HW-2 (due by EOD March10, 2021). Please cite your reference should you use any.

Your name (ID):

1. (20 pts) Let $\mathbb{X} = \mathbb{Y} = \{1, 2, \cdots, 10\}, \mathbb{A} = \{1, 2, \cdots, 10, 11\}$, and suppose the data distribution has marginal distribution $X \sim Uniform\{1, 2, \cdots, 10\}$. Furthermore, assume $Y = X$ (i.e., Y always has the exact same value as X). In the questions below we use square loss function $\ell(\hat{f}) = (\hat{f}(x) - y)^2$.

   (a) (5 pts) What is the Bayes risk?

   (b) (5 pts) What is the approximation error when using the hypothesis space of constant functions?

   (c) (10 pts) Suppose we use the hypothesis space F of affine functions $f(x) = a + bx$ for some $a, b \in \Re$.

   (i) What is the approximation error?

   (ii.) Denote f_F the best prediction function in F and consider function $\hat{f}(x) = x + 1$. Compute $R(\hat{f}) - R(f_F)$.

2. (20 pts) Consider a linear model with some Gaussian noise:

$$y_i = \sum_i w_i x_i + b + \epsilon_i, \tag{1}$$

where $\epsilon_i \sim \mathbb{N}(0, \sigma^2), i = 1, \cdots, n$. Where $y_i \in \Re$ is a scalar, $x_i \in \Re^d$ is a d-dimensional vector,$b \in \Re$ is a constant, $w \in \Re^d$ is d-dimensional weight on x_i, and ϵ_i is a i.i.d. Gaussian noise with variance σ^2. Given the data $x_i, i = 1, \cdots, n$, it is our goal to estimate w and b which specify the model.
We will show that solving the linear model (1) with MLE method is the same as solving the following Least Squares problem,

$$\underset{\beta}{argmin}(y - x^T\beta)^T(y - x^T\beta) \tag{2}$$

where $y = (y_1, \cdots, y_n)^T$,$x'_i = (1, x_i)^T$,$X' = (x'_1, \cdots, x'_n)$ and $\beta = (b, w)^T$.

   (a) (5 pts) From the model (1), derive the conditional distribution of $y_i|x_i, w, b$. Again, x_i is a fixed data point.

   (b) (5 pts) Assuming i.i.d. between each $\epsilon_i, i = 1, \cdots, n$ give an explicit expression for the loglikelihood, $\ln P(y|\beta)$ of the data.

(c) (5 pts) Now show that solving for β that maximizes the loglikelihood, i.e. MLE, is the same as solving the Least Square problem of (2).

(d) (5 pts) Derive β that maximizes the loglikelihood. (Assume x' has full rank on column space.)

3. (30 pts) Consider the Ridge regression with

$$\hat{\beta} = \underset{\beta}{argmin} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda ||\beta||_2^2,$$

where $x_i = [x_i^{(1)}, \cdots, x_i^{(p)}]$

(a) (10 pts) Show that a closed form expression for the ridge estimator is $\hat{\beta} = (A^T A + \lambda \mathbb{I})^{-1} A^T y$, where $A = [x_1, \cdots, x_n]$.

(b) (20 pts) An advantage of ridge regression is that a unique solution always exists since $(A^T A + \lambda \mathbb{I})$ is invertible. To be invertible, a matrix needs to be full rank. Argue that $(A^T A + \lambda \mathbb{I})$ is full rank by characterizing its p eigenvalues in terms of the singular values of A and λ.

4. (30 pts) Suppose we have training data of size n from an univariate Gaussian distribution of known variance σ^2 but unknown mean μ. Suppose further this mean itself is random, and characterized by a Gaussian distribution with mean μ_0 and variance σ_0.

(a) (10 pts) What is the MAP estimator for μ?

(b) (20 pts) Show the likelihood function of Gaussian sample is also Gaussian w.r.t. the parameter μ. That is, $p(D|\mu) \propto \mathcal{N}(\bar{x}|\mu, \frac{\sigma^2}{n})$.