Exam Review Session

William Cohen
What can I tell you about the exam?

• It’s qualitatively similar to the midterm
• Sample/old exams for 10-601 are probably pretty representative
• About 80% is on topics from after the midterm
• Questions might overlap with HWs
• Everyone can drop one HW but....
  – You should read through the HW you skipped carefully and
  – be familiar with the algorithms implemented
  – understand the questions
• Look over the review points in the lecture pages from the wiki
General hints in exam taking

• You **can** bring in one 8 ½ by 11” sheet (front and back)
• Look over everything quickly and **skip around**
  – probably nobody will know everything on the test
  – if it seems really complicated you’re likely missing something
• If you’re not sure what we’ve got in mind: state your **assumptions** clearly in your answer.
  – There’s room for this even on true/false
• If you look at a question and don’t know the answer:
  – we probably haven’t told you the answer
  – but we’ve told you enough to work it out
  – imagine arguing for some answer and see if you like it
Post mid-term topics

• Clustering
• Active learning
• Semi-supervised learning
• Graphical models
  – HMMs and sequential data
  – Topic models
• Deep learning
• PCA and matrix factorization
• Reinforcement learning
Post mid-term topics: clustering

What You Should Know Afterward

- Partitional Clustering. k-means and k-means ++
  - Lloyd’s method
  - Initialization techniques (random, furthest traversal, k-means++)
- Hierarchical Clustering.
  - Single linkage, Complete linkage

– You should understand the definitions and algorithmic outlines of the hierarchical methods as well
Post mid-term topics: clustering

• Clustering (3/2)
  – Later (3/28) we came back and looked EM as learning method for a “soft version” of k-means: _______
  – You should know how EM and k-means relate

What You Should Know Afterward

- Partitional Clustering. k-means and k-means ++
  - Lloyd’s method
  - Initialization techniques (random, furthest traversal, k-means++)
- Hierarchical Clustering.
  - Single linkage, Complete linkage
**Post mid-term topics: active learning**

- Understand the key points and the examples in the lecture
  - what’s the formal model (the API) for an active learner?
  - how can active learning be used to distinguish between two hypothesis that sometimes agree and sometimes disagree?
  - what kinds of queries does an active learner make? how do you design an algorithm to use those queries effectively? how do implement binary-search like algorithms for different types of concept classes?
Post mid-term topics: SSL

• Understand the key points and the examples in the lecture
  – what’s the formal model (the API) for SSL?

• Remember we also came back to this topic later:
  – SSL with naïve Bayes and k-means variants using EM and DGMs

• Free advice: there’s nothing on graph-based SSL on the exam
Post mid-term topics: DGMs

• Boy, did we talk about these a lot!
Post mid-term topics: DGMs

• Semantics
  – How does a graph encode a joint probability?
  – How compact (how many parameters) is a DGM compared to a naïve encoding?
  – How do you determine conditional independence in a DGM?
  – How do you learn parameters for a DGM? When do you use counts and MAPs and when do you need to use EM?
Post mid-term topics: DGMs

• Lots of (sadly important) terminology
  – blocking, d-separation, explaining away, Markov blanket, message passing, belief propagation, polytree, ...
  – can you define these? can you recognize them or draw them?
Post mid-term topics: DGMs

• DGMS for describing learning algorithms
  – parameters in a DGM
  – inference and learning
  – what does the DGM for ___ look like, and what do the individual variables mean?
    • supervised naïve Bayes?
    • mixture of Gaussians?
    • HMMs?
    • LDA?
Post mid-term topics: DGMs

• HMMs
  – how are they like (?) DGMs
  – how do you do inference and learning?
  – what are they useful for?

• Free advice: there are no questions on CRFs on the final
Post mid-term topics: DGMs

• DGMS and topic models
  – what are topic models? what sort of DGMs can be used model text?
  – what is Gibbs sampling, and when would you use it?

  – connection: belief propagation: what is it and when do you use it? when would you prefer Gibbs sampling to BP?
Post mid-term topics: deep learning

• Lots of terminology & equations:
  – vanishing gradients, saturation, reLU, ...
  – don’t spend too much time memorizing these they will all be different in a year or two

• Some architectures I went in to more details on with pictures and such
  – LSTMs
  – CNNs (which was a HW)
  – You should definitely understand how these are supposed to work and what the different pieces are
PCA and MF and CF

• What is PCA? what do those pictures mean?
• How do the different PCA vectors relate to each other?
  – are they always orthogonal?
• If I take data in N dimensions and re-express it in the first K dimensions formed by PCA, what does the new data look like?
• How are PCA and MF related?
• What are the algorithms used for PCA and for MF?
• What is collaborative filtering and how is MF used for it? E.g., what M gets F-ed?
PCA and MF

• What is PCA? what do those pictures mean?
• How do the different PCA vectors relate to each other?
  – are they always orthogonal?
• If I take data in N dimensions and re-express it in the first K dimensions formed by PCA, what does the new data look like?
• How are PCA and MF related?
• What are the algorithms used for PCA and for MF?
Post mid-term topics

• Clustering ✔
• Active learning ✔
• Semi-supervised learning ✔
• Graphical models ✔
  – HMMs and sequential data
  – Topic models
• Deep learning ✔
• PCA and matrix factorization ✔
• Reinforcement learning
Final notes

• I’m having office hours this week Tuesday at 11am not Thursday at 11am

• Good luck!
Sample questions

4. The EM procedure is guaranteed to converge to a global maximum.

6. When using Principle Components Analysis, the top $n$ Principle Components have the following property: they provide a lower-dimensional representation of the original data, and one that allows reconstructing the original data with the minimum sum of squared errors.

7. The Naive Bayes classifier is a special case of a Bayes Net.

8. The Hidden Markov Model is a special case of a Bayes Net.

9. When trained on the same data, Naive Bayes must estimate more parameters than Logistic Regression.
Sample questions

1. List some pairs $X,Y$ so that $I<X,A,Y>$

2. How many parameters are needed to define the joint?
Sample questions

Write down the joint distribution $P(X_1,\ldots,X_7)$
Sample Questions

Complete the diagram for:
• N documents of length L
• K topics
• \(Z_{ij}\) topic indicator for doc \(i\), position \(j\)
• \(W_{ij}\) word for doc \(i\), position \(j\)
• \(\alpha\) prior on topic multinomial
• \(\theta_i\) topic distribution for doc \(i\)
• \(\beta_i\) prior on work multinomials
• \(\phi_k\) word multinomial for topic \(k\)

Which variables have a discrete domain?
Sample questions

3. In this new basis you found above, (that is, reconstructing the points using only the first principal component), list, in order from closest to furthest, the three neighbors (A, B, C) of Q. Be sure to clearly note any ties:

Repeat for the second principle component
Sample questions

Repeat for the second principle component