CSC412 – Example Final Test

April 12, 2002
Time: 110 minutes
Total Points: 125
Worth: 25%

Name: Student Number:

1 Conjugate Priors

Consider performing Bayesian parameter estimation from a finite data sample $\mathcal{D} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$

- Assume the data $\mathbf{x}$ are real-valued and we want to use a Gaussian model (normal distribution) with fixed, known covariance $\mathbf{S}$ to model them:
  $$p(\mathbf{x}|\mu, \mathbf{S}) = \mathcal{N}(\mathbf{x}; \mu, \mathbf{S})$$
  If we use a Gaussian prior
  $$p(\mu|\alpha, \mathbf{Q}) = \mathcal{N}(\mu; \alpha, \mathbf{Q})$$
  on the mean parameter of the normal distribution, calculate the posterior over the mean $p(\mu|\mathcal{D}, \alpha, \mathbf{Q})$.
- Assume that the data x are non-negative integer counts and we want to use Poisson model with rate λ for them:
  $$p(x|\lambda) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}$$
  If we use a gamma prior on the rate parameter
  $$p(\lambda|s, c) = \frac{1}{sG(c)} \left(\frac{\lambda}{s}\right)^{c-1} \exp\left\{-\frac{\lambda}{s}\right\}$$
  compute the posterior $p(\lambda|\mathcal{D}, s, c)$ and show it is also gamma.
- Harder: Compute the predictive distibution $p(\mathbf{x}^{new}|\mathcal{D})$.

2 String-Edit HMMs

- Why are string-edit HMMs examples of “linear” or “left-to-right” HMMs?
- In a string-edit HMM, how many states are there for a template pattern of length M?
- How many nonzero entries are there (at most) in the transtion matrix?
- Assume symbols are emitted only upon arrival into match and insert states. In a dynamic programming grid for a string-edit HMM, what would be the cost (in negative log probability) of moving from state i at position t in the test sequence to state j at position $t+1$ in the test sequence:
(a) if state j is a match or an insert state
(b) if state j is a delete state

3 Junction Trees and the Hugin Algorithm

- Starting with an undirected graph, what are the steps to build a junction tree?
- Not all graphs have junction trees.
What is an equivalent graph property to having junction tree?
Give a definition of this property in terms of the original graph.
- Give the basic message passing equations in the Hugin algorithm when a clique V in the junction tree passes a message to a clique W.
Show that this message leaves the two cliques in agreement over the marginal distribution over their separator variables.
Show also that this message leaves the joint distribution over the two cliques unchanged.
- Harder: using the HMM as an example, explain why one might want to run the junction tree message passing algorithm in the absence of evidence.

4 Drawing Graphical Models

Convert the following generative model descriptions into a graphical model diagram using plates when necessary.

- A rate λ is drawn from a distribution $p(\lambda)$ and then using that rate, N counts i are drawn from a distribution $p(i|\lambda)$.
- A global covariance Σ is drawn from a distribution $p(\Sigma)$. Then, for each of K clusters, a cluster mean μ_k is drawn from a distribution p(μ), a number of data points n_k is drawn from a distribution $p(n)$ and then n_k points are drawn from a distribution $p(\mathbf{x}|\mu_k, \Sigma, n_k)$. All clusters have the same covariance.
- Same as above, except that each cluster has a separate covariance drawn from a distribution $p(\Sigma_k|\mu_k)$.

Alternate kind of question: Describe the generative model implied by the following graphical models.

5 Gradients of Feature-Based models

Consider a feature-based maximum-entropy classifier which uses the following classification rule to classify an input $\mathbf{x}$ in terms of its feature vector $\mathbf{f}(\mathbf{x})$:

$$p(c|\mathbf{x}) = \frac{\exp\{\mathbf{w}_c^\top \mathbf{f}(\mathbf{x})\}}{\sum_k \exp\{\mathbf{w}_k^\top \mathbf{f}(\mathbf{x})\}}$$

For a finite data sample $\mathcal{D} = \{\mathbf{x}^1, c^1, \mathbf{x}^2, c^2, \ldots, \mathbf{x}^N, c^N\}$ define the objective function $\ell(\mathcal{D}) = \sum_n \log p(c^n|\mathbf{x})$

- Compute the gradient $\frac{\partial \ell}{\partial \mathbf{w}_c}$
- For binary features, the components $\mathbf{f}_i(\mathbf{x})$ of the feature vector are either 0 or 1.
  What happens to the weight $\mathbf{w}_i$ if, using binary features, on the finite sample available:
  A certain component $\mathbf{f}_i$ is always 1?
  A certain component $\mathbf{f}_i$ is always 0?
  What happens if two components $\mathbf{f}_i$ and $\mathbf{f}_j$ are always the same?

6 Linear Regression

In a fully-observed linear regression model with Gaussian noise, we fit the observed outputs $\mathbf{y}$ as a function of the observed inputs $\mathbf{x}$. The model is a Gaussian density on $\mathbf{y}$ whose mean is a linear function of $\mathbf{x}$:

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mathbf{W}\mathbf{x} + \mathbf{w}_0, \Sigma)$$

- Write down the likelihood function of a finite training set $\{\mathbf{x}^1, \mathbf{y}^1, \mathbf{x}^2, \mathbf{y}^2, \ldots, \mathbf{x}^N, \mathbf{y}^N\}$ in terms of the parameters $\mathbf{W}$, $\mathbf{w}_0$ and Σ.
- Assuming the covariance matrix Σ is known, Solve for the maximum likelihood parameter estimates $\mathbf{W}^*$ and $\mathbf{w}_0^*$.

7 EM Algorithm Proof

Consider the familiar latent variable model setup for some data $\mathbf{x}$ with hidden variables $\mathbf{z}$ and parameters θ:

$$p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta)$$

In the EM algorithm, we iteratively update the parameters using two phases:
E-step: $q^{t+1}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \theta^t)$
M-step: $\theta^{t+1} = \text{argmax}_\theta \sum_{\mathbf{z}} q^{t+1}(\mathbf{z}) \log p(\mathbf{x}, \mathbf{z}|\theta^t)$
Prove that these iterations always either increases the likelihood of the data $\mathbf{x}$ or keep it the same.

8 QMR-DT

In the quick medical reference (QMR) graphical model, there are unobserved binary nodes representing diseases and partially observed binary nodes representing findings.

- How many disease nodes are there?
  How many findings are there?
- What is the model used for the joint probability of the findings and the diseases?
- Where did the parameters of the network come from?
- How does the complexity of inference scale with the total number of diseases/findings?
  number of observed findings?
  number of positive findings?