Appendices

Where to get Data

The data sets used in today's presentation are in the Data folder.

The best source for data to use in an education context is the UCI Machine Learning repository: https://archive.ics.uci.edu/ml. There are data sets here that have been used for training and evaluating learning systems since the 1990s.

The most popular data sets are good ones to try first. Many of these are available directly in sklearn so you don't have to read the data yourself.

Iris, Wine, and Breast Cancer Wisconsin

To find others, click on "View all Data Sets" and then click "Classification" and "Numerical" on the left hand side. Here's a direct link:

https://archive.ics.uci.edu/ml/datasets.php?format=&task=cla&att=num&area=&numAtt=&numIns=&type=&sort=nameUp&view=table

When you click on a data set, you usually get some information about it. Often this is a repeat of what is in the **.names** file for the data (see below). You click on the "Data Folder" link to find the raw data.

The format of the data varies, but usually there is a **.data** file and a **.names** file. The **.data** file is the data (usually in CSV format) and the **.names** file is a description of the data including what each column represents.

Reading and Graphing Data

In the Readings folder, you will find "Selections from Data Science from Scratch". This reading contains chapters 2, 3, and 9 of the textbook *Data Science from Scratch*. Chapter 3 shows you how to use the Python matplotlib library to graph data, and Chapter 9 shows you how to read text and csv files.