Predictive Analytics

Instructor: Cesar Acosta-Mejia

Course Description

The course focus is on building models for prediction and classification. The standard multiple linear regression model is extended to shrinkage models (ridge and lasso regression) for improved accuracy and dimension reduction. Overfitting, bias, cross validation, and AIC are used to evaluate the performance of these models. Models for classification including logistic regression, KNN, and multinomial regression, are reviewed and their prediction performance is estimated by means of error rates and gini index. Family of analytic models such as classification and regression trees (CART), ensembles of trees (random forests, bagging, and gradient boosting), support vector machines, and neural networks (for regression and classification) are evaluated.

Learning Objectives and Outcomes

- To understand the Data Analytics levels: Descriptive, Predictive, and Prescriptive Analytics and how they relate to Data Science.
- To understand the difference between supervised and unsupervised learning methods.
- To learn the most common data aggregation operations (cross tabulation and pivot tables).
- To build models for prediction and classification.
- To understand key concepts for data science modeling (overfitting, shrinkage, regularization, loss function, cross-validation).
- To learn how to apply cross validation to models with hyperparameters.
- To learn how to derive the loss function for shrinkage models.
- To compare the performance of different prediction and classification models.
- To build models to classify observations into two or more classes (categories).

Prerequisites: None

Recommended Preparation: An undergraduate course on Statistics, working knowledge of a programming language

Course Notes

The course material is available online.

Technological Proficiency and Hardware/Software Required

Required software: Python programming language. Jupyter Notebook is used as the main interface for documenting the scripts and results.

Textbook

- VanderPlas, *Python Data Science Handbook*, O'Reilly, 2017

Supplementary Materials (for reference)

- Kong, *Python Programming and Numerical Methods,* Academic Press, 2020

Description and Assessment of Assignments

Unless otherwise noted the assignments are individual. Dates are shown in the Course schedule. Submit on to Blackboard by the due date. No late homework is to be accepted.

Grading Policy

Assignment	Points	% of Grade
Homework	100 each (6^+ homework assignments)	30
Midterm	100	30
Final	100	40
TOTAL		100

Grading Scale (Course final grades will be determined using the following scale)

A	94-100
A-	90-94.9
B+	87-89
B	83-86.9
B-	80-82.9
C+	77-79
C	73-76.9
C-	70-72.9
D+	67-69
D	63-66.9
D-	60-62.9
F	59.9 and below

Assignment Submission Policy
Assignments should be typewritten and clean. Email submissions and late submissions are not allowed. No make-up exams are considered.

Timeline and Rules for submission
Assignments are to be returned the week after submission. Solutions will be released soon after the homework submission date.

Course Schedule: A Weekly Breakdown

Lab	Date	Topics/Daily Activities	Deliverables	slides	Files
1	**Aug 21**	**Introduction** to Analytics Descriptive, Predictive and Prescriptive Analytics. Python and Jupyter Notebook (JN)		overview.ppt analytics.ppt python.ppt	intro.ipynb dictionary.ipynb paragraph.txt
1	**Aug 23**	Python data structures. **Numpy** library. Operations on numpy arrays.		numpy.ppt	numpy.ipynb numpyreg.ipynb Odometer.csv
2	**Aug 28**	**Pandas** library. Data structures. Most Common Data Operations.		Pandas .ppt	
2	**Aug 30**	Pandas library. Pivot tables and cross tabulation.	HW1 Pandas		Example4.ipynb Cars93.csv
3	**Sep 4** (recorded)	**Data Visualization.** library matplotlib		matplot.ppt	matplot.ipynb
3	**Sep 6**	Web scraping with the pandas-datareader library. Data Visualization with pandas.	HW1 due HW2 Financial Analytics	returns3.ppt	finance13.ipynb
4	**Sep 11**	**Linear Regression.** OLS, regression assumptions, confidence and prediction intervals.		slr1.ppt mlr4.ppt	
4	**Sep 13**	**Linear Regression.** Examples with libraries sklearn and statsmodels	HW2 due HW3 regression	Cars93.csv Odometer.csv	slr3.ipynb finished3.ipynb
5	**Sep 18**	**Linear Regression** with categorical variables. Label encoding and one-hot encoding. Interaction terms. Examples.		categoricals.ppt	plots3.ipynb example1d.ipynb example2.ipynb
5	**Sep 20**	**Linear Regression applications**. Time Series forecasting.	HW3 due HW4 case		part2c.ipynb homes.ipynb
6	**Sep 25**	Overfitting. Cross validation strategies. Training/test sets, mean square prediction error (MSPE).		cv3.ppt part1.ppt	
6	**Sep 27**	**Linear Regression applications**. Polynomial regression. Feature selection. Scaling the data.	HW4 due	part2.ppt Auto.csv Credit.csv	Polynomial4.ipynb itertools.ipynb feature-cv7.ipynb
7	**Oct 2-4**	**MIDTERM**	E1		
8	**Oct 9**	**Classification Problems.** K-nearest neighbor (KNN). Hyperparameter search.		knn2.ppt	cancerknn.ipynb
8	**Oct 11**	**Classification Problems.** Logistic Regression. Cross Entropy Loss function. Pipelines for scaling with K-fold cross validation.	HW5 logistic	classification2.ppt logistic3.ppt	task.csv cancerlogistic.ipynb sklogis3.ipynb
9	**Oct 16**	**Regularization and Overfitting** Ridge regression and the LASSO. Hyperparameter tuning, validation set		rr2.ppt	
9	**Oct 18**	**Regularization and Overfitting.** Examples.	HW5 due HW6 ridge regression		ridge9.ipynb Hitters.csv cancerlogisticRR.ipynb

10	**Oct 23**	**Trees based Methods.** Predictors Space strategy. Tree pruning. Feature Selection. Regression trees. Examples.		trees3.ppt	regression4.ipynb Boston.csv
10	**Oct 25**	**Classification Trees.** Performance Measures for classification trees (gini index, cross entropy). Examples.	HW6 due	categ.ppt	cart4.ipynb carseats.csv
11	**Oct 30**	**Ensemble of Regression Trees.** Random Forest, Bagging, and Gradient Boosting.		ensembles4.ppt	ensemblereg4.ipynb polyboosting8.ipynb dataset.csv
11	**Nov 1**	**Ensemble of Classification Trees.** Applications and Examples.	HW7 Ensembles		ensembcancer3.ipynb
12	**Nov 6**	**Support Vector Machines.** Maximal Classifier, Support Vector Classifier, Support Vector Machine		svm.ppt	svm.ipynb function5.ipynb
12	**Nov 8**	**Support Vector Machines for regression.**	HW7 due	svmreg.ppt	optdigits.ipynb svmreg.ipynb
13	**Nov 13**	**Introduction to Neural Networks (NN).** Data representations for NN. Loss functions. Gradient descent.		nn4.ppt	perceptron5.ipynb multilayerp4.ipynb gradient5.ipynb
13	**Nov 15**	**NN Applications.** Library Keras. NN for binary classification. K-fold cross validation. NN for regression.			mnist.ipynb
14	**Nov 20-22**	**Thanksgiving Break**			
15	**Nov 27**	**Final Review**	E2 released		
		Final Exam TBD	E2 due		