Data Mining and Machine Learning: STAT 365 / STAT 665

Spring 2016 Monday, Wednesdays 14:30 - 15:45 DL 220

Instructor:	**Taylor Arnold**
E-mail:	taylor.arnold@yale.edu
Office:	24 Hillhouse, Rm 203
Office Hours:	tbd
Teaching Assistants:	Yu Lu, Jason Klusowski
TA Hours:	tbd
Website:	`http://www.stat.yale.edu/~tba3/stat665/`

Course Description:

Machine learning is an incredibly diverse field that sits at the intersection of computer science and applied statistics. This course will concentrate on the applied aspects of machine learning, centered around supervised classification problems. We will briefly cover regularized linear models and support vector machines before spending the majority of the semester on neural networks. Applications to problems in natural language processing and computer vision will serve as motivating examples.

Prerequisites:

- Proficient in R
- Introductory statistical theory
- Exposure to applied data analysis

References:

- Ian Goodfellow, Aaron Courville and Yoshua Bengio. *Deep Learning*. Book in preparation for MIT Press. `http://www.deeplearningbook.org/`.
- Jerome Friedman, Trevor Hastie and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, Berlin: Springer Series in Statistics, 2011.
- Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Book in preparation. `http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/`.
- Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Problem Sets:

There will be approximately 9 problem sets assigned throughout the semester, due on Thursdays. These will consist of both building custom implementations of machine learning algorithms, as well as applying established libraries to machine learning problems. You should expect to become comfortable working simultaneously in a number of programming languages. All submissions will be made electronically on the ClassesV2 site.

Grading:

Course grades will be determined based on scores from the problem sets. I want to make the grading extremely transparent, so these will all be graded on an 10 point scale (with the possibility of up to one additional point for truly exceptional work or extra credit questions). The final grade will be calculated by dropping the lowest grade, rounding the average of remainder to the nearest integer and reading off of the following table:

Numeric Score	Final Grade	
10	A	H
9	A-	H
8	B+	HP
7	B	HP
6	B-	HP
5	C+	P
4	C	P
3	C-	P
2	D	F
1	F	F
0	F	F

Because I will be dropping the lowest score, we will only accept late assignments in the event of exceptional circumstances (such as extended absences or family emergencies). In particular, Undergraduates enrolled in STAT 365 must submit a Dean's excuse for any late submission.

Tentative Schedule:

- 2016-01-20: Course introduction
- 2016-01-22: Linear classification methods I (EoSL 3 & 4)
- 2016-01-25: Linear classification methods II (EoSL 3 & 4)
- 2016-01-27: Random forests and gradient boosting (EoSL 10 & 15)
- 2016-02-01: Support vector machines I (EoSL 12)
- 2016-02-03: Support vector machines II (EoSL 12)
- 2016-02-08: Introduction to Neural Networks I (DL 6.1-6.2)
- 2016-02-10: Introduction to Neural Networks II (DL 6.3-6.4)
- 2016-02-15: Back-propagation (DL 6.4)
- 2016-02-17: Gradient Descent (DL 8.1-8.3)
- 2016-02-22: Adaptive Learning (DL 8.4)
- 2016-02-24: Introduction to Theano
- 2016-02-29: Computer Vision (DL 12.1)
- 2016-03-02: Convolution Networks (DL 9.1-9.2)

- 2016-03-07: Pooling in CNNs (DL 9.4)
- 2016-03-09: Unsupervised CNNs and Transfer Learning (DL 9.8)
- 2016-03-28: ILSVRC 2015 & MS COCO - Object Detection
- 2016-03-30: ILSVRC 2015 & MS COCO - Object Localization
- 2016-04-04: Deep Residual Learning (DL 20)
- 2016-04-06: Moving Images
- 2016-04-11: Natural Language Processing (DL 12.4)
- 2016-04-13: Word Embeddings
- 2016-04-18: Recurrent Neural Networks (DL 10)
- 2016-04-20: Dependency Parsers I
- 2016-04-25: Dependency Parsers II
- 2016-04-27: Natural Language Inference

Tentative Problem Sets:

- 2015-02-04: Linear models and tree-based methods
- 2015-02-11: Support vector machines
- 2015-02-18: Intro to neural networks
- 2015-02-25: Neural network optimization
- 2015-03-10: Neural networks with Theano
- 2015-03-31: Convolution networks and computer vision
- 2015-04-07: Applied problems in computer vision
- 2015-04-21: Recurrent neural networks
- 2015-04-28: Applied problems in natural language processing

ML Datasets of interest:

- Taxi Data: `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`
- Million Song Dataset: `http://labrosa.ee.columbia.edu/millionsong/`
- MNIST: `http://yann.lecun.com/exdb/mnist/`
- CIFAR-10/CIFAR-100: `https://www.cs.toronto.edu/~kriz/cifar.html`
- The Street View House Numbers (SVHN) Dataset: `http://ufldl.stanford.edu/housenumbers/`
- ILSVRC: `http://image-net.org/challenges/LSVRC/2015/`
- Microsoft Common Images in Context (MS COCO): `http://mscoco.org/dataset/`
- Stanford Natural Language Inference: `http://nlp.stanford.edu/projects/snli/`
- TIMIT Continuous Speech Corpus: `http://catalog.ldc.upenn.edu/LDC93S1`

Selection of neural network software:

- torch: http://torch.ch/
- Caffe: http://caffe.berkeleyvision.org/
- theano: http://deeplearning.net/software/theano/
- keras: https://github.com/fchollet/keras
- blocks: http://blocks.readthedocs.org/en/latest/
- Lasagne: https://lasagne.readthedocs.org/en/latest/
- tensorflow: https://github.com/tensorflow/tensorflow
- CNTK: https://cntk.codeplex.com/
- deeplearning4j: http://deeplearning4j.org/

Selection of conference proceedings in machine learning:

- NIPS: http://papers.nips.cc/
- ICML: http://www.machinelearning.org/icml.html
- ICDM: http://ieeexplore.ieee.org/xpl/conhome.jsp?punumber=1000179
- SIGKDD: http://dl.acm.org/citation.cfm?id=2783258
- JMLR: http://jmlr.csail.mit.edu/proceedings/
- AAAI: http://www.aaai.org/Library/conferences-library.php
- ICCV, WACV, CVPR (vision): http://pamitc.org/
- CoNLL (nlp): http://ifarm.nl/signll/conll/