notes (ideally on the same day they are released), make sure that everything is clear.
- If something is not clear, ask on Piazza, and help your classmates by answering their questions.
- After each Lecture, very briefly summarize my notes in Your Own Notes, extract the essence.
- Start working on each Problems Set on the same day it is released (or as soon as possible after that).
- Aim at finishing each Problem Set at least 1 day before it is due.
- If you get stuck with a problem, ask for hints on Piazza.

Keep in Mind

My goal is to help you understand and learn the material. Understanding is a creative and time- and effort-consuming process. If you don’t understand something, please talk to me. If you are struggling with balancing the workload please talk to me. If you have any concerns, please talk to me. Keep in mind that I am here to help.

Honor Code

“No member of the Caltech community shall take unfair advantage of any other member of the Caltech community.”