Statistical Machine Learning (W4400)
Spring 2016
https://courseworks.columbia.edu

John P. Cunningham
jpc2181
Ben Reddy, Phyllis Wan, Ashutosh Nanda
bmr2136, pw2348, an2655

Homework 5

Due: Thursday 28 April 2016

Students are encouraged to work together, but homework write-ups must be done individually and must be entirely the author's own work. Homework is due at the **beginning** of the class for which it is due. **Late homework will not be accepted under any circumstances.** To receive full credit, students must thoroughly explain how they arrived at their solutions and include the following information on their homeworks: name, UNI, homework number (e.g., HW03), and class (STAT W4400). All homework must be turned in online through Courseworks in PDF format, have a .pdf extension (not zip or other archive!), and be less than 4MB. If programming is part of the assignment, the code must be turned in in one or more .R files. Homeworks not adhering to these requirements will receive no credit. For your convenience (not required), a tex template for producing nice PDF files can be found on courseworks.

1. **Bit streams, with and without memory** (50 points)
   I wish to transmit information via ordered sequences of n bits with binary random variables $X_i \in \{0, 1\}$. Here we consider distributions over those sequences. To answer the following questions, please recall it is typical to calculate entropy using $\log_2$. Furthermore, the chain rule for entropy may be useful:

$$H\left(X_1, ..., X_n\right) = \sum_{i=1}^{n} H\left(X_i | X_{i-1}, ..., X_1\right).$$

   (a) (4 points) Among all probability distributions on the finite set $\{1, 2, 4, 8\}$, which has/have the largest entropy?

   (b) (4 points) Among all probability distributions on the finite set $\{1, 2, 4, 8\}$, which has/have the smallest entropy?

   (c) (5 points) Of all distributions over ordered sequences of n bits, one achieves maximum entropy. What is this maximum entropy value?

   (d) (5 points) Fully specify the Markov chain that produces sequences from this maximum entropy distribution. Hint: a graphical model is neither necessary nor sufficient.

   (e) (2 points) Sometimes, due to resource constraints (average power used, for example), the bit sequence must have memory. In this example, say that if a 1 is transmitted, then the next bit transmitted must be a 1 with probability 0.7; similarly, a transmitted 0 means the next bit must be a 0 with probability 0.6. Does this chain have higher or lower entropy than the chain above?

   (f) (12 points) What is the equilibrium distribution P_{eq} of the chain described in part (e)?

   (g) (6 points) If we draw $X_1 \sim F$, where F is an arbitrary distribution on $\{0, 1\}$, what is the probability $\lim_{n\to\infty} P(X_n = 1)$? What is $\lim_{n\to\infty} P(X_n = 0)$?

   (h) (12 points) If we draw $X_1 \sim P_{eq}$, what is the entropy of the ordered sequences of n bits that are drawn from this Markov chain?

2. **Conjugacy** (50 points)

Suppose observations $X_1, X_2, \ldots$ are recorded. We assume these to be conditionally independent and exponentially distributed given a parameter θ:

$$X_i \sim \text{Exponential}(\theta),$$

for all $i = 1, \ldots, n$. The exponential distribution is controlled by one *rate parameter* $\theta > 0$, and its density is

$$p\,(x; \theta) = \theta e^{-\theta x}$$

for $x \in \mathbb{R}_+$.

(a) (5 points) Plot the graph of $p\,(x; \theta)$ for $\theta = 1$ in the interval $x \in [0, 4]$.

(b) (5 points) What is the visual representation of the likelihood of individual data points? Draw it into the graph above for the samples in a toy dataset $\mathcal{X} = \{1, 2, 4\}$ and $\theta = 1$. How is the likehood of this toy dataset related to that of the individual data points?

(c) (4 points) Would a higher rate (e.g. $\theta = 2$) increase or decrease the likelihood for the toy data set?

(d) (10 points) For the next parts, we introduce a prior distribution $q(\theta)$ for the parameter. Our objective is to compute the posterior. In general, that requires computation of the evidence as the integral

$$p(x_1, \ldots, x_n) = \int_{\mathbb{R}_+} \Big(\prod_{i=1}^{n} p(x_i|\theta) \Big) q(\theta) d\theta \ .$$

We will not have to compute the integral in the following, since we choose a prior that is conjugate to the exponential.

The natural conjugate prior for the exponential distribution is the gamma distribution:

$$q(\theta|\alpha, \beta) = \theta^{\alpha-1} \frac{\beta^\alpha e^{-\beta\theta}}{\Gamma(\alpha)}$$

for $\theta \geq 0$ and $\alpha, \beta > 0$. We have already encountered this distribution in an earlier homework problem (where we computed its maximum likelihood estimator), and you will notice that we are using a different parametrization of the gamma density here.

The problem: first take a moment to convince yourself that the exponential and gamma distributions are exponential family models. Show that, if the data is exponentially distributed as above with a gamma prior

$$q(\theta) = \text{Gamma}(\alpha_0, \beta_0) \ ,$$

the posterior is again a gamma, and find the formula for the posterior parameters. (In other words, adapt the computation we performed in class for general exponential families to the specific case of the exponential/gamma model.) In detail:

- Ignore multiplicative constants and normalization terms, such as the evidence term in Bayes' formula.
- Show that the posterior is proportional to a gamma distribution.
- Deduce the parameters by comparing your result for the posterior to the definition of the gamma distribution.

(e) (10 points) Machine learning problems are often *online problems*, where each data point has to be processed immediately when it is recorded (as opposed to *batch problems*, where the entire data set is recorded first and then processed as a whole). Conjugate priors are particularly useful for online problems, since, roughly speaking, the posterior given the first $(n-1)$ observations can be used as a prior for processing the nth observation:

The problem: Show that, if $p(x|\theta)$ is an exponential family model and $q(\theta)$ its natural conjugate prior, the posterior $\Pi(\theta|x_{1:n})$ under n observations can be computed as the posterior given a single observation x_n using the prior $\tilde{q}(\theta) := \Pi(\theta|x_{1:n-1})$.

(f) (8 points) For the specific case of the exponential/gamma model, give the formula for the parameters (α_n, β_n) of the posterior $\Pi(\theta|x_{1:n}, \alpha_0, \beta_0)$ as a function of $(\alpha_{n-1}, \beta_{n-1})$.

(g) (8 points) Visualize the gradual change of shape of the posterior $\Pi(\theta|x_{1:n}, \alpha_0, \beta_0)$ with increasing n:

- Generate $n = 256$ exponentially distributed samples with parameter $\theta = 1$.
- Use the values $\alpha_0 = 2, \beta_0 = 0.2$ for the hyperparameters of the prior.
- Visualize the updated posterior distribution after $n = \{4, 8, 16, 256\}$, in the range $\theta \in [0, 4]$. Plot all curves into the same figure and label each curve.
  Hint: The gamma function Γ, which occurs in the definition of the gamma density, is implemented in R as `gamma`. When you have to compute a product over several data points, you might run into numerical problems with this function. One possible workaround to first compute the log-likelihood and then take its exponential $\exp(\log(p(x_{1:n}; \alpha, \beta)))$. The logarithm of the gamma function is implemented in R as a separate function `lgamma`.
- Comment on the behavior of the posterior distribution as n increases.