DS-GA 3001: Applied Statistics Practice Midterm
Solutions

Instructions:

- You have **100 minutes**, 4:55PM - 6:35PM
- The exam has 4 problems, totaling 100 points.
- Please answer each problem in the space below it.
- You are allowed to carry the textbook, your own notes and other course related material with you. Electronic devices are not allowed.
- Please read the problems carefully.
- We use boldcase letters $\boldsymbol{\theta}, \boldsymbol{x}, \cdots$ to distinguish vectors from scalars.
- Unless otherwise specified, you are required to provide explanations of how you arrived at your answers.
- You can use previous parts of a problem even if you did not solve them.
- The problems may not be arranged in an increasing order of difficulty. If you get stuck, it might be wise to try other problems first.
- Good luck and enjoy!

Full name: _____

N number: _____

1. Binary choice questions. *(40 points)*

For each of the statements, decide if it is "True" or "False". Provide explanations if you think it is "False". Each question is worth 5 points.

(a) For $\theta \in \mathbb{R}$, let $y \sim p_\theta$ denote the distribution where y is uniformly distributed on the interval $[\theta, \theta + 1]$. This family $(p_\theta)_{\theta \in \mathbb{R}}$ is an exponential family.

Solution: False. The support of p_θ changes with θ, so this family cannot be an exponential family.

(b) Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a sample of n i.i.d. observations from p_θ, and $D_n(\theta_1; \theta_2)$ be the deviance between two parameters $\theta_1, \theta_2 \in \mathbb{R}$ based on $\boldsymbol{y}$. Then $D_n(\theta_1; \theta_2) = nD_1(\theta_1; \theta_2)$, where $D_1(\theta_1, \theta_2)$ is the deviance based on a single observation y_1.

Solution: True. This is because

$$\begin{aligned} D_n(\theta_1; \theta_2) &= 2\mathbb{E}_{\theta_1}\left[\log \frac{p_{\theta_1}(\boldsymbol{y})}{p_{\theta_2}(\boldsymbol{y})}\right] \\ &= 2\mathbb{E}_{\theta_1}\left[\log \prod_{i=1}^{n} \frac{p_{\theta_1}(y_i)}{p_{\theta_2}(y_i)}\right] \\ &= 2\sum_{i=1}^{n} \mathbb{E}_{\theta_1}\left[\log \frac{p_{\theta_1}(y_i)}{p_{\theta_2}(y_i)}\right] \\ &= nD_1(\theta_1; \theta_2) \end{aligned}$$

(c) Let P be an unknown continuous distribution over $\mathbb{R}$. Given i.i.d. $Y_1, \cdots, Y_n \sim P$, Alice computes the following statistic

$$T = T(Y_1, \cdots, Y_n) = \text{number of distinct values in } (Y_1, \cdots, Y_n).$$

For example, $T(1, 2, 3) = 3$, and $T(1.4, 1, 1.4) = 2$. Alice would like to estimate the variance of T via bootstrap: she draws m bootstrap samples $\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(m)}$, where each sample $\mathbf{Y}^{(j)}$ is a collection of n uniformly random draws (with replacement) from $\{Y_1, \cdots, Y_n\}$. Alice proceeds to compute $T^{(j)} = T(\mathbf{Y}^{(j)})$, and uses the sample variance of $(T^{(j)} : j = 1, \cdots, m)$ to estimate the true variance of T.

Claim: for large (m, n), this bootstrap estimate is close to the true variance of T.

Solution: False. Since P is continuous, with probability one we will have $T = n$, and the true variance of T is zero. In contrast, $(Y_1, \cdots, Y_n)$ only takes n discrete values, so the random variable $T(\mathbf{Y}^{(j)})$ can take all possible values in $\{0, 1, \cdots, n\}$, and the bootstrap variance estimate is non-zero.
(Optional: using Poisson approximation one can show that as $n \to \infty$,

$$T^{(j)} = \sum_{i=1}^{n} \mathbb{1}(Y_i \text{ appears in } \mathbf{Y}^{(j)}) \approx \text{Poi}\left(n\left(1 - \left(1 - \frac{1}{n}\right)^n\right)\right) \approx \text{Poi}\left((1 - e^{-1})n\right),$$

so that the bootstrap variance estimate is $\approx (1 - e^{-1})n$. The main intuition of this problem is that the plug-in approach using a discrete distribution is problematic in this example with the true distribution being continuous).

(d) In a GLM with parameter $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$, Alice would like to test if $\beta_1 = \beta_2 = \cdots = \beta_p$. She runs the following procedure:

- compute the unrestricted MLE $\widehat{\boldsymbol{\beta}}^{(1)}$ and the corresponding log-likelihood ℓ_1;
- compute the restricted MLE $\widehat{\boldsymbol{\beta}}^{(2)}$ subject to the constraint $\widehat{\beta}_1^{(2)} = \widehat{\beta}_2^{(2)} = \cdots = \widehat{\beta}_p^{(2)}$, and compute the corresponding log-likelihood ℓ_2.

She then claims that under the null hypothesis $\beta_1 = \beta_2 = \cdots = \beta_p$, asymptotically one should have $2(\ell_1 - \ell_2) \sim \chi^2_{p-1}$. Is this claim correct?

Solution: True. The generalized likelihood ratio test tells that $2(\ell_1 - \ell_2) \sim \chi^2_{p_1 - p_2}$, where p_i is the dimension of the feasible set for $\boldsymbol{\beta}$ when computing ℓ_i. For the unrestricted MLE, we have $p_1 = p$; for the restricted MLE, the set $\{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_1 = \beta_2 = \cdots = \beta_p\}$ is a line so has dimension $p_2 = 1$. Hence $p_1 - p_2 = p - 1$.

(e) For model selection, intuitively speaking AIC aims to balance between two terms:

- the negative log-likelihood, which shrinks with an increasing model complexity;
- the number of model parameters, which grows with an increasing model complexity.

Solution: True. When the model gets larger, the negative log-likelihood becomes smaller (because the feasible set gets larger), and the number of model parameters also increases.

(f) Suppose $\mathbf{D}_1$ and $\mathbf{D}_2$ are two survival datasets for males and females, respectively. Then the following ways to plot the survival curves are equivalent:

- plot the Kaplan-Meier curves for males and females, respectively;
- fit the Cox model on $\mathbf{D}_1 \cup \mathbf{D}_2$ with the feature "gender", then plot the fitted survival curves for males and females, respectively.

Solution: False. The survival curves fitted by the Cox model must satisfy $h_1(t) = \alpha h_2(t)$ for every t, which may not be the case for the separate Kaplan-Meier curves. In HW4 P3 we have also observed that these curves are different.

(g) Recall that in the Cox model, the complete likelihood is $L(\boldsymbol{\beta}, h)$ and the profile likelihood is $pL(\boldsymbol{\beta}) = \max_h L(\boldsymbol{\beta}, h)$. Bob claims that computing the profile maximum likelihood $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} pL(\boldsymbol{\beta})$ is equivalent to one single iteration of the EM algorithm, where one first computes $\widehat{h}$ and then computes $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \widehat{h})$. Is this claim correct?

Solution: False. The profile likelihood looks for the optimal $h^\star$ for any given $\boldsymbol{\beta}$ and then computes $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, h^\star(\boldsymbol{\beta}))$. This is different from using a fixed $\widehat{h}$ and computing $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \widehat{h})$.

(h) For exponential families with missing data, the incomplete log-likelihood may *no longer* be concave in the natural parameter $\boldsymbol{\theta}$.

Solution: True. This is the reason why we apply the EM algorithm, rather than the gradient descent/Newton's method on the incomplete log-likelihood, when we have missing data.

2. EM algorithm with both covariates missing. *(20 points)*

Let the sample $(x_1, y_1), \cdots, (x_{n+m}, y_{n+m})$ be i.i.d. drawn from an exponential family $p_\theta(x, y) = \exp(\langle \theta, T(x, y)\rangle - A(\theta))h(x, y)$. However, the y-covariate is missing in the first n observations, and the x-covariate is missing in the last m observations. In other words, our observed sample is $(x_1, \cdots, x_n, y_{n+1}, \cdots, y_{n+m})$.

(a) Write out the incomplete log-likelihood for the observations (up to additive constants), in terms of $A(\theta)$ and its conditional variants. *(10 points)*

Solution: The incomplete log likelihood for the observations is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p_\theta(x_i) + \sum_{j=n+1}^{n+m} \log p_\theta(y_j) \\ &= \sum_{i=1}^{n} (A_{x_i}(\theta) - A(\theta)) + \sum_{j=n+1}^{n+m} (A_{y_j}(\theta) - A(\theta)) + C,\end{aligned}$$

where

$$\begin{aligned}A_x(\theta) &= \log \left[\int \exp(\langle \theta, T(x, y)\rangle) h(x, y) dy \right], \\ A_y(\theta) &= \log \left[\int \exp(\langle \theta, T(x, y)\rangle) h(x, y) dx \right].\end{aligned}$$

(b) Describe the EM algorithm for the MLE computation. You should give the details of both E and M steps; you need not give proofs. *(10 points)*

Solution: Similar to the EM algorithm in class, to move from $\theta^{(t)}$ to $\theta^{(t+1)}$:

- E-step: compute the vector $(\mu_1^{(t+1)}, \cdots, \mu_{n+m}^{(t+1)})$ with

$$\mu_i^{(t+1)} = \begin{cases} \mathbb{E}_{Y \sim p_{\theta^{(t)}}(\cdot|x_i)}[T(x_i, Y)] & \text{if } 1 \le i \le n, \\ \mathbb{E}_{X \sim p_{\theta^{(t)}}(\cdot|y_i)}[T(X, y_i)] & \text{if } n+1 \le i \le n+m. \end{cases}$$

- M-step: compute $\theta^{(t+1)}$ from the estimating equation

$$\nabla A(\theta^{(t+1)}) = \frac{1}{n+m} \sum_{i=1}^{n+m} \mu_i^{(t+1)}.$$

3. Mixture model with known locations. *(20 points)*

In a mixture model, we have a dataset $(y_1, \cdots, y_n)$ with $y_i \sim p_{\theta_i}$, where the unknown parameters $\theta_1, \cdots, \theta_n$ are i.i.d. drawn from an unknown distribution π. We can think of θ_i as the "locations" of the mixture, and the vector π as the "weights".

Throughout this problem we assume that both θ_i and y_i take discrete values, i.e.

$$\theta_i \in \Theta = \{\theta^1, \cdots, \theta^M\},$$
$$y_i \in \mathcal{Y} = \{y^1, \cdots, y^N\}.$$

Therefore, we may represent π as a probability vector $(\pi_1, \cdots, \pi_M)$, in the sense that $\mathbb{P}(\theta = \theta^j) = \pi_j$ if $\theta \sim \pi$. We will also use the notation $K_{j\ell} = p_{\theta^j}(y^\ell) = \mathbb{P}(y = y^\ell \mid \theta = \theta^j)$ to denote the conditional probability of observing $y = y^\ell$ when $\theta = \theta^j$.

(a) If $\theta \sim \pi$ and $y \sim p_\theta$, write down the marginal distribution of y in terms of (π, K). *(5 points)*

Solution: The marginal distribution of y is

$$p_\pi(y = y^\ell) = \sum_{j=1}^{M} \pi_j p_{\theta^j}(y^\ell) = \sum_{j=1}^{M} \pi_j K_{j\ell}.$$

Expressing in terms of indicators gives

$$p_\pi(y) = \sum_{\ell=1}^{N} \mathbb{1}(y = y^\ell) \cdot \sum_{j=1}^{M} \pi_j K_{j\ell}.$$

(b) Write down the log-likelihood of the dataset $(y_1, \cdots, y_n)$, as a function of π; we assume that $\Theta, \mathcal{Y}, K$ are known. Is the log-likelihood concave in π? *(5 points)*

Solution: The log-likelihood is

$$\log p_\pi(y_1, \cdots, y_n) = \sum_{i=1}^{n} \log p_\pi(y_i) = \sum_{i=1}^{n} \log \left(\sum_{\ell=1}^{N} \mathbb{1}(y_i = y^\ell) \cdot \sum_{j=1}^{M} \pi_j K_{j\ell} \right).$$

Since $x \mapsto \log(x)$ is concave, and the term inside the logarithm is affine in π, the log-likelihood is concave in π.

(c) Now suppose that π_j takes a form of a one-dimensional exponential family, i.e.

$$\pi_j = \exp(\beta T_j - A(\beta))h_j$$

for some known (T_j, h_j) and $A(\cdot)$. Is your log-likelihood in (b) concave in β? *(5 points)*

Solution: No. The log-likelihood becomes

$$\log p_\pi(y_1, \cdots, y_n) = \sum_{i=1}^{n} \log \left(\sum_{\ell=1}^{N} \mathbb{1}(y_i = y^\ell) \cdot \sum_{j=1}^{M} \exp(\beta T_j) h_j K_{j\ell} \right) - nA(\beta),$$

where the first term is convex in β, and the second term (i.e. $-A(\beta)$) is concave in β. So the overall log-likelihood may not be concave in β in general.

(d) Now suppose that Θ is unknown, so the matrix $K = K(\Theta)$ becomes a function of Θ. Is your log-likelihood in (b) jointly concave in π and Θ? *(5 points)*

Solution: No. Even in the simple Gaussian mixture model, the log-likelihood is not jointly concave in the weight parameters and location parameters. What we show in (b) is that if we fix the location parameters, then the log-likelihood becomes concave in the weights π.

4. Survival analysis. *(20 points)*

(a) Consider a survival dataset $\{(t_i, d_i, n_i)\}_{i=1}^N$, where n_i is the number of individuals who have survived through time t_i, and d_i is the number of deaths at time t_i. For simplicity we assume that there is no censoring.

Recall that the Kaplan-Meier estimator for the hazard rate at each time t_i is

$$\widehat{h}(t_i) = \frac{d_i}{n_i}.$$

We assume that $d_i \sim \mathrm{B}(n_i, h(t_i))$, where $\mathrm{B}(n, p)$ denotes the binomial distribution with n trials and success probability p, and $h(t_i)$ is the true hazard at t_i. Compute $\mathrm{Var}(\widehat{h}(t_i))$. *(5 points)*

Solution: As $\mathrm{Var}(X) = np(1-p)$ for $X \sim \mathrm{B}(n, p)$, we have

$$\mathrm{Var}(\widehat{h}(t_i)) = \frac{\mathrm{Var}(d_i)}{n_i^2} = \frac{h(t_i)(1 - h(t_i))}{n_i}.$$

(b) Recall that Kaplan-Meier estimator for the survival function is

$$\widehat{S}(t_i) = \prod_{j:t_j \le t_i} \left(1 - \widehat{h}(t_j)\right).$$

Suppose we know that

$$\mathrm{Var}(\log \widehat{S}(t_i)) \approx \sum_{j:t_j \le t_i} \left(\frac{1}{1-\widehat{h}(t_j)}\right)^2 \mathrm{Var}(\widehat{h}(t_j)).$$

Find the approximate variance $\mathrm{Var}(\widehat{S}(t_i))$ using the delta method and the plug-in approach.

You should use your result in (a), and express your final answer using $\widehat{S}(t_i)$ and $\{(t_i, d_i, n_i)\}_{i=1}^N$. *(5 points)*

Solution: Apply the delta method on $\log \widehat{S}(t_i)$ we get

$$\mathrm{Var}(\log \widehat{S}(t_i)) \approx \widehat{S}(t_i)^{-2} \mathrm{Var}(\widehat{S}(t_i)).$$

Hence

$$\begin{aligned}
\mathrm{Var}(\widehat{S}(t_i)) &\approx \widehat{S}(t_i)^2 \mathrm{Var}(\log \widehat{S}(t_i)) \\
&\approx \widehat{S}(t_i)^2 \sum_{j:t_j \le t_i} \left(\frac{1}{1-\widehat{h}(t_j)}\right)^2 \mathrm{Var}(\widehat{h}(t_j)) \\
&= \widehat{S}(t_i)^2 \sum_{j:t_j \le t_i} \left(\frac{1}{1-\widehat{h}(t_j)}\right)^2 \frac{\widehat{h}(t_j)(1-\widehat{h}(t_j))}{n_j} \\
&= \widehat{S}(t_i)^2 \sum_{j:t_j \le t_i} \frac{\widehat{h}(t_j)}{n_j(1-\widehat{h}(t_j))} \\
&= \widehat{S}(t_i)^2 \sum_{j:t_j \le t_i} \frac{d_j}{n_j(n_j-d_j)}
\end{aligned}$$

(c) Now consider a dataset $\{(t_i, \Delta_i, \boldsymbol{x}_i)\}_{i=1}^N$ with features $\boldsymbol{x}_i \in \mathbb{R}^d$, true-death indicators $\Delta_i \in \{0, 1\}$, and distinct stopping times t_i. The Cox model assumes that

$$h(t \mid \boldsymbol{x}) = e^{\boldsymbol{\beta}^\top \boldsymbol{x}} h(t),$$

where $\boldsymbol{\beta}$ independent of time.

Now suppose we would like to incorporate the time dependence by $\boldsymbol{\beta}(t) = g(t)\boldsymbol{\beta}$, or equivalently,

$$h(t \mid \boldsymbol{x}) = e^{g(t)\boldsymbol{\beta}^\top \boldsymbol{x}} h(t).$$

Write out the partial likelihood you will use to estimate $\boldsymbol{\beta}$. You may assume that g is a known function. *(5 points)*

Solution: The partial likelihood is

$$L(\boldsymbol{\beta}) \propto \prod_{i:\Delta_i=1} \frac{e^{g(t_i)\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{\sum_{k \in R_i} e^{g(t_i)\boldsymbol{\beta}^\top \boldsymbol{x}_k}},$$

where $R_i = \{j : t_j \geq t_i\}$ is the rist set at time t_i. Note that on the denominator we have $g(t_i)$ instead of $g(t_k)$, and this is because the current ratio is the probability that individual i is the chosen one who dies at time t_i among R_i. You can also derive the same formula using the profile likelihood.

(d) Propose a model for $h(t \mid \boldsymbol{x})$ if the features $\boldsymbol{x}$ are assumed to be time-dependent, i.e. $\boldsymbol{x} = \boldsymbol{x}(t)$. Do you think it is helpful to include both time-dependent features $\boldsymbol{x}(t)$ and the time-dependent coefficient $g(t)$ in (c)? *(5 points)*

Solution: When the features are time-dependent, we have

$$h(t \mid \boldsymbol{x}) = e^{\boldsymbol{\beta}^\top \boldsymbol{x}(t)} h(t).$$

The extra $g(t)$ factor is unnecessary as we can reparametrize $\widetilde{\boldsymbol{x}}(t) = g(t)\boldsymbol{x}(t)$ so that

$$e^{g(t)\boldsymbol{\beta}^\top \boldsymbol{x}(t)} h(t) = e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}(t)} h(t).$$