DSCI 102 Foundations of Data Science I, 4 Credits

University of Oregon, Data Science Program

Course Description

This course expands upon critical concepts and skills introduced in DSCI 101. Students apply increasingly sophisticated computational and statistical techniques to data across numerous domains.

Topics include the normal distribution, confidence intervals, regression, and classifiers. Ethical concerns resulting from use of the techniques in this course will be addressed.

Prerequisites: DSCI 101.

Major Requirement Required for all data science students

Format Two 1.5 hours classroom lectures / demos / discussions, one 1.0 hour lab.

Lecture: Demos, class questions, including iClicker participation, small group discussions, presentations.

Demos and slides will be posted on canvas before the lecture.

Lab: Quizzes will be given in labs, and you must take the quiz at your assigned lab time. Please attend your assigned lab.

Course Materials

Textbook: Computational and Inferential Thinking: The Foundations of Data Science, a free online textbook that includes Jupyter notebooks and public data sets for its examples.

Software: Python 3 with data science modules from an Anaconda installation such as Numpy and Matplotlib, as well as the datascience and otter-grader modules. With the Chrome or Firefox browser, you can use the UO Talapas super-computer and avoid the installing any software on your computer; it's also the only way you could use a Chromebook or similar machine with the course material.

Coaching Model: Instructors will play a role similar to a coach for athletes. We'll give you some information and feedback, but your active, engaged participation will lead to genuine learning. As with a sport, practice, practice, practice and your understanding will improve.

Learning Outcomes
Upon successful completion of this course each student should be able to:

1. Implement computational techniques to perform statistical analyses across large datasets
2. Characterize the normal distribution, including the mean, median, and standard deviation
3. Define confidence intervals and employ the square root law to calculate required sample size for specified confidence intervals.
4. Make quantitative predictions using regression techniques including calculating linear regression line equations and numerical least-squares minimization
5. Apply nearest-neighbor classifiers to predict categorical and quantitative values
6. Be able to enumerate ethical concerns resulting from use of the techniques in this course.

Assessment

Homework will be assigned regularly, due at start of class one week later (unless otherwise noted)

Labs sections meet weekly to work lab assignments and homework assignments. Attend all labs.

Quizzes every 2nd week for a total of 4 quizzes, specific notes will be available to you during the quiz.

Projects: Complete 2 in the second half of the term.

The final course grade will use these weighted components (or similar).

- Homework 14% (7 x 2% each)
- Lab Assignments 28% (7 x 4% each)
- Course Project 1 14% (1 x 14%)
- Quizzes 24% (4 x 6%)
- Final Quiz 20% (1 x 20%)

The course will be graded on the following scale

A+ = 96.67-100%  A = 93.34-96.66%  A- = 90.0-93.33%
B+ = 86.67-89.99%  B = 83.34-86.66%  B- = 80.0-83.33%
C+ = 76.67-79.99%  C = 73.34-76.66%  C- = 70.0-73.33%
D+ = 66.67-69.99%  D = 63.34-66.66%  D- = 60.0-63.33%
F = 0-59.99%
### Student Expectations

**Attend Class, Labs**
Regularly attend both classes and labs.
Actively engage in classroom discussions.

**Read**
Read text chapters before coming to class.
In class you will put that knowledge to work and discuss it.

### Course Outline  
*Subject to adjustments*

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Read CIT</th>
<th>Work or Quiz</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Review: Tables, arrays, p_value, hypotheses</td>
<td>Demo</td>
<td>Hw08, Lab00 Confidence Intervals,</td>
</tr>
<tr>
<td></td>
<td>Visualization with plots</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Confidence intervals: percentiles, Bootstrap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Center and Spread: Mean, standard deviation</td>
<td>13.0 – 13.4</td>
<td>Quiz 1 Hw09, Lab08 Bootstrap, Resample, Central Limit Theorem, Normal distribution, Variance of sample means</td>
</tr>
<tr>
<td></td>
<td>Normal distribution, Sample means, Central Limit Theorem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Design experiments, choose sample size</td>
<td>14,14.1,14.2, 14.3-5</td>
<td>Hw10, Lab19 Linear Regression</td>
</tr>
<tr>
<td></td>
<td>Correlation, standard (&quot;z&quot;) units</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Linear Regression</td>
<td>14.6</td>
<td>Quiz 2 Hw11, Lab10 Regression</td>
</tr>
<tr>
<td></td>
<td>Predict: Least Squares</td>
<td>15, 15.1</td>
<td>inference, Residuals, residual intervals</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>15.2, 15.3, 15.4</td>
<td>Hw12 Classification: Nearest Neighbor, k nearest neighbors</td>
</tr>
<tr>
<td>6</td>
<td>Residuals: Visual, numeric diagnostics</td>
<td>15.5, 15.6</td>
<td>Quiz3 Project 2 start – Cardiovascular disease studies</td>
</tr>
<tr>
<td></td>
<td>Regression: inference, slope of reg. line, bootstrap samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Classification, Training and Test sets</td>
<td>17, 17.3, 17.4</td>
<td>Project 2 finish</td>
</tr>
<tr>
<td>8</td>
<td>Classifiers: Decisions, Regression recap</td>
<td>18, 18.1</td>
<td>Quiz4 Project 3 Movie Classification or an independent project</td>
</tr>
<tr>
<td></td>
<td>Classify using Nearest neighbors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Decision trees, Bayes estimation</td>
<td>18.2</td>
<td>Project3: Examine ethics of your project</td>
</tr>
<tr>
<td></td>
<td>Gradient descent, Bias / Variance tradeoffs</td>
<td>r2d3 site (bias/variance tradeoff)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Ethics</td>
<td>Guest Lecture</td>
<td>Project 3 complete</td>
</tr>
<tr>
<td></td>
<td>Review</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Plan to commit time to work through labs, homework, projects and readings.

Learn by doing–and plan to get help when stuck on a topic.

Goal: master ideas and techniques of data science, think of ways to inform others without intimidating.

We encourage collaborative efforts, but collaboration means sharing ideas, techniques, but not copying code for assignments directly.

*Of course, quizzes must never involve collaboration.*