Machine Learning: Homework Assignment 6

E4525 Spring 2018,

IEOR, Columbia University

Due:Apr 19th, 2018

1. **Exponential Distribution**

   $s \in \mathbb{R}^+$ (s continuous and $s > 0$) has exponential distribution

   $$p(s; \lambda) = \lambda e^{-\lambda s} \qquad \text{for} \quad \lambda > 0 \tag{1}$$

   which is an exponential family distribution.

   To turn $p(s; \lambda)$ into the exponential family form we re-write it as

   $$h(s) e^{\eta T(s) - A(\eta)} = \lambda e^{-\lambda s} = e^{-\lambda s + \log \lambda} \tag{2}$$

   from that expression we can read the answer to all the questions below

   Give expressions for

   (a) Base measure $h(s)$.
   $h(s) = 1$

   (b) Sufficient statistic $T(s)$.
   $T(s) = s$

   (c) canonical parameter η as a function of the natural parameter λ.
   $\eta = -\lambda$

   (d) The cumulant function $A(\eta)$.

   $$A(\eta) = -\log(\lambda) = -\log(-\eta) \tag{3}$$

   (e) The average and variance of s as a function of λ

   $$\mathbb{E}(s) = A'(\eta) = \frac{-1}{\eta} = \frac{1}{\lambda} \tag{4}$$

   $$\mathrm{Var}(s) = A''(\eta) = \frac{1}{\eta^2} = \frac{1}{\lambda^2} \tag{5}$$

2. **Gamma Distribution, shape parameter α known**

$s \in \mathbb{R}^+$ (s continuous and $s > 0$), s has Gamma distribution

$$p(s;\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} \qquad \text{for} \ \ \alpha > 0,\ \beta > 0 \tag{6}$$

if we assume α is a known fixed number, $p(s;\beta)$ is has an exponential family distribution with a scalar canonical parameter η.

To turn $p(s;\alpha,\beta)$ into the exponential family form we re-write it as

$$h(s)e^{\eta T(s) - A(\eta)} = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} = \frac{s^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta s + \alpha \log \beta} \tag{7}$$

from that expression we can read the answer to all the questions below

Give expressions for

(a) Base measure $h(s)$.

$$h(s) = \frac{s^{\alpha-1}}{\Gamma(\alpha)} \tag{8}$$

(b) Sufficient statistic $T(s)$.

$$T(s) = s \tag{9}$$

(c) canonical parameter η as a function of the natural parameter β.

$$\eta = -\beta \tag{10}$$

(d) What does the constrain $\beta > 0$ in the definition of the Gamma distribution imply for the canonical parameter?

$$\eta < 0 \tag{11}$$

(e) The cumulant function $A(\eta)$.

$$A(\eta) = -\alpha \log(-\eta) \tag{12}$$

(f) The average and variance of s as a function of α and β

$$\mathbb{E}(s) = A'(\eta) = \frac{-\alpha}{\eta} = \frac{\alpha}{\beta} \tag{13}$$

$$\text{Var}(s) = A''(\eta) = \frac{\alpha}{\eta^2} = \frac{\alpha}{\beta^2} \tag{14}$$

(g) show that the exponential distribution of exercise 1 is an special case of the Gamma distribution.

setting $\alpha = 1$ the Gamma distribution reduces to the exponential one.

3. **Gamma Distribution, shape parameter α unknown**

$s \in \mathbb{R}^+$ (s continuous and $s > 0$) has distribution

$$p(s; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} \quad \text{for} \quad \alpha > 0, \ \beta > 0. \tag{15}$$

If we assume α is unknown number, $p(s; \alpha, \beta)$ is an exponential family distribution with a two dimensional canonical parameter $\eta = (\eta_1, \eta_2)$.

To turn $p(s; \alpha, \beta)$ into the exponential family form with two parameters we re-write it as

$$h(s) e^{\eta^T T(s) - A(\eta)} = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} = e^{-\beta s + (\alpha-1)\log s + \alpha \log \beta - \log \Gamma(\alpha)} \tag{16}$$

from that expression we can read the answer to all the questions below

Give expressions for

(a) Base measure $h(s)$.

$h(s) = 1$, or you could also write $h(s) = \frac{1}{s}$.

(b) Sufficient statistic $T(s) = (T_1(s), T_2(s))$.

$$T_1(s) = s \tag{17}$$

$$T_2(s) = \log s \tag{18}$$

(c) canonical parameter $\eta = (\eta_1, \eta_2)$ as a function of the natural parameters (α, β).

$$\eta_1 = -\beta \tag{19}$$

$$\eta_2 = \alpha - 1 \tag{20}$$

If you have chosen $h(s) = \frac{1}{s}$ then $\eta_2 = \alpha$.

(d) The cumulant function $A(\eta)$.

$$A(\eta) = -\alpha \log \beta + \log \Gamma(\alpha) = -(\eta_2 + 1)\log(-\eta_1) + \log \Gamma(\eta_2 + 1) \tag{21}$$

If we chose $h(s) = \frac{1}{s}$ we have the slightly simpler

$$A(\eta) = -\eta_2 \log(-\eta_1) + \log \Gamma(\eta_2) \tag{22}$$

4. **Generalized Linear Model for Gamma Distribution: canonical link**

We have two random variables $X \in \mathbb{R}^D$ and $Y \in \mathbb{R}^+$. We assume that $Y|_X$ is a Gamma distribution with the same assumptions of problem 2: Y has a Gamma distribution with α fixed

$$p(y|x; \eta, \alpha) = h(y)e^{\eta y - A(\eta)} \tag{23}$$

where $h(y)$ and $A(\eta)$ are given as the solutions of problem 2.

We further assume the canonical link (note the minus sign)

$$\eta(x) = -(b + x^T w). \tag{24}$$

where $w = (w_1, \ldots, w_D)$.

Given N observations $\{x_i, y_i\}$ give expression for the following

(a) The maximum likelihood loss function $E_{\text{Gam}}(b, w; \{x_i, y_i\})$

$$E_{\text{Gam}}(b, w; \{x_i, y_i\}) = \frac{1}{N} \sum_i A(\eta(x_i)) - y_i \eta(x_i) \tag{25}$$

$$= \frac{1}{N} \sum_i \left\{ -\alpha \log(b + x_i^T w) + y_i(b + x_i^T w) \right\} \tag{26}$$

$$\tag{27}$$

(b) The expected value $\hat{y}(x; b, w)$ of y conditional on x as a function of α, b and w

Using the result from exercise 2f, we have

$$\hat{y}(x; w, b) = \mathbb{E}(y|\eta(x)) = -\frac{\alpha}{\eta(x)} = \frac{\alpha}{b + x^T w} \tag{28}$$

(c) The gradient of the loss E_{Gam} for changes of b and w_d

$$\frac{\partial}{\partial b} E_{\text{Gam}} = \frac{1}{N} \sum_i \{\hat{y}(x; b, w) - y_i\} = \frac{1}{N} \sum_i \left\{ \frac{\alpha}{b + x^T w} - y_i \right\} \tag{29}$$

$$\frac{\partial}{\partial w_d} E_{\text{Gam}} = \frac{1}{N} \sum_i x_{i,d} \{\hat{y}(x; b, w) - y_i\} = \frac{1}{N} \sum_i x_{i,d} \left\{ \frac{\alpha}{b + x^T w} - y_i \right\} \tag{30}$$

(d) What does the constrain $\beta > 0$ in the definition of the Gamma probability distribution imply for the parameters b and w?

$\beta > 0$ implies $\eta(x_i) < 0$ for each observation using its definition

$$b + x_i^T w > 0 \quad \text{for} \quad i = 1, 2, \ldots, N \tag{31}$$

5. **Generalized Linear Model for Gamma Distribution: log link**

To avoid the constrained optimization that you analyzed in problem 4d We have two random variables X and Y. We assume that $Y|_X$ is a Gamma distribution with the assumptions of problem 2 For $s \in \mathbb{R}^+$ (s continuous and $s > 0$) has Gamma distribution

$$p(y|x; \eta, \alpha) = h(s)e^{\eta s - A(s)} \tag{32}$$

where η, $h(s)$ and $A(\eta)$ are given as the solutions of problem 2.

We further a non-canonical link defined by $\psi(\eta) = -e^{-\eta}$

$$\eta(x) = \psi(b + x^T b) = -e^{-b - x^T w}. \tag{33}$$

Given N observations $\{x_i, y_i\}$ give expression for the following

(a) The maximum likelihood loss function $E^{\log}_{\text{Gam}}(b, w; \{x_i, y_i\})$

$$E^{\log}_{\text{Gam}}(b, w; \{x_i, y_i\}) = \frac{1}{N}\sum_i A(\eta(x_i)) - y_i\eta(x_i) \tag{34}$$

$$= \frac{1}{N}\sum_i \left\{-\alpha \log(e^{-b-x_i^T w}) + y_i e^{-b-x_i^T w}\right\} \tag{35}$$

$$= \frac{1}{N}\sum_i \left\{\alpha(b + x_i^T w) + y_i e^{-b-x_i^T w}\right\} \tag{36}$$

(b) The expected value $\hat{y}(x; b, w)$ of y conditional on x as a function of α, b and w

Using the result from exercise 2f, we have

$$\hat{y}(x; w, b) = \mathbb{E}(y|\eta(x)) = -\frac{\alpha}{\eta(x)} = \alpha e^{b + x^T w} \tag{37}$$

As

$$\log \hat{y}(x; w, b) = \alpha + b + x^T w \tag{38}$$

This is known as the **log link**

(c) The gradient of the loss $E^{\log}_{\text{Gam}}$ for changes of b and w_d

Given that $\psi(\eta) = -e^{-\eta}$ we have that

$$\psi'(\eta) = e^{-\eta} \tag{39}$$

using that expression

$$\frac{\partial}{\partial b} E^{\log}_{\text{Gam}} = \frac{1}{N} \sum_i \{\hat{y}(x; b, w) - y_i\} \, \psi'(\eta) = \frac{1}{N} \sum_i \left\{\alpha - y_i e^{-b - x^T w}\right\} \tag{40}$$

$$\frac{\partial}{\partial w_d} E^{\log}_{\text{Gam}} = \frac{1}{N} \sum_i x_{i,d} \{\hat{y}(x; b, w) - y_i\} \, \psi'(\eta) = \frac{1}{N} \sum_i x_{i,d} \left\{\alpha - y_i e^{-b - x^T w}\right\} \tag{41}$$