WEEK 3

CLASS 5: DECISION TREES

HOMEWORK:

- Read the "Wisdom of the crowds" section from MLWave's post on Human Ensemble Learning.
- **Optional:**
  1. Read the abstract from Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?,
  2. Kaggle CTO Ben Hamner's comment about the paper, paying attention to the mentions of "Random Forests".

RESOURCES:

- scikit-learn's documentation on decision trees includes a nice overview of trees as well as tips for proper usage.
- For a more thorough introduction to decision trees, read section 4.3 (23 pages) of Introduction to Data Mining. (Chapter 4 is available as a free download.)
- This paper, The Science of Singing Along, contains a neat regression tree for predicting the percentage of an audience at a music venue that will sing along to a pop song.
- If you want to go deep into the different decision tree algorithms, this slide deck contains A Brief History of Classification and Regression Trees.
- **Installing GraphViz (optional):** * Mac: Download and install PKG file * Windows: Download and install MSI file, and then add GraphViz to your path:
  - Go to Control Panel, System, Advanced System Settings, Environment Variables
  - Under system variables, edit "Path" to include the path to the "bin" folder, such as: C:\Program Files (x86)\Graphviz2.38\bin

CLASS 5: ENSEMBLING

• Finish decision trees lesson

RESOURCES:

- scikit-learn's documentation on ensemble methods covers both "averaging methods"
- (such as bagging and Random Forests) as well as "boosting methods" (such as AdaBoost and Gradient Tree Boosting).
- For an intuitive explanation of Random Forests, read Edwin Chen's answer to How do random forests work in layman's terms?
- MLWave's Kaggle Ensembling Guide is very thorough and shows the many different ways that ensembling can take place.
- Browse the excellent solution paper from the winner of Kaggle's CrowdFlower competition for an example of the work and insight required to win a Kaggle competition.
- Interpretable vs Powerful Predictive Models: Why We Need Them Both is a short post on how the tactics useful in a Kaggle competition are not always useful in the real world.

CLASS 6: ADVANCED SCIKIT-LEARN AND CLUSTERING

HOMEWORK:
Optional: Read this classic paper, which may help you to connect many of the topics that we have studied throughout the course: A Few Useful Things to Know about Machine Learning.

SCIKIT-LEARN RESOURCES:

- Here is a longer example of feature scaling in scikit-learn, with additional discussion of the types of scaling you can use.
- Practical Data Science in Python is a long and well-written notebook that includes the use of scikit-learn's Pipeline.
- scikit-learn has an incredibly active mailing list that is often much more useful than Stack Overflow for researching a particular function.