ENGN 2520 / CSCI 1950-F Homework 3
Due Friday March 1 by 4pm

Problem 1

Consider binary classification ($Y = \{0, 1\}$) with logistic regression. In logistic regression we assume $p(y = 1|x) = \sigma(w^T \phi(x))$ for a feature vector $\phi(x) \in \mathbb{R}^k$ and parameters $w \in \mathbb{R}^k$.

The likelihood of the data $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ as a function of w is given by

$$p(D|w) = \prod_{i=1}^{n} p(x_i)p(y_i|x_i, w) = \prod_{i=1}^{n} p(x_i) \prod_{i=1}^{n} p(y_i|x_i, w).$$

Since $p(x)$ does not depend on w, to maximize the likelihood $p(D|w)$ we can maximize the conditional likelihood

$$p(D_y|D_x, w) = \prod_{i=1}^{n} p(y_i|x_i, w).$$

(a) Give a simplified formula for the error function $E(w) = -\log(p(D_y|D_x, w))$ in terms of σ, w, y_i and $\phi(x_i)$.

(b) Implement a matlab function that evaluates $E(w)$.

Your function should take:

1) an $n \times k$ matrix Φ where $\Phi(i, :) = \phi(x_i)$.

2) an n vector Y where $Y(i) = y_i$.

3) a vector of parameters w.

The function should return the value of $E(w)$. While you can implement this by iterating over data points in a for loop, doing so will be slow. For full credit your function should avoid for loops whenever possible. Instead you should use matrix operations and built in functions that operates on matrices and vectors like sum,max,mean,etc.

Turn in your matlab implementation.

Problem 2

There is no closed form solution for minimizing $E(w)$. But $E(w)$ is convex and we can minimize it using a general convex optimization package. The optimization package will need to repeatedly compute the value of $E(w)$ and also its gradient.

(a) Give a simplified formula for $\frac{\partial E(w)}{\partial w_j}$ in terms of σ, w, y_i and $\phi(x_i)$.

(b) Implement a matlab function that evaluates $\nabla E(w)$.
Your function should take:
1) an $n \times k$ matrix Φ where $\Phi(i,:) = \phi(x_i)$.
2) an n vector Y where $Y(i) = y_i$.
3) a vector of parameters w.
The function should return the gradient of $E(w)$. Note the gradient is a vector. Again you should avoid for loops whenever possible and use matrix/vector operations instead.

Turn in your matlab implementation.

Problem 3

Now we will use logistic regression for recognizing handwritten digits. We will work with the data from the last assignment, but only with digits 3 and 5.

Consider a binary classification task where $y = 0$ for a 3 and $y = 1$ for a 5. Let x_j be the value of the j-th pixel in image x. Let $\phi(x)$ be a $28 \times 28 + 1$ dimensional vector

$$\phi(x) = [x_1 \; x_2 \; \ldots \; x_{784} \; 1]^T.$$

(a) Use the training data and logistic regression to estimate $p(y|x)$. You should use the optimization routines available on the class website for minimizing $E(w)$. This will use your matlab functions that evaluate $E(w)$ and its gradient. Make a visualization of the resulting model w as a 28x28 image. What do you think the model is capturing?

(b) Suppose we predict the label for a new example by using $\hat{y} = \text{argmax}_y(p(y|x))$. What fraction of the test data is correctly classified by your logistic regresion classifier? Compare this to a generative model that assumes pixel values are independent conditional on the value of y (the approach you used for the last homework, but with only two digits).

Problem 4

Instead of MLE we can use MAP estimation for selecting w. Suppose we have a prior over w corresponding to a multivariate normal with covariance λI. To compute the MAP estimate of w given data D we have to maximize $p(w|D) \propto p(D_y|D_x, w)p(w)$.

(a) How can you modify $E(w)$ so that minimizing $E(w)$ finds the map estimate of w. What is the gradient of the modified $E(w)$?

(b) Give matlab functions that evaluate the modified error function and its gradient. Your functions should now take λ as a parameter as well.

(c) Repeat the experiments in Problem 3 using MAP estimation instead of MLE to obtain w. How does this change the estimated model w? How does this affect the test error?

You will need to pick a value for λ. You should try different values in the range of 0.01 to 10 and discuss the effect of λ on your results.