Statistical Machine Learning (W4400)
Spring 2014
http://stat.columbia.edu/∼porbanz/teaching/W4400/

Peter Orbanz
porbanz@stat.columbia.edu
Lu Meng
lumeng@stat.columbia.edu
Jingjing Zou
jingjing@stat.columbia.edu

Homework 5

Due: 1 May 2014

Homework submission: We will collect your homework **at the beginning of class** on the due date. If you cannot attend class that day, you can leave your solution in my postbox in the Department of Statistics, 10th floor SSW, at any time before then.

Problem 1 (Bayesian inference)

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

1. Plot the graph of $p(x;\theta)$ for $\theta = 1$ in the interval $x \in [0, 4]$.
2. 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?
3. Would a higher rate (e.g. $\theta = 2$) increase or decrease the likelihood for the toy data set?

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.

Question 1. 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.

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:

Question 2.

a. 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 $\tilde{q}(\theta) := \Pi(\theta|x_{1:n-1})$.

b. 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 $(\alpha_{n-1}, \beta_{n-1})$.

c. 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.

Question 3. Finally, we will show that the maximum a posteriori estimator

$$\theta_n^{\mathsf{MAP}} := \arg\max_{\theta} p\left(\theta|x_{1:n}\right)$$

asymptotically agrees with the ML estimator, which for the rate parameter of the exponential distribution is

$$\theta_n^{\mathsf{ML}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} \,.$$

a. Show that the gamma density attains its maximum at $x = \frac{\alpha-1}{\beta}$.

b. Plug in the values α_n and β_n that you have obtained for the posterior distribution to obtain θ_n^{MAP}.

c. Compare θ_n^{MAP} and θ_n^{ML} in the limit $n \to \infty$.