Final

SDS384

Spring 2019

This exam has five short and three long questions. You will have to answer **four short questions**, **two long questions**. The assigned points are noted next to each question; the total number of points is 50. You have 180 minutes to answer the questions.

Please answer all problems in the space provided on the exam. Use extra pages if needed. Of course, please put your name on extra pages.

Read each question carefully, **show your work** and **clearly present your answers**. Note, the exam is printed two-sided - please don't forget the problems on the even pages!

Good Luck!

Name: _____

UTeid: _____

1 Short questions (28 points)

Please answer any four of the short questions.

1. (7 pts) Consider a sequence of iid random variables $X_1, \ldots, X_n$ such that $X_i \sim Beta(\theta, 1)$, where $\theta > 0$. Let $\bar{X}_n$ denote the sample mean. The method of moments estimator of θ is $\hat{\theta}_n = \bar{X}_n/(1 - \bar{X}_n)$. Derive the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$. *You can use the fact that a* $Beta(\beta, 1)$ *random variable has mean* $\beta/(1 + \beta)$ *and variance* $\frac{\beta}{(\beta+1)^2(\beta+2)}$.

2. (7 pts) What is the VC dimension of the union of k intervals on the real line? Show your calculations.

3. (7 pts) Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of random variable with density f on the real line. A standard estimate of f is the kernel density estimate

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where $K : \Re \to [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(t)dt = 1$, and h is a bandwidth parameter. We will measure the quality of $\hat{f}$ using $\|\hat{f} - f\|_1 := \int_{-\infty}^{\infty} |\hat{f}(t) - f(t)|dt$. Prove that:

$$P(\|\hat{f} - f\|_1 \geq E\|\hat{f} - f\|_1 + \delta) \leq e^{-cn\delta^2},$$

where c is some constant.

4. (7 pts) Consider n i.i.d random variables $X_1, \ldots, X_n$ with mean μ. We are interested in estimating $E(X_1 - \mu)^3$. Construct a U statistic for estimating this. Explicitly write down the kernel.

5. (7 pts) Consider $X_1, \ldots, X_n$, n independent Uniform($[a, b]$) random variables. Obtain an upper bound on $E[\max_i X_i]$ in terms of a, b and n.

2 Long questions (22 points)

Please answer any two of the long questions.

1. (11 pts) Let $X_1, \ldots, X_n$ be independent random variables with $X_n \sim N(0, \sigma_n^2)$, where $\sigma_k^2 = 2^{-(k-1)}$.

   (a) (5 pts) Does the Lindeberg condition hold? Give a proof.

(b) (4 pts) Does $\sum_i X_i$ converge to a normal distribution? Give a proof of your answer. If yes, obtain the parameters of the limiting normal distribution.

(c) (2 pts) Write in a few sentences if this contradicts the Lindeberg-Feller theorem.

2. (11 pts) Let $X_1, \ldots, X_n$ be i.i.d random variables with mean μ and variance σ^2. We are interested in estimating the variance of the quantity $U = \frac{\sum_{i<j} X_i X_j}{\binom{n}{2}}$.

   (a) (6pts) Use the Efron-Stein inequality to obtain an upper bound on the variance of U.

(b) (4pts) What is the asymptotic variance of U?

(c) (1pt) Is the upper bound obtained using the Efron Stein inequality tight? Explain in a few sentences.

3. (11 pts) Let $X_i \in \mathbb{R}^p, i = 1 \dots n$ be i.i.d random variables such that $X_i \sim N(0, I_{p\times p})$ where $I_{p\times p}$ is the $p \times p$ identity matrix. Define the function class $\mathcal{F} = \{f : \mathbb{R}^p \to \mathbb{R} | f(x_1, \dots, x_p) = \beta^T x; \|\beta\|_1 \leq R\}$, where $\beta^T x = \sum_{i=1}^p \beta_i x_i$. We will do a direct proof of $\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i f(X_i) - E[f(X_1)]\right| \xrightarrow{P} 0$.

   (a) (4pts) Show that

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i f(X_i) - E[f(X_1)]\right| \leq \frac{R}{n}\|\sum_i X_i\|_\infty.$$

   Hint: you can use the fact that $\sup_{\|u\|_1\leq 1} |u^T v| = \|v\|_\infty$, *where* $u, v \in \mathbb{R}^p$.

(b) (2pts) Show that

$$\frac{R}{n}\|\sum_i X_i\|_\infty = \frac{R}{\sqrt{n}}\max_{1\le j\le p} Z_j,$$

where Z_j's are i.i.d standard normal random variables. Jack, there is a typo here it should be $|Z_j|$. So I would suggest a bit more lenient grading. If they prove Z_j thats fine.

(c) (5pts) Now show that, as long as $R\sqrt{\log p/n} \to 0$,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_i f(X_i) - E[f(X_1)]\right| \xrightarrow{P} 0.$$

If they show this without the absolute value, they should get full score.