INCLUDE THIS COVER PAGE WITH YOUR HOMEWORK

NAME:

BANNER ID:

BROWN EMAIL:

COLLABORATED WITH:

(LEAVE THIS BLANK)

<table>
<thead>
<tr>
<th>problem</th>
<th>grade</th>
<th>memo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IMPORTANT: Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently. All of the work submitted should be your own. Each student should write on the problem set the set of people with whom they collaborated.

Problem 1

Let \( f_w(x) = w^T \phi(x) \) be a function defined by a feature map \( \phi \) and a vector of parameters \( w \).

In class we showed that least squares regression with a training set \( T = \{(x_1, y_1), \ldots, (x_N, y_N)\} \) corresponds to the maximum likelihood estimate of \( w \) assuming

1. \( y_i = f_w(x_i) + e_i \)
2. the errors \( e_i \) are independent and distributed according to a Normal distribution with mean 0 and standard deviation \( \sigma^2 \).

Now suppose we have a prior distribution over the parameters \( w \), defined by a multivariate Normal with mean 0 and covariance matrix \( aI \) for some \( a \in \mathbb{R} \). The multivariate Normal (Gaussian) distribution is described in section 2.3 in the textbook.

The Maximum a posteriori (MAP) estimate of \( w \) is the vector maximizing the posterior probability of \( w \) given \( T \),

\[ w_{\text{MAP}} = \max_w p(w|T) \]

(a) Show that there exists a \( \lambda \) such that

\[ w_{\text{MAP}} = \min_w \lambda \frac{1}{2} w^T w + \frac{1}{2} \sum_{i=1}^{N} (f_w(x_i) - y_i)^2 \]

(b) What is the relationship between \( \lambda \), \( a \) and \( \sigma^2 \)?

Problem 2

Let \( x = (x_1, \ldots, x_n) \) be a random vector where each \( x_i \) is a binary random variable. The \( n \) random variables are independent, with \( x_i \) distributed according to a Bernoulli distribution with mean \( u_i \). This leads to a distribution \( p(x|u) \) over the random vectors that is parameterized by a vector of parameters \( u = (u_1, \ldots, u_n) \).

Suppose we have a training set \( T \) with \( k \) independent samples from \( p(x|u) \). Derive the maximum likelihood estimate for \( u \).

Problem 3

A naive Bayes classifier is a simple classifier obtained by assuming independence between features. In this assignment you will implement a naive Bayes classifier to recognize hand-
written digits. There are 10 classes \( y \) corresponding to digits 0 through 9. Each example \( x \) is a 28x28 binary image represented as a 784 dimensional binary vector.

The data for this problem is available on the course website. After loading the data, the training examples are available matrices “train?” where “?” is a digit and the test examples are similarly loaded into matrices “test?”. Each matrix has one example per row and the examples can be reshaped into a 28x28 matrix for visualization as follows.

```matlab
> load('digits');
> A = reshape(train3(43,:),28,28)';
> image(A);
```

Under the naive Bayes model we assume the features (pixel values) are independent conditional on the image class. We will use a different Bernoulli distribution to model each feature (pixel) of each class. Note that in this case \( p(x|y) \) is a product of Bernoulli distributions like in Problem 2.

Let \( u_{y,i} \) denote the mean of the Bernoulli distribution associated with the \( i \)-th feature (pixel) of class \( y \). You can assume that the \( p(y) = 1/10 \) for each class \( y \).

(a) What is the maximum likelihood estimate for the parameters \( u_{y,i} \) given training data \( \{(x_1,y_1), \ldots, (x_n,y_n)\} \)?

(b) Use ML estimation to train a model for each digit using the training data from the course website. Make a visualization of the model for each digit by displaying a 28x28 image where the brightness of each pixel specifies the mean of the Bernoulli distribution associated with that pixel. Include the visualization of the models in your writeup.

(c) Use the resulting models to classify the test data. What fraction of the test digits were correctly classified? Compute and turn in a 10x10 confusion matrix where entry \((i,j)\) specifies how often digit \( i \) was classified as digit \( j \).