0xdata

Movie Night: Data Science - Even on our Night Off

May 27, 2014

Anqi and Irene -

- Math Hackers for H2O
- Our Stats: Years on H2O: 2, Econ degrees: 2, EE degrees: 1, Paper cuts from math books: 1,237, How much we love H2O: SO much
- **Anqi** is a Math+Coder powerhouse who takes charge of H2O+R.
- **Irene** is a Pencil-and-Paper-Go-Big writer and stats geek. She handles checking the math and telling the world (H2O's support documentation)

H2O Makes This a Waterworld

- H2O does distributed analytics on big data

- Same statistics - new volumes of data
- Before H2O, it was common to spend full days waiting on models
- On a terabyte of data finishes in minutes on H2O

- Provide an interface for non-experts to build predictive models

The Untouchables: *No*, we do not replace data scientists

Q: Will H2O reduce the need for data scientists?

A: No! We want to make more people capable of analysis, and we want to enable capable analysts to conduct analysis faster

There is still a shortage of stats nerds, data scientists, and analysts.

Data Science and the Holy Grail

The ability to do analysis can’t replace the thinking person behind analysis

Example: Take a question as simple as Anqi and Irene have a night off - what movie should we go see?

Data? Yes. We need some stinkin' data.

Data is the 100000 observation data set from MovieLens

Dependent Information: Movie watcher ratings of movies on a 1 to 5 scale where 1 is dislike and 5 is like.

Independent Information: Demographics like age, gender and occupation; Movie tags (Comedy, Romance, Noir, etc...) (a movie can have more than one tag); ID/Tracking information.

WWAM: What Would Anqi Model?

Anqi believes that if we’re going to see a movie, we might as well see one we like.

Complexity doesn’t matter (go big)

Q: What movie are we most likely to enjoy?

Specification: Use all possible predictors, and let our choice of algorithm and hyper-parameters sort out importance.

Rise of the Gradients

DV = Like: derived from ratings. Ratings 4 or greater = 1, otherwise 0.

IV = All other predictors (omit information like movie title and user ID - those aren't useful)

It's likely that there are interaction effects. Comedy might score equally well among women and men, but Romantic Comedies might be well liked by women only.

Rise of the Gradients: The Sequel

The best tool for the job is GBM classification;

- our DV is 0/1 - two classes - easy enough

- we want to control for interactions, but we don't want to spend a lot of time figuring out that comedy and film-noir don't interact at all, but that romance and horror do.

- GBM allows us to specify complex interaction terms by controlling the depth of the model - which we will run now.

Friday Night GLMites

Irene thinks that a binomial regression will be more interpretable, so if there are two movies on the cusp we can choose better.

She cares more about avoiding movies that are going to be really awful, so here we'll pay more attention to how well the model predicts 0's.

Friday Night II

DV = Irene used the same DV as Anqi, but we care more now that we're correctly predicting 0s.

IV = All other predictors minus indicators.

While there are probably interaction effects, she wants to aim for quick interpretability. Regularization can help isolate the unique contributions of each piece of information, so binomial GLM is the best tool for the job.

Anqi's Predictions	*Irene's Predictions*
Monument Men *Her*	*Muppets* *Her*

Lights Out - Some Closing Points

We didn't address a serious problem here - but this does reflect the kinds of choices that arise in an applied setting.

H2O helped us make a decision, but it didn’t make a decision for us.

How different is the question that Anqi answered than the question Irene answered? Which model is going to be most useful for predicting preferences if we use it for the next year?

The Long Goodbye

We're giving a similar talk next week at MLconf.

We were invited to talk specifically as an example of women who are working in Data Science and Machine Learning.

On that note - 75% of the Data Science team at H2O are women. Nationally, women make up only 27% of Computer Scientists and 19% of Mathematicians.

H2O isn't just democratizing data science as a marketing gimmick - our executives have tried hard to put their money where their mouth is.

Stories change people, while statistics gives them something to argue about

- Bernie Siegel