# CSE Ph.D. Qualifying Exam, Fall 2023

This is a **closed book** exam. No books or notes are allowed.

# Data Analysis

Please answer three of the following four questions. All questions are graded on a scale of 10. If you answer all four, all answers will be graded and the three lowest scores will be used in computing your total. Show all your work and write in a readable way.

1. **Gaussian Statistics**

   The density of a multivariate normal random variable with mean $\mu$ and covariance matrix $\Sigma$ is given by

   $$P_{\mathcal{N}}(x|\mu,\Sigma) = (\det(2\Sigma))^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \quad (1)$$

   (1) [4pts] Consider the case where the random vector is split as $x = (x_1, x_2)$, with mean $\mu = (\mu_1, \mu_2)$ and (strictly positive definite) covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}$. Using the formula for conditional densities, show that the conditional distribution $x_2|x_1 = z$ is a multivariate normal with mean $\mu_2 + \Sigma_{2,1}(\Sigma_{1,1})^{-1} z$ and covariance matrix $\Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}\Sigma_{1,2}$ in the special case where $\Sigma$ is $2 \times 2$. *Hint: complete the square in the exponential. Use the formula* $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

   (2) [3pts] Assume that data points $(x_i, y_i)_{1 \le i \le N}$ are obtained as independent samples from a joint normal distribution with unknown mean and covariance. Derive the maximum likelihood estimates of its mean and covariance. Use part [1] to derive an estimate of $y$ for a previously unseen $x$.

   (3) [3pts] Consider the conditional model $y_i = \alpha^T x_i + \beta + \epsilon_i$ for independently distributed $\epsilon_i$ with mean zero and identity covariance matrix. Derive the maximum likelihood estimate of $\alpha, \beta$ and show how to use this model to predict $y$ for a previously unseen $x$. Compare the result to (2) and comment.

2. **Dimensionality Reduction**

   (1) [4pts] [**PCA**] There are many ways to "project" data $X \in \mathbb{R}^{n \times d}$ from high dimensions to lower dimensions $\hat{X} \in \mathbb{R}^{n \times p}$. $n$ is the number of data points, $d$ is the dimension of the original data, and $p$ is the dimension of the projected representations. PCA aims to finds the best linear projection i.e. the one that minimizes the reconstruction error $||X - \hat{X}||_F$ (the norm here is the Frobenius norm, described below). It does so by computing the covariance matrix of the data $C = \frac{1}{n} X^\top X$, and then projecting the data onto the first few eigenvectors of $C$. But why does finding the projection of the data onto the largest eigenvectors of the covariance

matrix minimize the reconstruction error? Please prove this mathematically. For simplicity, we will consider the case of PCA from $d$ dimensions to 1 dimension. We will also assume that $X$ is "centered," meaning that the mean across all samples of every data dimension is 0. (*Hint, establish the connection between minimizing the reconstruction error and maximizing projected variance*).
** The Frobenius norm, sometimes also called the Euclidean norm (a term unfortunately also used for the vector $L^2$-norm), is matrix norm of an $m \times n$ matrix $A$ defined as the square root of the sum of the absolute squares of its elements, $||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$. **

(2) [3pts] [**KernelPCA**] Show how to use kernels in PCA, i.e., derive kernelPCA and its projected low-dimensional representation.

(3) [3pts] [**KernelPCA**] We want to apply KernelPCA to the 2D raw data in Figure 1, which is centered. The projected data should also be 2-dimensional. Which kernel function should be used so that the projected data would be linearly separable? The kernel function doesn't need to be precise. You can use {a, b, c, ...} to replace the unknown coefficients. Please also draw a sketch of the 2D representation with KernelPCA applied to the data in Figure 1.

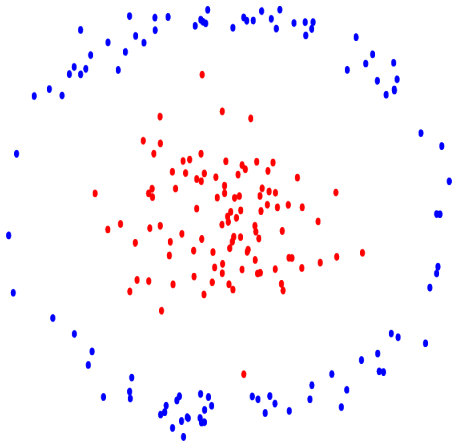

Figure 1: 2D raw data.

## 3. Maximum Likelihood and Maximum A Posteriori Estimations

Consider a biased coin with an unknown probability $\theta \in [0, 1]$ of landing heads after a flip. After conducting multiple coin flips, the resulting sequence is denoted as $\boldsymbol{x} =$ {H, H, T, H, T}, where 'H' represents heads and 'T' represents tails.

(1) [3pts] Determine the maximum likelihood estimation (MLE) for the parameter $\theta$ based on the observed flips.

(2) [2pts] If we know that the probability $\theta$ of the fixed coin must be one of the following values: $\theta \in \{0.2, 0.5, 0.8\}$, then what is the MLE for $\theta$?

(3) [3pts] Consider the same restricted set of possible $\theta$ values as in (2). Additionally, you have access to prior probabilities for each value: $p(\theta = 0.2) = 0.1$, $p(\theta = 0.5) = 0.05$, and $p(\theta = 0.8) = 0.85$. Determine the maximum a posteriori (MAP) estimation for $\theta$.

(4) [2pts] Given an infinite number of flips of the biased coin, discuss the relationship between the results obtained from MLE and MAP estimations. Provide a concise explanation.

4. **Neural Networks**

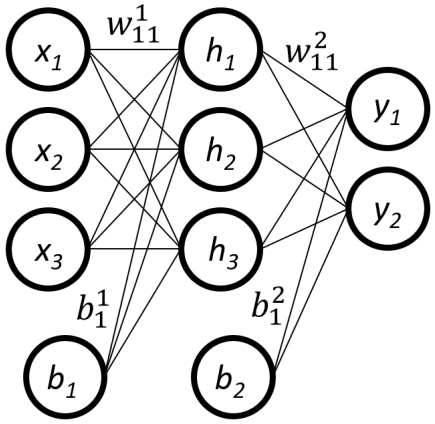


Figure 2: Feed forward neural network.

Figure 1 shows a feed forward neural network with one hidden layer containing three neurons $h_1$, $h_2$, $h_3$ with a sigmoid activation function, with three inputs $x_1$, $x_2$, $x_3$ and two linear output layers $y_1$, $y_2$.

(1) [1pts] How many total parameters are in this model? Suppose we add an additional hidden layer with 4 neurons, how many total parameters do we have now?

(2) [2pts] What is the forward expression to compute $y_1$?

(3) [2pts] Suppose we train the model on the squared loss $L = 1/2(y - y')^2$. What is the expression for $\frac{\partial L}{\partial w_{ij}^2}$?

(4) [3pts] What is the expression for $\frac{\partial L}{\partial w_{ij}^1}$?

(5) [2pts] Name any three strategies to reduce overfitting for this model.