Please remember to justify your answers carefully.

Last name: ___________________ First name: ___________________
Question 1 (Linear regression.)
Consider the following data with one input and one output.

<table>
<thead>
<tr>
<th>$x$ (input)</th>
<th>$y$ (output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

(a) Consider modeling the dependence between the input, $x$, and output, $y$, using a linear model, $y = ax + c$. What is the squared error, $\sum_i (y_i - (ax_i + c))^2$, obtained by running linear regression on the entire data?

(b) Assume that the 2 left most points in the plot above are used for training, and the 2 right most are the test set. What is the squared error on the test set after running linear regression on the training data?
(c) We now consider a new set of data below. What is the squared error using linear regression on this data? (You may assume that the best fit is a horizontal line, i.e., $a = 0$ in our linear model, meaning that $y = c$.)

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Question 2 (Bayesian classification.)
Consider the following classification problem. First, we generate $X \sim \text{Bernoulli}(\frac{1}{2})$, where Bernoulli random variables take the values 0 or 1. Next, we generate $Y$ that depends on the value of $X$; if $X = 1$, then $Y \sim \text{Bernoulli}(p)$, else $Y \sim \text{Bernoulli}(q)$, where you can assume that $p > q$. (Following Homework 5, it may be easier for you to visualize that $Y = 1$ represents a red class, and $Y = 0$ represents a blue class.)

(a) Express $\Pr(X = 0)$, $\Pr(X = 1)$, $\Pr(Y = 0|X = 0)$, $\Pr(Y = 1|X = 0)$, $\Pr(Y = 0|X = 1)$, and $\Pr(Y = 1|X = 1)$ using $p$ and $q$.

(b) Express the joint probabilities, $\Pr(X = 0, Y = 0)$, $\Pr(X = 0, Y = 1)$, $\Pr(X = 1, Y = 0)$, and $\Pr(X = 1, Y = 1)$.

(c) Compute $\Pr(Y = 0)$ and $\Pr(Y = 1)$. 
(d) Express $\Pr(X = 0|Y = 0)$, $\Pr(X = 0|Y = 1)$, $\Pr(X = 1|Y = 0)$, and $\Pr(X = 1|Y = 1)$ using $p$ and $q$.

(e) What is the Bayes optimal classifier? That is, given $Y = 0$ (or $Y = 1$), is $X = 0$ or $X = 1$ more likely? (Recall that $p > q$.)
Question 3 (Ridge regression.)
Consider a vector $\beta$ observed through noisy measurements $y$,

$$y = X\beta + z.$$  

Our goal is to recover or estimate $\beta \in \mathbb{R}^N$, given $X \in \mathbb{R}^{M \times N}$ and $y \in \mathbb{R}^M$. Below, you will show in several steps that when $\beta$ and $z$ are modeled as independent and identically distributed (i.i.d.) Gaussian, maximum a posteriori (MAP) estimation of $\beta$,

$$\beta_{MAP} = \arg \max_{\beta} f(\beta | X, y),$$

is a special case of ridge regression.

(a) We begin deriving the solution $\beta_{MAP}$ that maximizes $f(\beta | X, y)$,

$$\beta_{MAP} = \arg \max_{\beta} f(\beta | X, y)$$

$$= \arg \max_{\beta} \frac{f(\beta, X, y)}{f(X, y)}$$

$$= \arg \max_{\beta} f(\beta, X, y).$$  \hspace{1cm} (1)

Why does the last step, (1), hold?

(b) Focusing on the last term, $f(\beta, X, y) = f(X)f(\beta|X)f(y|\beta, X)$, but $\beta$ is independent of $X$, i.e., $f(\beta|X) = f(\beta)$, and so

$$\beta_{MAP} = \arg \max_{\beta} f(\beta, X, y)$$

$$= \arg \max_{\beta} f(X)f(\beta)f(y|\beta, X)$$

$$= \arg \max_{\beta} f(\beta)f(y|\beta, X).$$  \hspace{1cm} (2)

Why does the last step, (2), hold?
(c) To compute $f(\beta)$ and $f(y|\beta, X)$ in (2), we model each of the $M$ scalar entries in $z \in \mathbb{R}^M$, which can be interpreted as a noise or error vector, as i.i.d. zero-mean Gaussian with variance $\sigma_z^2$. That is, $Z_m \sim \mathcal{N}(0, \sigma_z^2)$, where $Z_m$ is the random variable corresponding to entry $m$ of the noise vector, and $m \in \{1, \ldots, M\}$. The probability density function (pdf) for $Z_m$ can be expressed,

$$f(Z_m = z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{z^2}{2\sigma_z^2}}.$$ 

You are also given that the $N$ entries of $\beta$ are i.i.d. with pdf $f(\beta_n) \sim \mathcal{N}(0, \sigma^2)$, and $n \in \{1, \ldots, N\}$. Derive expressions for $f(\beta)$ and $f(y|\beta, X)$ in terms of $X, y, \sigma_z, \sigma$.

(d) Because the logarithm is a monotone function, it suffices to maximize log ($f(\beta|y, X)$),

$$\beta_{MAP} = \arg \max_{\beta} f(\beta)f(y|\beta, X)$$

$$= \arg \max_{\beta} \text{log}(f(\beta)f(y|\beta, X))$$

$$= \arg \max_{\beta} \text{Polynomial}(\beta, y, X, \sigma_Z, \sigma).$$

Express Polynomial($\beta, y, X, \sigma_Z, \sigma$).
(e) Recall that ridge regression has the form

\[ \beta_{\text{ridge}} = \arg \min_\beta \| y - X \beta \|^2 + \lambda \| \beta \|^2, \]

where \( \| \cdot \|^2 \) is the squared \( \ell_2 \) norm. (Note that the arg max of previous parts becomes an arg min by adding a minus sign.) The ridge regression form should correspond to your expression above; what is \( \lambda \)? What intuition can you draw from the expression for \( \lambda \)? If you could not derive the expression for \( \beta_{\text{MAP}} \), please explain in general how \( \lambda \) impacts the ridge regression solution.

(f) For a ridge regression problem, the following three figures illustrate the solutions for \( \lambda = 1, 10, 100 \), but in the wrong order. Match each of the three figures with its corresponding \( \lambda \) value. Make sure to justify your answer.

![Figure 1](image)

(a) \( \lambda = \)  
(b) \( \lambda = \)  
(c) \( \lambda = \)

Figure 1: Suggest the corresponding \( \lambda \) values from 1, 10, 100.
Question 4 (Image deblurring.)
Imagine a person taking a picture while their hand is shaking. Instead of the camera’s focal plane $y$ measuring the image, $x$, there will be a 2D convolution between $x$ and some convolution kernel $h$, which can be denoted by $h \ast x$ and defined as

$$\{ h \ast x \}_{n_1,n_2} = \sum_{k_1,k_2} h_{k_1,k_2} x_{n_1-k_1,n_2-k_2}.$$

Note that $n_1$ and $n_2$ are horizontal and vertical indices in an image, and $k_1$ and $k_2$ are horizontal and vertical indices used to compute 2D convolution. The measurements are further contaminated by noise,

$$y_{n_1,n_2} = \{ h \ast x \}_{n_1,n_2} + z_{n_1,n_2},$$

which is more conveniently denoted in 2D form, $y = h \ast x + z$.

Imagine that several years from now you work in digital camera design. The camera array gives you the noisy blurry measurements, $y$, an accelerometer provides an estimate of the convolution kernel, $h$, and suppose further that you know the distribution of noise, $z$. How would you estimate the unknown image, $x$?

In your answer, please consider the following aspects. First, how is the deblurring problem, $y = h \ast x + z$, related to other problems we discussed; is this a classification problem, clustering, linear regression, other? Second, how can sparsity information for images help in estimating $x$? Ideally, you should be able to propose a scheme that deblurs $x$ using algorithms and concepts we learned in class.