Homework 1 - Decision Stumps

# Instructions

Most homework assignments for this class will consist of two tasks:

1. **Reading/Video**: This will be an article or video about a topic in machine learning related to what we have learned so far in the class. After you read the article or watch the video, we will ask you to answer some comprehension questions (typed or handwritten responses are fine), which will relate to the class content, coding activity, and to the reading/video. We will start off class next week by discussing these comprehension questions (remind us if we forget!)
2. **Coding**: This consists of completing the coding file you were working on in class today. This does not include the bonus challenge, although you are welcome to complete that if you want!

# Reading

Please read the below article. The first part of it (up to “The Best Split”) relates to material we covered in class today. The second part (after that) relates to material we will cover next week. And some parts scattered through the whole piece (about adding multiple dimensions to the data) relate to material we will cover two weeks from now. However, this article should give you a good review of what we covered and overview of what is to come.

A Visual Introduction to Machine Learning, <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

## Comprehension Questions

1. The article has what it calls a **scatterplot matrix** to show the relations between each pair of dimensions. The way to interpret that is as follows:
   1. For each graph (each square with green and blue datapoints), determine the y axis by looking left from the graph until you see a label, and determine the x axis by looking down until you see a label.
   2. The graph is a scatterplot relating those dimensions (i.e. each point is a single house, where its (x,y) position determines the values it has for those dimensions, and the color indicates what city it is in).

Now, remember what they showed earlier about boxes separating the New York data from the San Francisco data. This is a way of understanding how **predictive** particular dimensions are at separating the data. If it is easy to draw the boxes, that means that if you are given those two dimensions about a house, it is easy to predict whether it is in New York or San Francisco. On the other hand, if it is hard to draw those boxes (because the New York and San Francisco dots are all mixed up on the scatterplot), those two dimensions are not very predictive.

Looking at the **scatterplot matrix**, which two dimensions seem the most **predictive** about whether a house is in New York or San Francisco? Which seem the least **predictive**? If you were building a machine learning classifier to determine whether a house is in New York or San Francisco, which dimensions (information) would you want to have about the house to provide the best prediction?

1. In class, we discussed one setting that decision stumps could be good at -- predicting cold or flu based on temperature. In the article, they discuss another -- predicting where a house is located (New York or San Francisco) based on elevation, year built, price, etc. What are other settings where decision stumps/trees may be useful at classifying information? Write your ideas below.
2. In class and in the coding activity, we determine the best split by trying out every possible splitting point that yields two lists. However, this may take a long time if we have a lot of data (like multiple million data points). What are faster ways to find split points that may not be the best, but are “good enough”? Write down your ideas, including how you would find the splitting point and how you would determine that it is “good enough.”