Formative Assignments

Probabilistic and Unsupervised Learning

Peter Orbanz

These problems will not be marked, but you should attempt to solve them before they are discussed in the tutorial.

1. **[28 marks] Statistics and Distributions**.

   Prepare for: Tutorial on 20 October 2021

   In the coming weeks we will be making extensive use of the following distributions, all of which belong to the exponential family. For each of these distributions, find:

   (a) The standard exponential form, identifying the natural parameters in terms of the conventional parameters given in the table (i.e. the function $\phi(\theta)$), and the sufficient statistic (i.e. $\mathsf{T}(x)$).

   (b) The expected value of the sufficient statistics in terms of the natural or conventional parameters (i.e. $\langle \mathsf{T}(x) \rangle_{p(x|\theta)}$). These expectations are often called the "mean" or "moment" parameters of the distribution. [Note: show your derivation of the expectations; don't just look them up.]

   The distributions to consider are:

Name	Domain	Symbol	Density or Probability fn
Multivariate Normal	$\mathbb{R}^D$	$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$	$\lvert 2\pi\Sigma\rvert^{-1/2}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
Binomial	$\mathbb{Z}_{0-N}$	$x \sim \mathsf{Binom}(p)$	$\binom{N}{x} p^x (1-p)^{N-x}$
Multinomial	$[\mathbb{Z}_{0-N}]^D$	$\mathbf{x} \sim \mathsf{Multinom}(\mathbf{p})$	$\dfrac{N!}{x_1!\, x_2! \ldots x_D!} \prod_{d=1}^{D} p_d^{x_d}$
Poisson	$\mathbb{Z}_{0+}$	$x \sim \mathsf{Poisson}(\mu)$	$\mu^x e^{-\mu}/x!$
Beta	$[0,1]$	$x \sim \mathsf{Beta}(\alpha, \beta)$	$\dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Gamma	$\mathbb{R}_+$	$x \sim \mathsf{Gamma}(\alpha, \beta)$	$\dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Dirichlet	$[0,1]^D$	$\mathbf{x} \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$	$\dfrac{\Gamma(\sum_{d=1}^{D} \alpha_d)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1}$

   [4 marks each]

2. **[5 marks] Basic spectral properties**.

   Prepare for: 20 October 2021

   Let A be a symmetric $n \times n$-matrix, with eigenvalues $\lambda_1, \ldots, \lambda_n$.

   (a) Show that the matrix $B = A + cI$, where I is the identity matrix and $c \in \mathbb{R}$, has eigenvalues $\lambda_1 + c, \ldots, \lambda_n + c$. [3 marks]

   (b) Suppose v and w are eigenvectors of A, with the same eigenvalue λ. Show that any linear combination of v and w is again an eigenvector of A. What is its eigenvalue? [2 marks]

3. **[7 marks] ML in the exponential family.**

   Prepare for: 27 October 2021

   Express the maximum-likelihood value of the *mean* parameters (as defined in the question above) of the general exponential family distribution

   $$p(\mathbf{x}|\theta) = g(\theta)f(\mathbf{x})e^{\theta^{\mathsf{T}}\mathsf{T}(\mathbf{x})}$$

   as a function of a data set of iid observations $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$.

4. **[20 marks] Modelling Data.**

   Prepare for: 27 October 2021

   (a) Download the data file called `geyser.txt` from the course web site. This is a sequence of 295 consecutive measurements of two variables from Old Faithful geyser in Yellowstone National Park: the duration of the current eruption in minutes (rounded to the nearest second), and the waiting time since the last eruption in minutes (to the nearest minute). Examine the data by plotting the variables within (`plot(geyser(:,1),geyser(:,2),'o');`) and between (e.g. `plot(geyser(1:end-n,1),geyser(n+1:end,1 or 2),'o');` for various `n`) time steps. Discuss and justify based on your observations what kind of model might be most appropriate for this data set. Consider each of the models we have encountered in the course through week 3: a multivariate normal, a mixture of Gaussians, a Markov chain, a hidden Markov model, an observed stochastic linear dynamical system and a linear-Gaussian state-space model. Can you guess how many discrete states or (continuous) latent dimensions the model might have? [10 marks]

   (b) Consider a data set consisting of the following string of 160 symbols from the alphabet $\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$:

   ```
   AABBBACABBBACAAAAAAAAABBBACAAAAABACAAAAAABBBBACAAAAAAAAAAAABACABACAABBACAAABBBBA
   CAAABACAAAABACAABACAAABBACAAAABBBBACABBACAAAAAABACABACAAABACAABBBACAAAABACABBACA
   ```

   Study this string and suggest the structure of an HMM model that may have generated it. Specify the number of hidden states in the HMM, the transition matrix with any constraints and estimates of the transition probabilities and the output or emission matrix probalities, and the intial state probabilities. You need to provide some description/justification for how you arrived at these numbers. We do **not** expect you to implement the Baum-Welch algorithm—you should be able to answer this question just by examining the sequence carefully. [10 marks]

5. **[15 marks] Zero-temperature EM.**

   Prepare for: 27 October 2021

   In the automatic speech recognition community HMMs are sometimes trained by using the Viterbi algorithm in place of the forward–backward algorithm. In other words, in the E step of EM (Baum–Welch), instead of computing the expected sufficient statistics from the posterior distribution over hidden states: $p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}, \theta)$, the sufficient statistics are computed using the single *most probable* hidden state sequence: $\mathbf{s}^*_{1:T} = \arg\max_{\mathbf{s}_{1:T}} p(\mathbf{s}_{1:T}|\mathbf{x}_{1:T}, \theta)$.

   (a) Is this algorithm guaranteed to converge (in the sense that the free-energy or some other reaches an asymptote)? To answer this you might want to consider the discussion of the real EM algorithm and what happens if we constrain $q(s)$ to put all its mass on one setting of the hidden variables. Support your arguments. [10 marks]

   (b) If it converges, will it converge to a maximum of the likelihood? If it does not converge what will happen? Support your arguments. [5 marks]

   (c) [Bonus (just for culture)] Why do you think this question is labelled "Zero-temperature EM" Hint: think about where temperature would appear in the the free-energy. [no marks]

6. **[35 points] Deriving Gibbs Sampling for LDA.**

Prepare for: 3 November 2021

In this question we derive two Gibbs sampling algorithms for latent Dirichlet allocation (LDA). LDA is a topic model that defines multiple mixtures of discrete distributions with shared components. The archetypical application is to words in documents. Suppose there are W possible words, D documents and K topics. The LDA model specifies the distribution of the ith word in the dth document, $x_{id} \in \{1 \dots W\}$, in terms of the hyperparameters α and β, by way of latent Dirichlet parameters:

$$\begin{aligned}
&\text{topic distribution for } d\text{th document} && \boldsymbol{\theta}_d | \alpha \sim \text{Dirichlet}(\alpha, \dots, \alpha) && (1)\\
&\text{word distribution for } k\text{th topic} && \boldsymbol{\phi}_k | \beta \sim \text{Dirichlet}(\beta, \dots, \beta) && (2)\\
&\text{topic for } i\text{th word in } d\text{th document} && z_{id} | \boldsymbol{\theta}_d \sim \text{Discrete}(\boldsymbol{\theta}_d) && (3)\\
&\text{identity of } i\text{th word in } d\text{th document} && x_{id} | z_{id}, \boldsymbol{\phi}_{z_{id}} \sim \text{Discrete}(\boldsymbol{\phi}_{z_{id}}) && (4)
\end{aligned}$$

Let $A_{dk} = \sum_i \delta(z_{id} = k)$ be the number of z_{id} variables taking on value k in document d, and $B_{kw} = \sum_d \sum_i \delta(x_{id} = w)\delta(z_{id} = k)$ be the number of times word w is assigned to topic k across all the documents. Let N_d be the total number of words in document d and let $M_k = \sum_w B_{kw}$ be the total number of words assigned to topic k.

(a) Write down the joint probability over the observed data and latent variables, expressing the joint probability in terms of the counts N_d, M_k, A_{dk}, and B_{kw}. [5 points]

(b) Derive the Gibbs sampling updates for all the latent variables z_{id} and parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$. [10 points]

(c) Integrate out all the parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ from the joint probability in (a), resulting in a joint probability over only the z_{id} topic assignment variables and x_{id} observed variables. Again this expression should relate to z_{id}'s and x_{id}'s only through the counts N_d, M_k, A_{dk}, and B_{kw}. [5 points]

(d) Derive the Gibbs sampling updates for the z_{id} with all parameters integrated out. This is called **collapsed Gibbs sampling**. You will need the the following identity of the Gamma function: $\Gamma(1 + x) = x\Gamma(x)$ for $x > 0$. [10 points]

(e) What hyperpriors would you give to α and β? How would you generate samples of α and β from the appropriate conditionals? [You should suggest an algorithm and justify its feasibility, but do not need to derive the update equations; 5 points]