Final Exam IEDA 5270

Name _____ Student ID _____

Question 1 (15 points)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with unknown μ, σ^2. Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. Assuming $\alpha < 1/2$, show that the one-sided t-test is equivalent to the likelihood ratio test.

Question 2 (15 points)

Let Y_1 be binomial(n, θ), so that $Y_1 = \sum_{i=1}^n X_i$, where X_i is Bernoulli with success probability θ. We know that Y_1 is a complete and sufficient statistic for θ. Since $\mathbb{E}[Y_1] = n\theta$, we know that Y_1/n is a UMVUE of θ. Now, let $Y_2 = (X_1 + X_2)/2$. In an effortless manner, find $\mathbb{E}[Y_2|Y_1]$.

Question 3 (25 points)

Suppose (X, Y) has the joint density

$$f(x, y|\theta) = \exp\{-(\theta x + y/\theta)\}, \ x, y > 0.$$

(a) (10 points) For an i.i.d sample $\{(X_i, Y_i) : i = 1, \cdots, n\}$, show that the Fisher information is $I(\theta) = 2n/\theta^2$.

(b) (5 points) For the estimators

$$T = \sqrt{\left(\sum Y_i\right) / \left(\sum X_i\right)} \text{ and } U = \sqrt{\left(\sum Y_i\right)\left(\sum X_i\right)}$$

show that the information in T alone is $[n/(2n+1)]I(\theta)$;

(c) (5 points) Show that the information in (T, U) is $I(\theta)$;

(d) (5 points) Show that (T, U) is jointly sufficient but not complete.

Question 4 (25 points)

We have seen in class that the Poisson regression is particularly useful in modeling the count of a specific event of interest. However, in real-world applications, the numbers of zeros in the sample may not be properly modeled by a Poisson distribution. That is to say, conditioning on having a positive count, a Poisson fits the data well, but that is not true for the probability of having a zero count. To address this problem, one may modified the Poisson regression as follows: Y_i is exactly zero with probability $p(\boldsymbol{X}_i)$, and Y_i is a Poisson$(\lambda(\boldsymbol{X}_i))$ random variable with probability $1 - p(\boldsymbol{X}_i)$. Recall that the Poisson density is given by

$$f(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{for } \lambda > 0.$$

(a) (5 points) Let $p_i = p(\boldsymbol{X}_i)$ and $\lambda_i = \lambda(\boldsymbol{X}_i)$, calculate $\mathsf{P}(Y_i = 0|\boldsymbol{X}_i)$ and $\mathsf{P}(Y_i = k|\boldsymbol{X}_i)$ for $k > 0$.

(b) (5 points) Under the GLM framework, we model the probability by

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \boldsymbol{X}^T\boldsymbol{\beta}, \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{X}^T\boldsymbol{\gamma}.$$

Write down the likelihood function of this model, given data $\{(\boldsymbol{X}_i, Y_i), i = 1 \ldots, n\}$.

(c) (10 points) To find the MLE for $\boldsymbol{\beta}, \boldsymbol{\gamma}$, calculate the gradient and Hessian of the likelihood function you derived from part (b).

(d) (5 points) Describe a procedure to justify the use of this modified model instead of the naive Poisson regression.

Question 5 (10 points)

Given data $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ where $X_i \in \mathbb{R}^1$, consider the model

$$Y = f(X) + \epsilon,$$

where ϵ is a Gaussian noise with zero-mean and variance σ^2 and

$$f(x) = \sum_{i=-D\cdot 2^D}^{D\cdot 2^D} \theta_i \phi_{\xi_i}(x), \quad \text{with } \phi_{\xi_i}(x) = e^{-(x-\xi_i)^2}.$$

That is, we model Y as the sum of $2 \cdot D \cdot 2^D$ basis functions $\phi_{\xi_i}(\cdot)$, which are placed equally in the interval $[-D, D]$ with gaps 2^{-D}. Apparently the number of parameters is huge and we will definitely overfit the model. We learn the parameters using ridge regression with $\lambda = \sigma^2 \cdot 2^D$, that is, the cost function is given by

$$J(\boldsymbol{\theta}) = \|\boldsymbol{X\theta} - \boldsymbol{Y}\|^2 + \sigma^2 2^D \|\boldsymbol{\theta}\|^2.$$

Let $D \to \infty$, show that the prediction can be computed as

$$\widehat{f}(x) = \boldsymbol{k}(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y},$$

where $\boldsymbol{k} = (k_1, \ldots, k_N)$, $\boldsymbol{K} = \{K_{ij}\}_{0 \le i,j \le N}$ with

$$k_i = k(x, X_i), \quad k_{i,j} = k(X_i, X_j)$$

and

$$k(x, y) = \sqrt{\frac{\pi}{2}} e^{-(x-y)^2/2}.$$

Question 6 (10 points)

The principal component analysis (PCA) is effective for dimensionality reduction. Indeed, PCA can be used in conjunction with linear regression to achieve parameter shrinkage. This is done by first applying PCA to the data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, keep the first $k < p$ principal components $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k\}$, and apply a multiple linear regression of $\boldsymbol{Y}$ using the first k PC $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k) \in \mathbb{R}^{n\times k}$. This is called *principal component regression*. Recall that for PCA, we use the singular-value decomposition $\boldsymbol{X} = \boldsymbol{UDV}^T$, where $\boldsymbol{D}^2$ is a diagonal matrix with elements d_i^2 in descending order, and $\boldsymbol{U} \in \mathbb{R}^{n\times p}$, $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices.

(a) (5 points) For PC regression, write down the predicted value $\widehat{\boldsymbol{Y}}^{\text{PCR}}$ in the form of $\widehat{\boldsymbol{Y}}^{\text{PCR}} = \boldsymbol{SY}$, where $\boldsymbol{S}$ depends explicitly on $\boldsymbol{U}$ but not on $\boldsymbol{X}$.

(b) (3 points) Repeat part (a) for the ridge regression.

(c) (2 points) Compare the shrinkage effect of PC regression and ridge regression.