Regression

Forward selection: why/when does it work? Compare forward selection with Lasso/Ridge on a variety of linear regression settings. Understand sub-modularity.

http://stat.wharton.upenn.edu/~kord/papers/submod-stat.pdf

Bootstrap and its uses:

Bootstrap and Lasso
http://www.di.ens.fr/~fbach/fbach_bolasso_icml2008.pdf

Bag of little bootstraps https://arxiv.org/abs/1112.5016

Random Lasso http://dept.stat.lsa.umich.edu/~jizhu/pubs/Wang-AOAS11.pdf

Clustering (how to estimate k)

Choosing k for network clustering https://arxiv.org/abs/1311.2694

Learning k with AIC https://papers.nips.cc/paper/2526-learning-the-k-in-k-means.pdf

Learning k with BIC http://www.cs.cmu.edu/~dpelleg/download/xmeans.pdf

The gap statistic https://web.stanford.edu/~hastie/Papers/gap.pdf

Distance metric learning

So far we looked at Euclidian distances, what if you could learn the distance metric?

http://ai.stanford.edu/~ang/papers/nips02-metric.pdf
http://www.cs.cmu.edu/~liuy/frame_survey_v2.pdf

Classification with imbalanced clusters

http://www.jmlr.org/papers/volume8/owen07a/owen07a.pdf
http://sci2s.ugr.es/keel/pdf/specific/articulo/xue_do_2008.pdf

Topic Models

Data - 20 newsgroup data
Data - Webkb data

Naive Bayes and document clustering
http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html

LDA
http://jmlr.csail.mit.edu/papers/v3/blei03a.html

EM for document clustering
http://www.cs.umass.edu/~mccallum/papers/emcat-mlj2000.ps

Supervised topic models
http://www.cs.princeton.edu/~blei/papers/BleiMcAuliffe2007.pdf

Divide and conquer approaches:

Can you come up with methods that divide the data and combine the results to do clustering/regression on enormous datasets?

Parallel kmeans/spectral clustering
http://ntucsu.csie.ntu.edu.tw/~cjlin/papers/psc08.pdf

Parallel Lasso
http://iie.fing.edu.uy/~gmateos/pubs/dlasso/D_LASSO_TSP.pdf

Parallel clustering with core-sets
http://www.cs.princeton.edu/~yingyul/distributedClustering.pdf

Network models and inference

Blockmodels survey:

https://www.cs.umd.edu/class/spring2008/cmsc828g/Slides/block-models.pdf

Mixed membership block models:
http://jmlr.csail.mit.edu/papers/volume9/airoldi08a/airoldi08a.pdf

Spectral clustering: https://arxiv.org/abs/1007.1684

Cross validation:
http://www.tandfonline.com/doi/abs/10.1080/01621459.2016.1246365?journalCode=uasa20