Andrew Ho <kironide@gmail.com>

Plan for Week 3

2 messages

Jonah Sinick <jsinick@gmail.com> Fri, Feb 26, 2016 at 1:50 PM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Hi All,

We're finishing up the first week. I've been impressed by your progress to date!

What we've done so far

We've talked about or done work with

- **Programming in R** –

  Basic programming syntax (for loops, if-else, function syntax, optional arguments in functions).

  Data types (arrays, lists, matrices, dataframes, character vectors, numeric vectors, factors).

  Data manipulation – subsetting (with and without dplyr), lapply / sapply, joining, cbind, rbind.

- **Graphing** – Surface familiarity with the libraries ggplot2, GGally, corrplot, mosaicplot, gridExtra.

  Histograms, density plots, scatter plots, smoothed plots, arranging plots in a grid.

- **Basic statistics** – t-tests, p-values, multiple hypothesis testing, the central limit theorem

- **Ordinary least square regression** using the function lm() in R. Regression coefficients, residuals, R^2 as a measure of goodness of fit, adjusted-R^2.

- **Basic data transformations** – turning categorical variables into dummy variables (e.g. using dummy() in the lme4 package), Box-Cox transformations (and the logarithm) for dealing with skewed data, handling missing values by removing rows containing them, or filling them in with column means.

- **Regularized linear regression** for handling overfitting. L^1 and L^2 regularization schemes, using cv.glmnet to find the optimal strength of Bayesian prior on coefficients.

- **Cross validation** for estimating the generalizable predictive power of a model.

- **Logistic regression** for predicting which of two classes an example is in, using a linear model.

- **Principal component analysis** for summarizing a family of features by reducing it to a smaller number of weighted averages of them that capture a lot of the variance.

You'll be spending more time solidifying your understanding of these concepts in the coming weeks. I'd especially like to ensure that you have a clear understanding of how to interpret regression coefficients with multiple linear regression.

Remaining material

Things that we still need to talk about:

1. **Basic natural language processing** – classification using a bag of n-grams model.
2. **Recommender systems** using collaborative filtering techniques
3. **More sophisticated methods for handling missing values.**
4. **Nonlinear methods** K-nearest neighbors, Multivariate Adaptive Regression Splines random forests and boosting. Possibly Support Vector Machines (SVMs) & shallow neural nets.
5. **Clustering** – k-means & Gaussian mixtures. Latent Dirichlet Allocation in NLP
6. **Parallelizing large scale computations**. Amazon web services, multicore processing.

It's not essential that we do all of these before you start working on larger scale projects. The plan for next week is to cover 2-3 of them, and spend the rest of the time consolidating what we've done to date and/or working on a large scale project.

The order in which we cover them the above is somewhat arbitrary. Please indicate on a scale from 1-5 how interested you are in doing each one of them next week in the Excel file here.

Jonah Sinick <jsinick@gmail.com> Fri, Feb 26, 2016 at 2:59 PM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

(Just sent out a link to a version that can be edited.)

[Quoted text hidden]