MSDS 7331 – Machine Learning I

Course Overview and Schedule

Please find the schedule for the course below. Each week is marked with the section titles for videos, along with the number of questions. Graded assignments are marked for the week they are due. The key for the outline is as follows:

V	Video with slides, length is also given
S	Screen capture, typically of coding examples, length is also given
LB	Light Board or white board derivations or examples, length is also given
Y	YouTube video (of myself or guest instructor)
Q	Multiple Choice or Short Answer Question, Number of questions is also given
Q-e	Essay style question
Q-u	Upload question

1 Introduction to Data Mining in Python
1.3 Why Mine Data (V8:29 and 7 Qs)
1.4 Types of Data (V1:35, V3:22, 7 Qs, V2:28, 4Qs)
1.5 Representing Features in Python Pandas (S35:18)
1.6 Feature Manipulation (S13:59)
Total: 66 minutes
Live: Introductions and Numpy through Jupyter Notebook

2 Data Exploration in Python
2.3 Types of Data Summary Plots (V2:35, V2:08, V2:03, V2:12, V1:57)
2.4 Plotting Made Easy (S28:21)
2.5 PCA/LDA for Continuous Variables (V11:47, S20:12)
Total: 70 minutes
Live: Example Student Assignment

3 Project Work Week
Live: Individual Team Meetings
Project One Due Sunday Following Class

4 From Linear Regression to Classification
4.3 Linear Regression (LB5:23, LB8:08, LB7:52)
4.4 Gradient Based Optimization (LB8:37)
4.5 A New Objective: Classification (LB6:36, LB4:44)
4.6 Logistic Regression Optimization Setup (LB4:39)
Total: 45 minutes
Live: *—First Live Session Assignment—*

5 Logistic Regression and Support Vector Machines
5.3 Optimizing Two Class Logistic Regression (LB18:13)
5.4 Multiclass Logistic regression (LB5:14)
5.5 Gradient Optimization: Stochastic, Mini-Batch, Batch (LB7:22)
5.6 Support Vector Machines (LB15:14, LB14:43)
5.7 Kernel Tricks (LB12:26)
Total: 66 minutes
Live: SVM, LR, and SGD Notebook
Mini Project Due, SVM and LR

6 Decision Trees
6.3 What is a Decision Tree (V2:45, 2Qs, V1:40, 1Q)
6.4 Building a Decision Tree (V2:29, 1Q, V2:03, 1Q, V5:55, 1Q)
6.5 Criteria for Determining the Best Split (V3:36, 1Q, V3:04, 1Q, V3:11, V3:13, 2Q, V2:35)
6.6 Decision Boundaries and Stopping Criteria (V2:56, 1Q, V0:59, 1Q, V3:05, 1Q)
6.7 Decision Tree Generalization (V3:14, V6:22, 2Q, V1:43)
Total: 50 minutes
Live: *—Second Live Session Assignment—*

7 Model Evaluation and Ensemble Classification
7.3 Model Evaluation (V2:28, 1Q-e, V2:35, 1Q, V2:34, 2Q-e)
7.4 Methods for Estimating Performance (V3:37, V1:34, 1Q, V2:10, 1Q, V1:29, 1Q, V4:53)
7.5 Model Comparison (LB2:48, LB5:21, LB4:38, LB2:18) 7.6 Ensemble Classifiers (V4:22, 1Q, V7:52, 1Q, V5:48, 2Q)
Total: 55 minutes
Live: Grand Poobah Classification Notebook

8 Alternative Classifiers
8.3 Nearest Neighbors Classification (V5:21, 1Q, V3:19, 2Q-e, Y21:44, V3:44)
8.4 Nearest Neighbor Regression (V4:58, 1Q)
8.5 KD Trees (V2:36, 1Q, V2:51, 1Q, V5:18)
8.6 Naive Bayes (V1:46, 1Q, V4:53, 1Q, V4:21, 1Q, V1:45)
Total: 60 minutes
Live: *—Third Live Session Assignment—*

9 Project Work Week
Live: Individual Team Meetings
Project Two Due Sunday Following Class

10 Basic Clustering Methods
10.3 What is clustering (V2:18, 2Q, V4:37, 2Q)
10.4 K-means Clustering (V1:13, 1Q, V5:07, 1Q-u, V1:58, V4:27, 2Q, V3:22, V2:32, 1Q)
10.5 Hierarchical Agglomerative Clustering (HAC) (V6:40, 1Q, V2:26)
10.6 Density Based Clustering (V2:301Q, V4:36)
10.7 Clustering Examples (S23:47)
10.8 Clustering Validity (V4:46, V5:06)
Total: 50 minutes
Live: *—Fourth Live Session Assignment—*

11 More Clustering Methods
11.3 Gaussian Mixtures (V2:01, 2Q, V1:28, 1Q, V3:45, 1Q, V3:43)
11.4 Graph Clustering (V3:37, 1Q-u, V2:23, V2:12, V4:14)
11.5 Spectral Clustering (V3:15, V4:53)
Total: 30 minutes
Live: Spectral and GMM Clustering Notebook

12 Association Rule Mining
12.3 What is Association Rule Mining (V4:11, 1Q, V1:45, 1Q, V1:01)
12.4 Frequent Itemset Generation (V2:02, V0:52, V1:27, V1:50, 1Q, V1:29, V3:05, V2:51, V1:35)
12.5 Frequent Pattern Trees (V3:46, 1Q, S15:31, 1Q, S5:11)
12.6 Rules Generation (V3:52, 1Q, V1:35)
12.7 Maximal and Closed Sets (V4:11)
12.8 Evaluating Association Rules (V4:40, 1Q-e, V3:37, V2:42)
Total: 66 minutes
Live: —Fifth Live Session Assignment—

Recommender Systems (Optional, can be over Thanksgiving)
13.3 Intro to collaborative filtering (V1:33, V3:11, V2:04, V2:41, V1:21)
13.4 SFrames in Dato (Y32:45)
More In-depth Explanation of Collaborative Filtering in GraphLab Page:
https://www.youtube.com/watch?v=I-xPNMCqW2o&feature=youtu.be (V49:00)
Total: 92 minutes
Live: Recommender Systems Notebook

14 Data Parallel and Out of Core
14.3 Data Parallel and Grid Searching
Live: Individual Team Meetings
Total: Two Hours
Final Project Due Sunday Following Class

Unit One: An Introduction to Data Mining in Python

Lecture Units:

1.3 Why Mine Data (V8:29 and 7 Qs)
1.4 Types of Data (V1:35, V3:22, 7 Qs, V2:28, 4Qs)
1.5 Representing Features in Python Pandas (S35:18)
1.6 Feature Manipulation (S13:59)
Total Video: 66 Minutes
Jupyter notebook(s):

- Look at notebook 1 on the course github page.
- Look at notebook 2 on the course github page. (up to visualization)

Unit Objectives

- Understand examples of data mining data sets
- Introduction to attribute types
- Familiarize and install the iPython/Jupyter ecosystem
- Working with data in Pandas: Feature normalization, One hot encoding, Imputation

Optional Reading

- Chapter I and Chapter II from Tan, Steinbach, and Kumar
- Pandas ecosystem introduction:
- 10-minute tour: http://vimeo.com/59324550
- http://pandas.pydata.org/pandas-docs/version/0.15.2/tutorials.html

Installations

- Python Installation: https://www.python.org/
    - You can use with Python 2 or Python 3 for this course.
    - Great package with most of the batteries included: Anaconda
        - When using anaconda, be sure to use the included “conda” or “pip” package manager
- Install Jupyter notebook on your work station: Jupyter Notebook (interactive python)
- http://jupyter.org
- this is included in the anaconda distribution
- Install Pandas:
- Pydata Pandas Eco-system
- this is included in the anaconda distribution

Unit Assignments

- Install all required packages on your development system (Linux, Mac or, if need be, Windows)
    - Linux and mac will be slightly better operating systems when working with the packages in this class. Although Windows will be fine for the majority of cases, some packages run more slowly or are not optimized for this OS.
- Have an idea about the dataset you would like to use during the semester
- During Live Session:
    - Introductions
    - Introduce course grading schema
- Dataset selection: have a brief explanation of your data, how you plan to preprocess it, and why it is useful for third parties
    - Pay attention and start to feel around for others to work with on teams during semester. Teams are due by next week!!
- Come with questions about the Pandas ecosystem and variable representations!
- In-class example: what is Numpy and why do we use it? This is example E00 on the course github page.

Dataset Selection for Semester

You will select (as a team) which dataset you wish to work on during the semester. You have a lot of free reign for the analysis, but the dataset must be sufficiently complex.

This means:

1. there must be 10 or more attributes for analyzing the data (both continuous and discrete),
2. there must be at least 30,000 records to classify in the data, and
3. there must be a good documentation of the dataset attributes.

Select a classification dataset from one of the following sources:

- **Public Record:** Chosen from a public repository. Some examples include the Dallas Crime dataset, UCI machine learning repository, and many others.
- **Kaggle:** This website hosts machine learning competitions with prizes to the best performers. You can choose one of the open and enterable competitions-as long as it meets the criteria for complexity outlined above.
    - **Note:** you need to sign up to use Kaggle. Many of the competitions require more than just machine learning to solve. For instance, it is usually up to you to get the dataset in a format that is readable for python. Additionally, you may need to extract features to use (e.g., they give you images and expect you to process the images and extract features).
    - This option is probably more work than the others but has more reward also.
    - As students, you usually qualify for getting additional prizes and/or job offers, so keep this kind of thing in mind for the future, even if you do not use it for class.
- **Your Own:** If you want to use data from your own research or data that you have collected, that's fine. It still must meet the above-mentioned criteria. Verify your dataset with the instructor before choosing this option!

Unit Two: Data Exploration in Python

Lecture Units:

2.3 Types of Data Summary Plots (V2:35, V2:08, V2:03, V2:12, V1:57)

2.4 Plotting Made Easy (S28:21)

2.5 PCA/LDA for Continuous Variables (V11:47, S20:12)

Total Video: 70 minutes

Jupyter notebook(s):

- Look at notebook 2 on the course github page (visualization).
- Look at notebook 3 on the course github page (dimensionality reduction).

Unit Objectives

- Understand how to run aggregated statistics in python Pandas
- Understand how to create and use visualizations with matplotlib and Pandas
- Exposure to other plotting interfaces like seaborne, MPLD3, and Plotly
- Employ methods for dimensionality reduction in scikit-learn (PCA, LDA, and randomized PCA)

Installation

- Install scipy, numpy, scikit-learn, and matplotlib (these are included with anaconda)

Optional Reading

- Appendix B and Chapter III from Tan, Steinbach, and Kumar
- Look at the initial steps in the CRISP-DM framework: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- Data Science Blog by Sebastian Raschka
  - Great intuitions for PCA and LDA in python
- Other visualization software (built to integrate with matplotlib in python):
  - Seaborn
  - MPLD3
  - Plotly (now open source!!)
- Running aggregated statistics (approximating the median) out-of-core:
  - http://www.silota.com/site-search-blog/approximate-median-computation-big-data/
- Other python visualization examples:
  - http://pandas.pydata.org/pandas-docs/stable/visualization.html
  - http://matplotlib.org/examples/index.html

Unit Assignments

- Make sure you have formed teams for the group lab assignments (groups can be size up to three)
- Look at first project work week assignment (it is due at the end of next week)
- Start analyzing your dataset using simple visualizations and statistics using Pandas and a visualization package like matplotlib l;

During live session:

- Go over a good example from the past of the first assignment
- Come with questions regarding the visualization and dimensionality notebooks

Unit Three: First Team Project Work Week, Data Exploration and Analysis

No lecture content: Use Jupyter and pandas ecosystem to explore chosen data sets

Turn in the first project work week assignment by Sunday at midnight following the live session

During Live Session (optional):

- Come with questions about using pandas on your datasets
- I will answer questions to each team individually in breakout sessions, please work on refining assignments as a team during class time
- Show some of what your team has been working on and any problems you have run into. Others will benefit greatly from these types of questions.

Due Sunday before midnight following the Unit 3 live session:

First Project Work Week Assignment

diabetic
28M
Pre-diabetic
79M
300 million people
United States
kaggle
The Home of Data Science
COMPETITIONS - ENERGY INDUSTRY SPECIALISTS
PREDICTIVE MODELING SOLUTIONS - TUTORIALS
KAGGLE IN CLASS - DATA SCIENCE JOBS BOARD

You are to perform analysis of a data set: exploring the statistical summaries of the features, visualizing the attributes, and making conclusions from the visualizations and analysis. Follow the CRISP-DM framework in your analysis (you are not performing all of the CRISP-DM outline, only the portions relevant to understanding and visualization). This report is worth 20% of the final grade.

Please upload a report (one per team) with all code used, visualizations, and text in a single document. The format of the document can be PDF, *.ipynb, or HTML. You can write the report in whatever format you like, but it is easiest to turn in the rendered Jupyter notebook.

A note on grading: This lab is mostly about visualizing and understanding your dataset. The largest share of the points is from how you interpret the visuals that you make. Making the visuals is not enough to satisfy each of the rubrics below—you should appropriately explain what the implications of the visualizations are. In other words, expect about 20% of the available points for visuals that have no substantive discussion.

Grading Rubric

Business Understanding **(10 points total).**

- Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). Describe how you would define and measure the outcomes from the dataset. That is, why is this data important and how do you know if you have mined useful knowledge from the dataset? How would you measure the effectiveness of a good prediction algorithm? Be specific.

Data Understanding **(80 points total)**

- **[10 points]** Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file.
- **[15 points]** Verify data quality: Explain any missing values, duplicate data, and outliers. Are those mistakes? How do you deal with these problems? Give justifications for your methods.
- **[10 points]** Visualize appropriate statistics (e.g., range, mode, mean, median, variance, counts) for a subset of attributes. Describe anything meaningful you found from this or if you found something potentially interesting. Note: You can also use data from other sources for comparison. Explain why the statistics run are meaningful.
- **[15 points]** Visualize the most interesting attributes (at least 5 attributes, your opinion on what is interesting). Important: Interpret the implications for each visualization. Explain for each attribute why the chosen visualization is appropriate.
- **[15 points]** Visualize relationships between attributes: Look at the attributes via scatter plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate. Explain any interesting relationships.
- **[10 points]** Identify and explain interesting relationships between features and the class you are trying to predict (i.e., relationships with variables and the target classification).
- **[5 points]** Are there other features that could be added to the data or created from existing features? Which ones?

Exceptional Work **(10 points total)**

- You have free reign to provide additional analyses.
- One idea: implement dimensionality reduction, then visualize and interpret the results.

Unit Four: From Linear Regression to Classification

Lecture Units:

4.3 Linear Regression (LB5:23, LB8:08, LB7:52)

4.4 Gradient Based Optimization (LB8:37)

4.5 A New Objective: Classification (LB6:36, LB4:44)

4.6 Logistic Regression Optimization Setup (LB4:39)

Total Video: 45.5 minutes

A Note to those feeling overwhelmed by the mathematics:
First, your book has good (but compact) explanations of regression and optimization techniques in appendices D and E, respectively. They cover more topics than I do in a short number of pages. It’s a great section for reviewing or referencing. We will cover a lot of Appendix E when we hit logistic regression (logistic regression is NOT covered in your book). At the end of the day, we are covering these mathematics because they are extremely useful for understanding how tweaking the different parameters of the algorithms changes generalization and the speed of training. Understanding the ramifications of these mathematics is what might separate a “Good” data scientist from an “Exceptional” data scientist.

There are also great examples of multivariate linear regression on Dr. Andrew Ng's iTunesU page (a Stanford course). https://itunes.apple.com/WebObjects/MZStore.woa/wa/viewPodcast?id=384233048#ls=1 He uses some of the same notation that I use. I have posted some of Dr. Ng's PDF's for reference, see below.

For logistic regression, you can also see some methods on the Stanford course, but I don't prefer their method of introduction. Instead, I use notation and intuition from Dr. Carlos Guestrin's course on Big Data (from Carnegie Mellon 7 University of Washington). I will be going over a modified version of Carlos's notation. We will focus only on gradient based optimization for logistic regression. Next week we will hit SVM’s and more logistic regression. The book covers SVM's fairly well, but I will make a few modifications so that their use is more intuitive.

Unit Objectives

- Review the derivation of the linear regression solution
- Understand the ramifications of optimizing for linear regression with a “large number of instances” versus “a large number of attributes”
- Understand how perceptron neural networks build upon the iterative linear regression solution
- Derive perceptron back-propagation (indirectly)
- Setup derivation of Logistic Regression and Support Vector Machines

Optional Reading

- Appendix D: Linear Regression, from Tan, Steinbach, and Kumar
- Appendix E: Optimization, from Tan, Steinbach, and Kumar
- Chapter V (5.4, Artificial Neural Networks), from Tan, Steinbach, and Kumar
- Andrew Ng’s Linear Regression Overview, CS229 Notes 1: Part I, Sections 1-3: https://www.dropbox.com/s/eriypxr001pfso6/cs229-notes1.pdf?dl=0

Unit Assignments

During Live Session:

- First live session assignment: Linear Regression
- You will complete an interactive Jupyter notebook assignment (working in teams if you prefer) during the live session
- You will work on a Jupyter notebook containing questions about linear regression and linear classification from the videos.
- An early version of the notebook will be sent out during the week for you to answer the questions before class starts.
- The instructor will then work out these preliminary questions during the first 15 minutes of class.
- After this introduction, another notebook will be made available for you to work on as a team. You will be asked to live code (mostly using Numpy) some of the equations we went over in the videos.
- The live session assignment is meant to reinforce the mathematics with programming so that the concepts become more concrete.
- You will complete the coding assignment and the questions and upload the updated python notebook at the end of live session.
- Ask questions to the instructor! This is meant to facilitate one-on-one time!!

Unit Five: Logistic Regression and Support Vector Machines

Lecture Units:

5.3 Optimizing Two Class Logistic Regression (LB18:13)
5.4 Multiclass Logistic regression (LB5:14)
5.5 Gradient Optimization: Stochastic, Mini-Batch, Batch (LB7:22)
5.6 Support Vector Machines (LB15:14, LB14:43)
5.7 Kernel Tricks (LB12:26)
Total Video: 73 minutes

Unit Objectives

- Derive optimization procedure for logistic regression and support vector machines
- Understand alternative optimization methods using iterative gradients: stochastic, batch, and mini-batch gradient descent
- Understand how kernels can be used to represent dot products in infinite dimensional space

Optional Reading

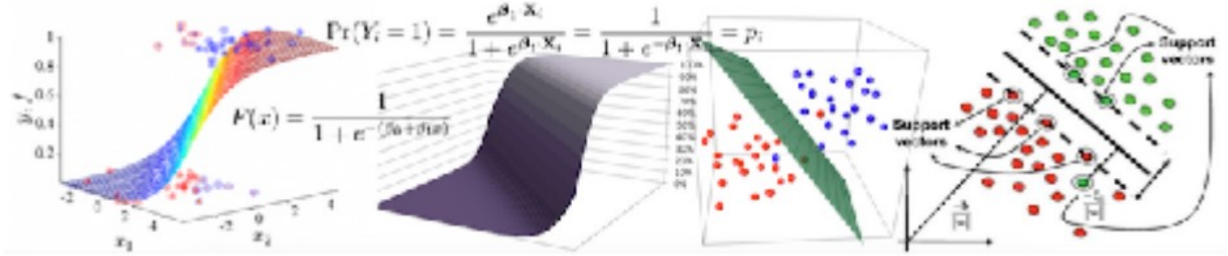
- Chapter V (5.5 SVMs), from Tan, Steinbach, and Kumar
- Andrew Ng's Linear Regression Overview, CS229 Notes 1: Part II, Sections 5 and 6 only
- https://www.dropbox.com/s/eriypxr001pfso6/cs229-notes1.pdf?dl=0
- Andrew Ng's SVM Overview, CS229 Notes 3, Section 1-8 https://www.dropbox.com/s/60pbyjos01mj3zj/cs229-notes3.pdf?dl=0
- Chapter Four (4.4, Logistic Regression only) of The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, and Jerome Friedman

Unit Assignments

- Finish the introductory assignment for using SVMs and Logistic regression with scikit-learn. It is due Sunday before midnight following live session.
- During live session:
- Live coding of SVMs and LR in scikit-learn with batch and stochastic gradient descent. The example notebook will be notebook 4 on the course github page.
- Review how to interpret the weights of logistic regression and linear support vector machines

Due Sunday at Midnight following Unit Five:

Mini-Project: SVM&LR Classification



$\Pr(Y_i = 1) = \frac{e^{\beta_1 X_i}}{1+e^{\beta_1 X_i}} = \frac{1}{1+e^{-\beta_1 X_i}} = p_i$
$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$
Support vectors
Support vectors

You are to perform predictive analysis (classification) upon a data set: model the dataset using methods we have discussed in class: logistic regression & support vector machines and making conclusions from the analysis. Follow the CRISP-DM framework in your analysis (you are not performing all of the CRISP-DM outline, only the portions relevant to the grading rubric outlined below). This report is worth 10% of the final grade. You may complete this assignment in teams of as many as three people.

Write a report covering all the steps of the project. The format of the document can be PDF, *.ipynb, or HTML. You can write the report in whatever format you like, but it is easiest to turn in the rendered Jupyter notebook. The results should be reproducible using your report. Please **carefully describe every assumption and every step in your report.**

A note on grading: A common mistake I see in this lab is not investigating different input parameters for each model. Try a number of parameter combinations and discuss how the model changed.

SVM and Logistic Regression Modeling

- **[50 points]** Create a logistic regression model and a support vector machine model for the classification task involved with your dataset. Assess how well each model performs (use 80/20 training/testing split for your data). **Adjust parameters of the models to make them more accurate**. If your dataset size requires the use of stochastic gradient descent, then linear kernel only is fine to use. That is, the SGDClassifier is fine to use for optimizing logistic regression and linear support vector machines. For many problems, SGD will be required in order to train the SVM model in a reasonable timeframe.
- **[10 points]** Discuss the advantages of each model for each classification task. Does one type of model offer superior performance over another in terms of prediction accuracy? In terms of training time or efficiency? Explain in detail.
- **[30 points]** Use the weights from logistic regression to interpret the importance of different features for the classification task. **Explain your interpretation in detail.** Why do you think some variables are more important?
- **[10 points]** Look at the chosen support vectors for the classification task. Do these provide any insight into the data? Explain. If you used stochastic gradient descent (and therefore did not explicitly solve for support vectors), try subsampling your data to train the SVC model—then analyze the support vectors from the subsampled dataset.

Unit Six: Decision Trees

Lecture Units:

6.3 What is a Decision Tree (V2:45, 2Qs, V1:40, 1Q)
6.4 Building a Decision Tree (V2:29, 1Q, V2:03, 1Q, V5:55, 1Q)
6.5 Criteria for Determining the Best Split (V3:36, 1Q, V3:04, 1Q, V3:11, V3:13, 2Q, V2:35)
6.6 Decision Boundaries and Stopping Criteria (V2:56, 1Q, V0:59, 1Q, V3:05, 1Q)
6.7 Decision Tree Generalization (V3:14, V6:22, 2Q, V1:43)
Total Video: 50 minutes

Unit Objectives

- Understand the overall process of Hunt's Algorithm
- Understand the splitting criteria and different methods for splitting on binary, discrete, and continuous attributes
- Get an intuition for how decision trees divide a space (i.e., the decision boundaries)
- Understand the basic overview of different decision tree algorithms: ID3, CART, C4.5/J48
- Understand pruning methods of decision trees

Optional Reading

- Chapter IV (4.1-4.3, Decision Trees), from Tan, Steinbach, and Kumar
- scikit-learn decision trees documentation, http://scikit-learn.org/stable/documentation.html (tips, algorithms, mathematical formulation)
- Chapter 14 (Classification Trees), from M. Kuhn and K. Johnson, Applied Predictive Modeling
- Notebook 5 on the course github page (decision trees)

Unit Assignments

During Live Session:

- **Second live session assignment:** Decision Trees and Splitting Criteria
- You will complete an interactive Jupyter notebook assignment (working in teams if you prefer) during the live session
- You will receive a link to a Jupyter notebook containing questions about decision trees and splitting criteria from the videos.
- An early version of the notebook will be sent out during the week for you to answer the questions before class starts.
- The instructor will then work out these preliminary questions during the first 15 minutes of class.
- After this introduction, another notebook will be made available for you to work on as a team. You will be asked to live code (mostly using Numpy and scikit-learn) some of the equations we went over in the videos.
- You will complete the coding assignment and the questions and upload the updated python notebook at the end of live session.
- Ask questions to the instructor! This is meant to facilitate one-on-one time!

Unit Seven: Ensembles and Statistical Model Comparison

Lecture Units

7.3 Model Evaluation (V2:28, 1Q-e, V2:35, 1Q, V2:34, 2Q-e)

7.4 Methods for Estimating Performance (V3:37, V1:34, 1Q, V2:10, 1Q, V1:29, 1Q, V4:53)

7.5 Model Comparison (LB2:48, LB5:21, LB4:38, LB2:18)

7.6 Ensemble Classifiers (V4:22, 1Q, V7:52, 1Q, V5:48, 2Q)

Total Video: 55 minutes

Unit Objectives

- Understand various methods for assessing performance: precision, recall, F-measure, ROC, and AUC
- Understand how to use different cross validations and holdout in scikit-learn
- Statistical model comparisons: confidence intervals in K-fold and holdout
- Understand ensemble classification methods: boosting, bagging, stacking, and cascading • Use pipelining in scikit-learn for easy chaining of methods

Optional Reading

- Chapter IV (4.4-4.6, Overfitting, Evaluation, Comparing Classifiers), from Tan, Steinbach, and Kumar
- Chapter V (5.6-5.7, Ensemble Methods and Class Imbalance), from Tan, Steinbach, and Kumar
- Chapter Seven and Chapter Fifteen of The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, and Jerome Friedman
- Chapter 11, from M. Kuhn and K. Johnson, Applied Predictive Modeling
- Pipelines in sklearn: http://scikit-learn.org/stable/modules/pipeline.html

Unit Assignments

- Look at how you might use different classifiers like decision trees and ensemble methods on your chosen data sets

During live session:

- Go over the grand Poobah classification notebook of the Olivetti faces dataset
- Notebook 6 on the course github page (classification)

Unit Eight: Alternative Classification Methods

Lecture Units

8.3 Nearest Neighbors Classification (V5:21, 1Q, V3:19, 2Q-e, Y21:44, V3:44)
8.4 Nearest Neighbor Regression (V4:58, 1Q)
8.5 KD Trees (V2:36, 1Q, V2:51, 1Q, V5:18)
8.6 Naive Bayes (V1:46, 1Q, V4:53, 1Q, V4:21, 1Q, V1:45)
Total Video: 60 minutes

Unit Objectives

- Understand the complexities associated with k-nearest neighbors classification and regression
- Employ different distance metrics and understand their advantages
- Understand how data structures can be used to effectively reduce comparisons to the naive “brute force” approach
- Introduce the concept of KNN regression and enclosing/encapsulated k-neighbor regression (e-KNN)
- Understand Naive Bayes classification methods for various types of data

Optional Reading

- Chapter V (5.2 and 5.3, KNN and Naive Bayes), from Tan, Steinbach, and Kumar
- scikit-learn documentation for KNN using KD-trees and Naive Bayes:
- http://scikit-learn.org/stable/modules/naive_bayes.html
- http://scikit-learn.org/stable/modules/neighbors.html
- Chapter 13 (KNN only) of The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, and Jerome Friedman
- Chapter 13 (KNN and Naive Bayes), from M. Kuhn and K. Johnson, Applied Predictive Modeling
- Gupta, Maya R., Eric K. Garcia, and Erika Chin. "Adaptive local linear regression with application to printer color management." Image Processing, IEEE Transactions on 17, no. 6 (2008): 936-945
- http://www.countbayesie.com/blog/2015/2/18/bayes-theorem-with-lego
- Look at the grand Poobah Notebook to see how KNN and Naive Bayes can be used in scikitlearn:
- Included in notebook 6 of the course github page.

Unit Assignments

During live session:

- Third live session assignment: KNN and Naive Bayes
- You will complete an interactive Jupyter notebook assignment (working in teams if you prefer) during the live session
- You will receive a link to a Jupyter notebook that I made containing questions about using KNN and Naive Bayes in scikit-learn.
- An early version of the notebook will be sent out during the week for you to answer the questions before class starts.
- The instructor will then work out these preliminary questions during the first 15 minutes of class.
- After this introduction, another notebook will be made available for you to work on as a team.
- You will be asked to live code (mostly using Numpy and scikit-learn) some of the algorithms we went over in the videos.
- You will complete the coding assignment and the questions and upload the updated python notebook at the end of live session.
- Ask questions to the instructor! This is meant to facilitate one-on-one time!

Unit Nine: Second Team Project Work Week, Classification

- No lecture content: Use scikit-learn to properly predict from your dataset
- Turn in the second project work week assignment by Sunday at midnight following live session
- Please note that this assignment should be much more rigorous than the previous min-project on classification. More explanations and data exploration is required for this assignment.

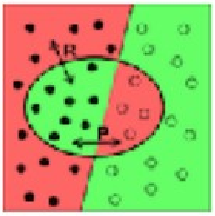
During Live Session (Optional):

- Come with questions about using pandas on your datasets
- I will answer questions to each team individually in breakout sessions, please work on refining assignments as a team during class time
- Show some of what your team has been working on and any problems you have run into. Others will benefit greatly from these types of questions.

This is due the Sunday (midnight) following the ninth live session of the course:

Second Project Work Week Assignment

15/1 Split
15/5 Split
1/19 Split
All Data
Training Data
Holdout Data
Training Data
Holdout Data
Training Data
Holdout Data



You are to build upon the predictive analysis (classification) that you already completed in the previous mini-project, adding additional modeling from new classification algorithms as well as more explanations that are inline with the CRISP-DM framework. You should use appropriate cross validation for all of your analysis (explain your chosen method of performance validation in detail). Try to use as much testing data as possible in a realistic manner (you should define what you think is realistic and why).

This report is worth 20% of the final grade. Please upload a report (one per team) with all code used, visualizations, and text in a single document. The format of the document can be PDF, *.ipynb, or HTML. You can write the report in whatever format you like, but it is easiest to turn in the rendered Jupyter notebook. The results should be reproducible using your report. Please carefully describe every assumption and every step in your report.

Dataset Selection

Select a dataset identically to the way you selected for the first project work week and mini-project. You are not required to use the same dataset that you used in the past, but you are encouraged. You must identify two tasks from the dataset to regress or classify. That is:

- Two classification tasks OR
- Two regression tasks OR
- One classification task and one regression task

For example, if your dataset was from the diabetes data you might try to predict two tasks: (1) classifying if a patient will be readmitted within a 30 day period or not, and (2) regressing what the total number of days a patient will spend in the hospital, given their history and specifics of the encounter like tests administered and previous admittance.

Grading Rubric

Data Preparation **(15 points total)**

- **[10 points]** Define and prepare your class variables. Use proper variable representations (int, float, one-hot, etc.). Use pre-processing methods (as needed) for dimensionality reduction, scaling, etc. Remove variables that are not needed/useful for the analysis.
- **[5 points]** Describe the final dataset that is used for classification/regression (include a description of any newly formed variables you created).

Modeling and Evaluation **(70 points total)**

- **[10 points]** Choose and explain your evaluation metrics that you will use (i.e., accuracy, precision, recall, F-measure, or any metric we have discussed). Why are the measure(s) appropriate for analyzing the results of your modeling? Give a detailed explanation backing up any assertions.
- **[10 points]** Choose the method you will use for dividing your data into training and testing splits (i.e., are you using Stratified 10-fold cross validation? Why?). Explain why your chosen method is appropriate or use more than one method as appropriate. For example, if you are using time series data then you should be using continuous training and testing sets across time.
- **[20 points]** Create three different classification/regression models for each task (e.g., random forest, KNN, and SVM for task one and the same or different algorithms for task two). Two modeling techniques must be new (but the third could be SVM or logistic regression). Adjust parameters as appropriate to increase generalization performance using your chosen metric. You must investigate different parameters of the algorithms!
- **[10 points]** Analyze the results using your chosen method of evaluation. Use visualizations of the results to bolster the analysis. Explain any visuals and analyze why they are interesting to someone that might use this model.
- **[10 points]** Discuss the advantages of each model for each classification task, if any. If there are not advantages, explain why. Is any model better than another? Is the difference significant with 95% confidence? Use proper statistical comparison methods. You must use statistical comparison techniques—be sure they are appropriate for your chosen method of validation as discussed in unit 7 of the course.

- **[10 points]** Which attributes from your analysis are most important? Use proper methods discussed in class to evaluate the importance of different attributes. Discuss the results and hypothesize about why certain attributes are more important than others for a given classification task.

Deployment **(5 points total)**

- How useful is your model for interested parties (i.e., the companies or organizations that might want to use it for prediction)? How would you measure the model's value if it was used by these parties? How would you deploy your model for interested parties? What other data should be collected? How often would the model need to be updated, etc.?

Exceptional Work **(10 points total)**

- You have free reign to provide additional analyses.
- One idea: grid search parameters in a parallelized fashion and visualize the performances across attributes. Which parameters are most significant for making a good model for each classification algorithm?

Unit Ten: Basic Clustering

Lecture Units

10.3 What is clustering (V2:18, 2Q, V4:37, 2Q)
10.4 K-means Clustering (V1:13, 1Q, V5:07, 1Q-u, V1:58, V4:27, 2Q, V3:22, V2:32, 1Q)
10.5 Hierarchical Agglomerative Clustering (HAC) (V6:40, 1Q, V2:26)
10.6 Density Based Clustering (V2:301Q, V4:36)
10.7 Clustering Examples (S23:47)
10.8 Clustering Validity (V4:46, V5:06)
Total Video: 50 minutes
Notebook 9 on the course github page (basic clustering)

Unit Objectives

- Be able to identify the various types of clusters that exist (globular, partitional, etc.)
- Understand efficiency of K-means and the different implementations
- Understand the limitations of k-means
- Understand hierarchical agglomerative clustering and the different merging schemes
- Understand the basics of DBSCAN and how to find Eps given a MinPts
- Understand how to evaluate the validity of clusters using various methods

Optional Reading

- Chapter VIII, from Tan, Steinbach, and Kumar
- scikit-learn documentation on clustering: http://scikit-learn.org/stable/modules/clustering.html
- Only look at k-means, hierarchical agglomerative clustering, DBSCAN, and cluster validity (That is, skip spectral clustering, birch, etc.)
- Chapter Thirteen (K-means only) of The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, and Jerome Friedman Unit

Assignments

During Live Session:

- Fourth live session assignment: Basic Clustering
- You will complete an interactive Jupyter notebook assignment (working in teams if you prefer) during the live session
- You will receive a link to a Jupyter notebook that I made containing questions about using k-means, HAC, and DBSCAN in scikit-learn. You will be asked to live code (mostly using scikit-learn) some of the algorithms we went over in the videos.
- **A note for Windows users:** k-means clustering is not optimized in the current release of scikit for the windows OS using python 2. It therefore takes many minutes to run each iteration (instead of milliseconds). The easiest way to run the notebook is via mac or linux or to upgrade to python 3 on windows.
- You will complete the coding assignment and the questions and upload the updated python notebook at the end of live session.
- Ask questions to the instructor! This is meant to facilitate one-on-one time!

Unit Eleven: More Clustering

Lecture Units

11.3 Gaussian Mixtures (V2:01, 2Q, V1:28, 1Q, V3:45, 1Q, V3:43)
11.4 Graph Clustering (V3:37, 1Q-u, V2:23, V2:12, V4:14)
11.5 Spectral Clustering (V3:15, V4:53)
Total Video: 30 minutes

Unit Objectives

- Understand limitations of current clustering schemes
- Introduce Expectation Maximization and Probabilistic clustering with mixture models
- Graph based clustering:
- Understand two common algorithms: spectral clustering and chameleon

Optional Reading

- Chapter IX (9.1, 9.2.2, and 9.4.1-9.4.4, Characteristics and Graph based clustering), from Tan, Steinbach, and Kumar
- scikit-learn's documentation of GMMs: http://scikit-learn.org/stable/modules/mixture.html
- scikit-learn's documentation of using the BIC: http://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_selection.html#example-mixture-plot-gmm-selection-py
- http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf
- Chapter Eight (8.2-8.5, Expectation Maximization and Bayes) of The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, and Jerome Friedman

Unit Assignments

- Look for at final assignment for the course and start to make selections about what topic you would like to do: Clustering, Association Rule Mining, or Collaborative Filtering
- During Live Session:
- Go over the example clustering notebook in class:
- Notebook 11 on the course github page (advanced clustering)
- Introduce Dato for clustering and the SFrame from graphlab-create

Unit Twelve: Association Rule Mining

Lecture Units

12.3 What is Association Rule Mining (V4:11, 1Q, V1:45, 1Q, V1:01)

12.4 Frequent Itemset Generation (V2:02, V0:52, V1:27, V1:50, 1Q, V1:29, V3:05, V2:51, V1:35)

12.5 Frequent Pattern Trees (V3:46, 1Q, S15:31, 1Q, S5:11)

12.6 Rules Generation (V3:52, 1Q, V1:35)

12.7 Maximal and Closed Sets (V4:11)

12.8 Evaluating Association Rules (V4:40, 1Q-e, V3:37, V2:42)

Total Video: 66 minutes

Unit Objectives

- Understand frequent itemset generation and how it can be decoupled from association rule generation
- Understand the apriori algorithm and how it reduces the complexity of finding itemsets that are frequent
- Get an intuition about frequent pattern trees
- Understand rules generation apriori algorithm
- Understand compact rules representations: maximal and closed sets

Reading/Installation

- Chapter VI, from Tan, Steinbach, and Kumar
- Install R, and libraries arules and arulesviz
  - http://www.r-project.org
- Here is a good blog on installing packages with R: http://www.r-bloggers.com/ installing-r-packages/
- Install Rpy2 (windows users may not be able to install this correctly)
- http://cran.r-project.org/web/packages/arules/index.html
- http://cran.r-project.org/web/packages/arules/arules.pdf
- http://aimotion.blogspot.com/2013/01/machine-learning-and-data-mining.html
- Using R and the apriori package: https://www.youtube.com/watch?v=DQGJhZNhG4M
- Here is a Jupyter notebook I made based on an example from Dr. Michael Hahsler. This might be useful for those wanting to do their final project on association rule mining:
- Notebook 12 on the course github page (association analysis) Unit Assignments

During live session:

- Final live session assignment: Association Rules
- You will complete an interactive Jupyter notebook assignment (working in teams if you prefer) during the live session

- You will receive a link to a Jupyter notebook that I made containing questions about using association rule mining in Jupyter using R. You will be asked to live code (mostly using Numpy and R) for frequent itemsets and rules generation we talked about in the videos.
- **A note for Windows users:** Rpy2 ended its support for windows. It therefore will likely not run on the windows OS. The easiest way to run the notebook is via mac or linux. If you cannot install Rpy2, an alternative notebook will be give to you.
- You will complete the coding assignment and the questions and upload the updated python notebook at the end of live session.
- To test if you have everything installed correctly on your system: please download the following notebook with data and try to run it:
- ICA5_Test folder from the course github page
- Ask questions to the instructor! This is meant to facilitate one-on-one time!!

Unit Thirteen: Collaborative Filtering, Optional

Lecture Units:

13.4 Intro to collaborative filtering (V1:33, V3:11, V2:04, V2:41, V1:21)
13.5 SFrames in Dato (Y32:45)
More In-Depth Explanation of Collaborative Filtering in GraphLab Page:
https://www.youtube.com/watch?v=I-xPNMCqW2o&feature=youtu.be (V49:00)

Total Video: 92 minutes

Unit Objectives

- Understand the basics of collaborative filtering using item-item similarity and matrix factorization for user-item similarity
- Understand the utility of out-of-core computations without parallelization
- Understand how to evaluate different recommender systems in graph lab create

Optional Reading

- Look at dato's online documentation: https://dato.com/products/create/docs/
- In particular, look at the scalable data structure (SFrame) and also the recommender API or examples in their notebooks
- Look at the open source core of Dato: https://dato.com/products/create/open_source.html
- http://guidetodatamining.com/guide/ch2/DataMining-ch2.pdf
- http://michael.hahsler.net/research/Recommender_SMU2011/slides/Recomm_2011.pdf
- https://dato.com/learn/gallery/index.html

Installation

Install Dato's graphlab create on your system: https://dato.com/products/create/

Unit Assignments

Keep working on your final projects!

During Live session:

- Explain evaluation methods for recommender systems
- Go over collaborative filtering in Dato notebook:
- Notebook 13 from the course github page (collaborative filtering)

Unit Fourteen: Data Parallel in Python

Lecture Units:

14.3 Data Parallel (V5:10, V3:08)
14.4 Pycon Tutorial (Y1:33:00)
Total Video: 2 hours

As an alternative to the sklearn video content:
Look at Dask, Dato, or parallelism with PySpark

Unit Objectives:

- Understand advantages and limitations of map-reduce in a iterative data mining algorithms
- Understand grid searching and parametrization with scikit-learn

Optional Reading:

- Data parallel in Association Rules and Clustering: http://users.eecs.northwestern.edu/~yingliu/papers/para_arm_cluster.pdf
- SVMs in parallel (approximate): http://infolab.stanford.edu/~echang/CIKM-tutorial.pdf
- Chapter Two and Twelve of Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman
- The professional version of Graphlab allows configuring parallel environments on things like AWS and hadoop. If the money is there, these are excellent and scalable implementations that can be run in just a few lines of code. But be sure you actually need parallelism. Often it’s just out-of-core computation that is needed.

Unit Assignments:

Keep working on the final assignment! It is due at the end of next week!!

During Final Live Session:

- Individual team meetings
- Come with questions about using clustering, rule mining, or recommendations on your datasets
- I (and other teams) will answer as many questions as possible during the live session

- Use scikit-learn/R/Dato to properly cluster/associate/recommend from your dataset
- Turn in the final project work week assignment by Sunday at midnight following live session
- Please make this assignment a breadth of what you have learned in the final third of the class, showing me as much as possible in terms of different techniques and proper evaluation on your datasets.
- I will answer questions to each team individually in breakout sessions, please work on refining assignments as a team during class time

- Show some of what your team has been working on and any problems you have run into. Others will benefit greatly from these types of questions.

Final Team Project Work Week

CRISP-DM Capstone: Association Rule Mining, Clustering, or Collaborative Filtering

In the final assignment for this course, you will be using one of three different analysis methods:

- Option A: Use clustering on an unlabeled dataset to provide insight or features
- Option B: Use transaction data for mining associations rules
- Option C: Use collaborative filtering to build a custom recommendation system

Your choice of dataset will largely determine the task that you are trying to achieve, though the dataset does not need to change from your previous tasks.

- For example, you might choose to use clustering on your data as a preprocessing step that extracts different features. Then you can use those features to build a classifier and analyze its performance in terms of accuracy (precision, recall) and speed.
- Alternatively, you might choose a completely different dataset and perform rule mining or build a recommendation system.

Dataset Selection and Toolkits

As before, you need to choose a dataset that is not small. It might be massive in terms of the number of attributes (or transactions), classes (or items, users, etc.) or whatever is appropriate for the task you are performing. Note that scikit-learn can be used for clustering analysis, but not for Association Rule Mining (you should use R) or collaborative filtering (you should use graphlabcreate from Dato). Both can be run using Jupyter notebooks as shown in lecture.

- One example of a recommendation dataset is the movie lens rating data: http://grouplens.org/ datasets/movielens/
- Some examples of association rule mining datasets: http://fimi.ua.ac.be/data/

Write a report covering in detail all the steps of the project. The results need to be reproducible using only this report. Describe all assumptions you make and include all code you use in the Jupyter notebook or as supplemental functions. Follow the CRISP-DM framework in your analysis (you are performing all of the CRISP-DM outline).

This report is worth 20% of the final grade.

Grading Rubric:

Business Understanding **(10 points total).**

- **[10 points]** Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?). How will you measure the effectiveness of a good algorithm? Why does your chosen validation method make sense for this specific dataset and the stakeholders needs?

Data Understanding **(20 points total)**

- **[10 points]** Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file. Verify data quality: Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems?
- **[10 points]** Visualize the any important attributes appropriately. Important: Provide an interpretation for any charts or graphs.

Modeling and Evaluation **(50 points total)**

Different tasks will require different evaluation methods. Be as thorough as possible when analyzing the data you have chosen and use visualizations of the results to explain the performance and expected outcomes whenever possible. Guide the reader through your analysis with plenty of discussion of the results. Each option is broken down by:

- **[10 Points]** Train and adjust parameters
- **[10 Points]** Evaluate and Compare
- **[10 Points]** Visualize Results
- **[20 Points]** Summarize the Ramifications

<u>Option A: Cluster Analysis</u>

- Train: Perform cluster analysis using several clustering methods (adjust parameters).
- Eval: Use internal and/or external validation measures to describe and compare the clusterings and the clusters— how did you determine a suitable number of clusters for each method?
- Visualize: Use tables/visualization to discuss the found results. Explain each visualization in detail.
- Summarize: Describe your results. What findings are the most interesting and why?

<u>Option B: Association Rule Mining</u>

- Train: Create frequent itemsets and association rules (adjust parameters).
- Eval: Use several measures for evaluating how interesting different rules are.
- Visualize: Use tables/visualization to discuss the found results.
- Summarize: Describe your results. What findings are the most compelling and why?

<u>Option C: Collaborative Filtering</u>

- Train: Create user-item matrices or item-item matrices using collaborative filtering (adjust parameters).
- Eval: Determine performance of the recommendations using different performance measures (explain the ramifications of each measure).

- Visualize: Use tables/visualization to discuss the found results. Explain each visualization in detail.
- Summarize: Describe your results. What findings are the most compelling and why?

Deployment **(10 points total)**

Be critical of your performance and tell the reader how you current model might be usable by other parties.

- Did you achieve your goals? If not, can you reign in the utility of your modeling?
- How useful is your model for interested parties (i.e., the companies or organizations that might want to use it)?
- How would you deploy your model for interested parties?
- What other data should be collected?
- How often would the model need to be updated, etc.?

Exceptional Work **(10 points total)**

You have free reign to provide additional analyses or combine analyses.