Summary of COS424

David M. Blei

May 3, 2012

- Data are everywhere. We need tools to solve problems with data.
  - Making predictions about what will happen / what happened
  - Building better tools to explore data
- Data are complicated and there is a lot of it.
  - We want to be able to express what we know about the data
  - We need to be able to compute with large data sets
- We have started to learn about this field—statistics, machine learning, data science, data mining—that tries to address these problems.
- My goal today: Review what we have learned, make connections between the ideas, highlight some of the recurring themes.
- Our syllabus was divided into four sections
  - Classification: For example, spam filtering.
  - Clustering: Grouping documents/images/genes into groups
  - Prediction: Forming real-valued estimates of the future
  - Dimension reduction: Finding patterns, e.g., collaborative filtering
- But today I want to talk about the general process of doing data anlaysis. We alluded to this when we talked about topic models.

1. Form assumptions about your data
2. Learn about your data using those assumptions
3. Use what you learned to do things, like predict and explore

1 Forming assumptions

- *Probability models are a language for assumptions about data.*

- What are the probability models we have learned about?
  - Exponential families
  - Naive Bayes classification
  - Probabilistic classification
  - Mixtures of Gaussians, Multinomials, anything
  - Hidden markov models
  - Mixed-membership models
  - Linear regression, logistic regression, generalized linear models
  - Principal component analysis, factor analysis, NMF
  - Kalman filters

- Theme: Models generalize through the data generating distribution.
  - Naive Bayes classification to probabilistic classification
  - Mixtures of Gaussians to mixtures of anything
  - Linear regression to generalized linear models
  - PCA/FA to NMF (and latent Dirichlet allocation)

  When you read about a new model, think about what other kinds of data you can apply it to. When you encounter new data, think about how you can embed it in existing models.

- Notice the role of the exponential family. We can adapt our models to many kinds of observations—continuous, discrete, categorical, multivariate, ordinal.

- Theme: Models can be connected and combined
  - Mixture models in sequence give us hidden Markov models
  - Factor analyzers in sequence give us Kalman filters
  - LDA models in sequence give us dynamic topic models

- Theme: Regularization
  - Helps us moderate bias/variance trade-off
  - Can give sparse solutions, for interpretability and computation
  - Relates to *Bayesian models*, e.g., smoothed multinomial probabilities and regression models with priors.

- Theme: Graphical models
  - A visual language of probabilistic assumptions.
  - Encourage generalization, modularity
  - Connect assumptions to step #2, algorithms.

2 Learning about data

- *The problem of how to discover patterns in data—also called "learning"—can be formulated as an optimization of an appropriate objective function.*

- We focused on *maximum likelihood estimation*.
  - The model is indexed by parameters.
  - The observed data have a probability under each setting.
  - Find the parameters that make the data most likely—exact or gradient-based.

- Other objectives are related to likelihoods under a model.
  - k-means
  - PCA: maximizing variance and minimizing reconstruction error

  - Traditional regression, i.e.., minimizing the sum of squares)

- We also discussed *regularized likelihood*, which related to MAP estimation in a Bayesian set-up.

- The EM algorithm let us cope with hidden quantities in a general way, via coordinate ascent of the likelihood function.
  - Compute the conditional distribution of hidden variables
  - Expected complete likelihoods are easier than marginal likelihoods.

- In modern data analysis we need to think about scale.
  - Map-reduce: Take advantage of clusters of computers
  - Stochastic optimization: Repeatedly subsample the data

  These methods take advantage of the special structure of the objective function.

- This perspective—learning as optimization—is also prevelant in *approximate posterior inference*. This is an advanced subject we didn't discuss.

- Inference is important when we don't differentiate between estimating parameters and computing conditionals of hidden variables. (Or, if you insist on differentiating, think about posterior inference as being "local learning", e.g., in the EM algorithm.)

- Variational inference and MCMC both can be construed as optimizations. (E.g., applying stochastic optimization to variational inference lets us learn topic models with millions of documents.)

3 Predict, explore, check

- You have made assumptions and fit a model. What next?

- *With a fitted model in hand, solve your problem with a probabilistic computation that uses the model and some input.*

- Solve your problem:
  - Cast what you are trying to do in terms of a predictive quantity
  - Cast that quantity as a probabilistic computation

- Examples:
  - Classification: The maximum probability class
  - Speech recognition: The maximum probability sequence
  - Target tracking: The expected location at the next time point
  - Collaborative filtering: The expected ratings of unrated movies
  - Exploration: A clustering of the data to form a navigator of it

- Evaluation based on cross validation is important to predict how well your model will do on future data.

- Other evaluations could be used: human studies, money earned

- Note: Causal inference is difficult from observational data. Finding likely patterns of errors is less difficult.

- Error analysis is important
  - Where did my method go wrong?
  - Are there patterns to this?
  - Does this suggest directions of improvement?

  Note the iterative nature of this process—now we go back to step #1.

4 Learn more

- Take classes

  - Foundations of probabilistic models (COS513A and B)
  - Courses in CS, EE, ORFE, MoBio, Political Science, others

- Join the ml-stat-talks mailing list

- Go to graduate school, or don't go to graduate school

- Read
  - ESL, PRML, others
  - Conference proceedings: ICML, NIPS, EMNLP, CVPR, ...
  - Journals: JMLR, JASA, AAS, AS, BA, JCGS, ...

- Continue to play with data and solve problems.