MATH 389: Statistical Learning

Instructor:	**Taylor Arnold**
E-mail:	tarnold2@richmond.edu
Office:	Jepson Hall, Rm 218
Office hours:	TBD

Catalogue Description:

Develops an understanding of methods and algorithms for building predictive models from data. Topics include model complexity, hyper-parameter tuning, over- and under-fitting, and the evaluation of predictive performance. Models covered include linear regression, penalized regression, additive models, gradient boosted trees, and neural networks. Applications are drawn from many areas, with a particular focus on processing unstructured text and image corpora.

Computing:

The focus of this course will be on developing an understanding of model and algorithms for building predictive models from data. To facilitate this, nearly every class assignment and exam will involve some form of computing. No prior programming experience is assumed or required.

We will use the **R** programming environment throughout the semester. It is freely available for all major operating systems and is pre-installed on many campus computers. You can download it and all supporting files for your own machine via these links:

```
https://cran.r-project.org/
https://www.rstudio.com/
```

I strongly recommend using your own machine for this course and bring a laptop or tablet to each class meeting. The lab computers in Jepson are available and contain some, though not all, of the required software.

Course Website:

All of the materials and assignments for the course will be posted on the class website:

```
https://statsmaths.github.io/stat395
```

At the end of the semester, this version of the course will be archived and available for your reference.

GitHub:

All of your work for this semester will be submitted through GitHub, the same platform that hosts our website. You'll need to set up a free account, which we will cover during the week of class.

Labs:

Every class will have an associated file named lab00.Rmd, with the appropriate class number replaced for the 00. By noon before the start of the next class, you must complete the questions contained within the lab notebook. Assignments will be submitted through GitHub; this process will explained in more detail during class.

During most class meetings, we will do some combination of presenting your results in small groups or to the class in general. Note that your presence and attention in class will be an important aspect of your lab grade.

Final Project:

There will also be a final project for this course consisting of both written and oral components. The overarching goal of the assignment is to apply what we have learned this semester to a dataset that has not been nicely cleaned for analysis. The skills we have covered are very applicable. I want students to leave the course feeling comfortable applying these skills to real datasets outside of the relatively clean format I gave you in the course.
There are three main approaches that can be taken for the final project:

- **data collection**: create a new dataset yourself, load it into R, and apply statistical learning techniques to the data. This could in theory be from any source, but to be interesting it will probably be either collecting another photo dataset or curating a dataset of texts
- **predictive modelling competition**: choose a Kaggle predictive learning competition and write up your attempts to build a predictive model. Generally, make sure that the dataset is sufficiently difficult by choosing something that includes multiple input files or works with text, image, or other non-tabular datasources.
- **new method**: find a new method and/or R package that we have not covered in class and apply this to your dataset. Ideas include working with network data, or using Bayesian models. Your write-up should include background on the method similar to what my lecture notes cover.

Weekly Topics:

Each week of the semester focuses on a new topic in statistical learning. Topics and several references are given below for each week's materials (exact coverge may change from year to year based on pace and interest). The following textbooks are referenced below with the respective abbreviations:

- Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. *The Elements of Statistical Learning.* **(EoSL)**
- Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. **(DL)**
- Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View.* **(ADA)**
- Garrett Grolemund and Hadley Wickham. *R for Data Science.* **(R4DS)**

All of these texts are available as free digital resources. Direct links to all of the material will be given in the course notes. Note that these references are meant to supplement, not replace, the lecture notes.

WEEK 01 - Introduction to R, RMarkdown, and Graphics

- **R4DS**, Chapters 1-3
- W. N. Venables, D. M. Smith and the R Core Team. An Introduction to R.
- **R:** dplyr, readr, ggplot2

WEEK 02 - Linear Regression

- **EoSL**, Chapter 3
- **ADA**, Chapters 1-2
- **R4DS**, Chapter 3
- **R:** stats::lm

WEEK 03 - Logistic regression and classification

- **EoSL**, Chapters 4 and 7.1-7.3
- **ADA**, Chapters 3 and 11
- Ye, Jianming (1998). "On Measuring and Correcting the Effects of Data Mining and Model Selection." Journal of the American Statistical Association, 93: 120–131.
- **R:** stats::glm

WEEK 04 - Incorporating Non-Linear Effects

- **EoSL**, Chapter 5

- **ADA**, Chapter 7
- L.J.P. van der Maaten, E.O. Postma, H.J. van den Herik. Dimensionality Reduction: A Comparative Review
- **R:** stats::poly

WEEK 05 - Adaptive, local models

- **EoSL**, Chapter 6
- **ADA**, Chapter 4
- Buja, Andreas, Trevor Hastie and Robert Tibshirani (1989). "Linear Smoothers and Additive Models." Annals of Statistics, 17: 453–555.
- Ledoux, Michel. The concentration of measure phenomenon. No. 89. American Mathematical Soc., 2005.
- **R:** stats::smooth.spline

WEEK 06 - Penalized regression (Ridge and lasso)

- **EoSL**, Chapter 4
- **ADA**, Chapter 16 and Appendix H
- Robert Tibshirani (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B.
- **R:** glmnet

WEEK 07 - Additive models

- **EoSL**, Chapter 9
- **ADA**, Chapter 9
- Buja, Andreas, Trevor Hastie and Robert Tibshirani (1989). "Linear Smoothers and Additive Models." Annals of Statistics, 17: 453–555.
- Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion), Annals of Statistics 28: 337–307
- **R:** mgcv

WEEK 08 - Decision trees, random forests, gradient boosted trees

- **EoSL**, Chapter 10
- **ADA**, Chapter 13
- Breiman, L. (2001). "Random forests", Machine Learning 45: 5–32
- Buhlmann, P. and Hothorn, T. (2007). "Boosting algorithms: regularization, prediction and model fitting (with discussion)", Statistical Science 22(4): 477–505

- Friedman, J. (2001). “Greedy function approximation: A gradient boosting machine”, Annals of Statistics 29(5): 1189–1232
- **R:** randomForest, xgboost

WEEK 09 - Support vector machines

- **EoSL**, Chapter 12
- Chang, Chih-Chung, and Chih-Jen Lin. “LIBSVM: A library for support vector machines.” ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 (2011): 27.
- Zhu, Kaihua, et al. “Parallelizing support vector machines on distributed computers.” Advances in Neural Information Processing Systems. 2008.
- **R:** e1071

WEEK 10 - Dense neural networks

- **EoSL**, Chapter 11
- **DL**, Chapters 6 and 7
- Michael A. Nielsen, “Neural Networks and Deep Learning”, Determination Press, 2015.
- Hinton, GE; Osindero, S; Teh, YW (Jul 2006). “A fast learning algorithm for deep belief nets.”. Neural computation 18 (7): 1527–54.
- **R:** keras

WEEK 11 - Featurizing textual data

- **R4DS**, Chapter 14
- **DL**, Chapter 14
- Taylor Arnold, “A Tidy Data Model for Natural Language Processing Using cleanNLP.” The R Journal, 9.2, 1-20 (2017).
- **R:**, cleanNLP, tokenizers

WEEK 12 - Convolutional neural networks

- **DL**, Chapters 8 and 9
- Ciresan, Dan; Meier, Ueli; Schmidhuber, Jürgen (June 2012). “Multi-column deep neural networks for image classification”. CVPR 2012.
- Krizhevsky, A., Sutskever, I., and Hinton, G. “ImageNet classification with deep convolutional neural networks.” In Advances in Neural Information Processing Systems 25 (NIPS 2012).

WEEK 13 - Transfer learning with CNNs

- **DL**, Chapter 10
- He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” arXiv preprint arXiv:1512.03385 (2015).
- Yosinski, Jason, et al. “How transferable are features in deep neural networks?.” Advances in Neural Information Processing Systems. 2014.
- Razavian, Ali S., et al. “CNN features off-the-shelf: an astounding baseline for recognition.” Computer Vision and Pattern Recognition Workshops (CVPRW).

WEEK 14 - Recurrent neural networks and word embeddings

- **DL**, Chapter 10
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality.” In Advances in neural information processing systems, pp. 3111-3119. (2013).
- Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. “Improving word representations via global context and multiple word prototypes.” In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.
- Pascanu, R., Gulcehre, Ç., Cho, K., and Bengio, Y. “How to construct deep recurrent neural networks. (ICLR 2014).
- **R:** fasttextM