# STATS303 Midterm Exam

Fall’20 Session 2

NAME: ___________________________

NetID: ___________________________

Time: 75min

Total points: 200pt

Please keep this exam confidential and do not share the problems with others.

## Problem 1. (60pt)

Assume you are the principal data scientist of a company named Dorakitty. A manager from another division needs some help from you and asks you the following questions. Answer his / her questions in plain language.

1. (20pt) Our division have some sales data and we trained a linear regression model with 100 independent variables. But we don't believe we need to interpret our sales using so many variables. What should we do?

2. (20pt) Our sales data can be partitioned into several clusters and we plan to use K-means for this task. However, we don't know exactly how many clusters we have. What should we do? Give me **one** idea.

3. (20pt) Our division fit a Nadaraya-Watson kernel weighted average for our sales data. However, at boundary there seems to be a big bias. What should we do?

## Problem 2. (50pt)

Assume a regression model $r = f(x) + \epsilon$ where $x, r \in \mathbb{R}$, $f(x)$ is some deterministic but unknown function and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Suppose $g(x|\theta)$ is our estimator to $f$ where $\theta$ denotes the parameters.

1. (20pt) Write the density $p(r|x)$ in terms of $g(x|\theta)$ and $\sigma$.

2. (10pt) Suppose there is an unknown joint density $p(x, r)$ for $x$ and $r$. Explain why the log likelihood $\mathcal{L}(\theta|\mathcal{X})$ of $p(x, r)$, where the sample $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$ contains i.i.d. data points, can be written as

$$\mathcal{L}(\theta|\mathcal{X}) = \log \prod_{t=1}^{N} p(r^t|x^t) + C \ .$$

3. (20pt) According to Parts 1 and 2, show that the maximum likelihood estimator is given by minimizing

$$\frac{1}{2} \sum_{t=1}^{N} [r^t - g(x^t|\theta)]^2 \ .$$

## Problem 3. (50pt)

Consider the data points $\boldsymbol{x}_1 = (0,1,2)^{\mathrm{T}}$, $\boldsymbol{x}_2 = (-1,3,4)^{\mathrm{T}}$, $\boldsymbol{x}_3 = (0,0,1)^{\mathrm{T}}$ and $\boldsymbol{x}_4 = (2,3,-2)^{\mathrm{T}}$.

1. (10pt) Write a data matrix $\boldsymbol{X}$ for the data points where each row correspond to a data point.
2. (10pt) What is the first step if we want to apply PCA to the data points? Choose from the following.
   (A) Center the data around the origin;
   (B) Perform SVD on $\boldsymbol{X}$;
   (C) Perform K-means on $\boldsymbol{X}$;
   (D) Perform dimensionality reduction on $\boldsymbol{X}$.
3. (10pt) Suppose a system gives output $r_j$ if we input $\boldsymbol{x}_j$ for $j = 1,2,3,4$. We fit a ridge regression model by solving

   $$\min_{\boldsymbol{w}\in\mathbb{R}^4} \frac{1}{2}\left\|\boldsymbol{r}-\tilde{\boldsymbol{X}}\boldsymbol{w}\right\|^2 + \frac{\lambda}{2}\left\|\boldsymbol{w}\right\|^2 \ , \tag{🕵}$$

   where $\boldsymbol{r} = [r_1, r_2, r_3, r_4]^{\mathrm{T}}$. What is $\tilde{\boldsymbol{X}}$?
4. (20pt) By taking the gradient with respect to $\boldsymbol{w}$, derive the solution of (🕵) in terms of $\tilde{X}$,$\lambda$ and $\boldsymbol{r}$.

**Problem 4. (40pt)**

1. (20pt) Let $\{x^t\}_{t=1}^N$ be given. The $K$-NN density estimator is given by $\hat{p}(x) = \dfrac{K}{2Nd_K(x)}$ where $d_K(x)$ is the distance between $x$ and its $K$-th closest neighbor in $\{x^t\}_{t=1}^N$. Prove that $\hat{p}$ is NOT a density.

2. (20pt) Consider applying $K$-means with $K = 2$ clusters to the five points $(0, 0)$, $(1, 2)$, $(2, 0)$, $(3, 2)$, $(4, 0)$. Suppose the initial centers are set to be $(0, 0)$ and $(3, 0)$. Write the E-step and the M-step for the first iteration. You need to clearly state the locations of the centers and the labels of the points.