Homework 4

Handed out: Saturday, March 02, 2019

Due: Monday, March 18, 2019 11:55 pm

Notes:

- We *highly* encourage typed (Latex or Word) homework. Compile as single report containing solutions, derivations, figures, etc.
- Submit all files including report pdf, report source files (e.g. .tex or .docx files), data, figures produced by computer codes and programs files (e.g. .py or .m files) in a **.zip** folder. Programs should include a Readme file with instructions on how to run your computer programs.
- Zipped folder should be turned in on Sakai with the following naming scheme: **HW4_LastName_FirstName.zip**
- Collaboration is encouraged however all submitted reports, programs, figures, etc. should be an individual student's writeup. Direct copying could be considered cheating.
- Homework problems that simply provide computer outputs with no technical discussion, Algorithms, etc. will receive no credit.
- Software resources set can be downloaded from this link.

1 Exponential family

A. The exponential family of distributions over y given parameter θ is given as

$$p\left(y|\theta\right) = h\left(y\right)\exp\left\{y\theta - A\left(\theta\right)\right\} \tag{1}$$

Present each of the following distributions in the exponential family form. Identify the relevant components necessary for use in a GLM: (i) the canonical parameter θ, (ii) $h\left(y\right)$, (iii) $A\left(\theta\right)$. Show your work

(a) Normal distribution

(b) Binomial distribution

(c) Poisson distribution

(d) Gamma distribution with distribution function

$$f\left(y\right) = \frac{1}{\Gamma\left(\nu\right)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}e^{-y\nu/\mu}, \;\; y > 0 \tag{2}$$

(e) Inverse Gaussian distribution with density function

$$f\left(y\right) = \sqrt{\frac{\lambda}{2\pi y^3}}\exp\left\{-\frac{\lambda\left(y-\mu\right)^2}{2\mu^2 y}\right\}, \;\; y, \mu, \lambda > 0 \tag{3}$$

B. Data are generated for the exponential distribution with density $f(y) = \lambda \exp(\lambda y)$ where $\lambda, y > 0$. The distribution is a member of the exponential family was shown before.

   (a) Identify the specific form of θ, $h(y)$ and $A(\theta)$ for the exponential distribution.

   (b) Whats the canonical link for a generalized linear model (GLM) with a response following the exponential distribution?

   (c) Identify a practical difficulty that may arise when using the canonical link in this instance

C. The Conway-Maxwell Poisson distribution has the probability function

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad y = 0, 1, 2, \ldots \tag{4}$$

where

$$Z(\lambda, \nu) = \sum_{i=1}^{\infty} \frac{\lambda^i}{(i!)^\nu} \tag{5}$$

   (a) Place this distribution in an exponential family form with respect to both parameters, and identify all the relevant components

   (b) In statistics, overdispersion is the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model. Explain why this distribution can be used to model overdispersion for count data.

2 Generalized linear model - Probit regression

In this homework, we consider the Bank data set problem discussed in detail in Chapter 4 (Generalized Linear Models) of the Bayesian Core book by Jean-Michel Marin, and Christian P. Robert. The data set can be downloaded from this link. In this data set, we have measurements on 100 genuine Swiss banknotes and 100 counterfeit ones. The response variable y is thus the status of the banknote, where 0 stands for genuine and 1 stands for counterfeit, while the explanatory factors are the length of the bill x_1, the width of the left edge x_2, the width of the right edge x3, and the bottom margin width x_4, all expressed in millimeters. We want a probabilistic model that predicts the type of banknote (i.e., that detects counterfeit banknotes) based on the four measurements above. In this context, do the following:

A. In GLMs, we have

$$\boldsymbol{y}|\mathbf{x}, \boldsymbol{w} \sim f\left(y|\mathbf{x}^T \boldsymbol{w}\right) \tag{6}$$

The above model is defined by two functions – a conditional density f of y given x that belongs to the exponential family and and that is parameterized by an expectation parameter $\mu = \mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$, and a a link function g that relates the mean $\mu = \mu(\mathbf{x})$

of f and the covariate vector, $\mathbf{x}$ as $g(\mu) = \mathbf{x}^T \boldsymbol{w}$ One of the popular link function is the probit link function $g(\mu_i) = \Phi^{-1}(\mu_i)$ where Φ is the standard normal cdf. The corresponding likelihood is given as

$$l(\boldsymbol{w}|\boldsymbol{y}, \mathbf{x}) \propto \prod_{i=1}^{n} \Phi\left(\mathbf{x}^{iT}\boldsymbol{w}\right)^{y_i} \left[1 - \Phi\left(\mathbf{x}^{iT}\boldsymbol{w}\right)\right]^{1-y_i} \tag{7}$$

Considering non-informative G-priors for the weights

$$\boldsymbol{w}|\sigma^2, \mathbf{x} \sim \mathcal{N}_k\left(0_k, \sigma^2\left(\mathbf{x}^T\mathbf{x}\right)^{-1}\right), \tag{8}$$

and

$$\pi\left(\sigma^2|\boldsymbol{x}\right) \propto \sigma^{-3/2}, \tag{9}$$

compute an expression for the posterior.

B. In part A of this problem, you will see that it is not possible to provide analytical expression for the posterior. Therefore, one has to compute the posterior approximately. In this context, one option is to use the Metropolis Hasting (MH) algorithm as described in Algorithm 4.2 of [1]. Use MH algorithm to compute the posterior distribution of the weights (Similar to Fig. 4.5 in the book mentioned above). Run MH for 10,000 iteration. Consider the first 1000 iterations to be the burn-in period. Plot the histogram of the weights.

C. As an extension of the part B, write a code for drawing samples from the predictive distribution of y^*, $p(y^*|x^*)$. Plot histogram of the predictive distribution of y^* for a randomly generated sample of $\boldsymbol{x}^*$ (within the bound of the data set).

D. Computing marginal distribution of $\boldsymbol{y}$ is important to provide approximations to the Bayes factor. Unfortunately, this can not be computed in a closed form.

Provide an importance sampling based approximation to for the marginal distribution of $\boldsymbol{y}$, $p(\boldsymbol{y})$. Write a code for drawing samples from the marginal distribution of $\boldsymbol{y}$. Use it to compute the Bayes factor corresponding to the null hypothesis, $H_0 : w_i = 0$.

3 Generalized linear model - Logit regression

Consider the same data set as previous problem and perform the following tasks

A. An alternative to the probit regression model is the logit regression model. In this case, the link function $g(\mu_i) = \log(\mu_i/(1-\mu_i))$. For this case, compute the posterior. For this model, the likelihood can be written as

$$l(\boldsymbol{w}|\boldsymbol{y}, \mathbf{x}) = \frac{\exp\left\{\sum_{i=1}^{n} y_i \mathbf{x}^{iT}\boldsymbol{w}\right\}}{\prod_{i=1}^{n}\left[1 + \exp\left(\mathbf{x}^{iT}\boldsymbol{w}\right)\right]} \tag{10}$$

The prior is similar to that defined in the previous problem. Compute an expression for the posterior of $\boldsymbol{w}$.

B. n part A of this problem, you will see that it is not possible to provide analytical expression for the posterior. Therefore, one has to compute the posterior approximately. In this context, one option is to use the Metropolis Hasting (MH) algorithm as described in Algorithm 4.2 of [1]. Use MH algorithm to compute the posterior distribution of the weights (Similar to Fig. 4.5 in the book mentioned above). Run MH for 10,000 iteration. Consider the first 1000 iterations to be the burn-in period. Plot the histogram of the weights.

C. As an extension of the part B, write a code for drawing samples from the predictive distribution of y^*, $p\left(y^*|x^*\right)$. Plot histogram of the predictive distribution of y^* for a randomly generated sample of $\boldsymbol{x}^*$ (within the bound of the data set).

4 K-means algorithm

For the yeast gene expression data provided at this link, use K-means algorithm to compute the cluster centers. Assume, you have 16 clusters

5 Gaussian mixture, expectation maximization and mixture of experts

A. Show the student's t distribution can be represented as infinite mixture of Gaussian. For simplicity, assume one dimensional-distribution.

B. It is possible to interpret the probit regression model as a latent variable model. In this setup, we associate each otem x_i with two utilities u_{0i} and u_{1i}, corresponding to the possible choices of $y_i = 0$ and $y_i = 1$. We then assume that the observed choice is whichever action has larger utility. Further details on this representation can be found in Section 9.4.2 of [2].

   In this context, show how the probit regression, represented as latent variable model, can be trained using expectation maximization.

C. For the data set given at this link, train a probit regression model by using expectation maximization. Compare with conventional probit regression model. Provide your observations.

D. Consider the linear regression model:

$$y = w_0 + w_1 x \tag{11}$$

   For the given data set in this link, compute the unknown coefficients using the student's T model. This is the same problem as 1A of HW3. However, in this case, assume all the parameters to be unknown.

E. Often a single regression model may not be sufficient for tracking variability of a function in the overall problem domain. Under such circumstances, a good option is to use multiple regression model, each applied to a separate input space. We can model this by allowing the mixing weights and the mixture densities to be input-dependent. Such models are known as mixture of experts (MOE). Mixtures of experts are useful in solving inverse problems. These are problems where we have to invert a many-to-one mapping.

   For the data-set provided this link, fit a mixture of experts with three experts. Plot the predictive mean and mode.

   Consider each of the three experts to be a linear regression model of the form

$$y = w_0 + wx \tag{12}$$

References

[1] J.-M. Marin and C. P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics. Springer New York, New York, NY, 2007.

[2] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.