B565 Homework 6

1. Consider two-dimensional vectors from two different classes. Class 1 consists of $\{(0,1),(1,1),(2,2)\}$ while class 2 consists of $\{(0,-1),(1,-2),(2,-3)\}$. We want to consider a maximum margin classifier that separates these points.

   (a) Are these points linearly separable? Explain why or why not.

   (b) What is the maxium margin: $\max_{w,b} \min_i \frac{c_i(w^t x_i + b)}{||w||}$ subject to $c_i(w^t x_i + b) \geq 0$ for $i = 1, \ldots, n$? Suggestion: do this by reasoning from a figure rather than by computation.

   (c) Write an equation giving the separating hyperplane for the maximum margin classifier.

   (d) Reasoning from the figure, which are the support vectors. That is, which are the vectors whose distance from the separating hyperplane is the margin?

   (e) Using the fact that the $\{x_i\}$ satisfy $c_i(w^t x_i + b) \geq 1$ with equality for the support vectors, solve for w and b.

2. Create a linearly separable collection of 50 two-dimensional labeled data points $\{(x_1, c_1) \ldots, (x_n, c_n)\}$ with $x_i \in [-\pi, +\pi] \times [-\pi, +\pi]$ and $c_i \in \{+1, -1\}$ such that

$$c_i(w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 \cos(x_{i1}) + w_4 \sin(x_{i1})) > 0$$

   for $i = 1, \ldots, n$ for randomly chosen parameters (independent $N(0,1)$) $w_0, \ldots, w_4$.

   (a) Plot the data using different symbols for the two classes.

   (b) Find the optimal parameters $\hat{w}_0, \ldots, \hat{w}_4$ that maximize the margin generated by the level set

$$\{(x_1, x_2) : \hat{w}_0 + \hat{w}_1 x_{i1} + \hat{w}_2 x_{i2} + \hat{w}_3 \cos(x_{i1}) + \hat{w}_4 \sin(x_{i1}) = 0\}$$

   (c) Add the decision boundary to your points by drawing the resulting level set.

3. The College Mall Chipotle store often has long lines. Presumably a customer's willingness to join the line depends on its length. We will measure this length as a real number, x, in meters. We model the probability that a person is willing to join the line as

$$p(\text{join}|x) = \frac{1}{1 + \exp\{w_0 + w_1 x + w_2 x^2\}}$$

   (a) Suppose $w_0 = 3$, $w_1 = -.05$, $w_2 = -.08$ and plot the probability of a customer joining the line as a function of x for x in the range of 0 to 10 meters.

   (b) Estimate the weight parameters using the data set "chipotle.dat" on Canvas, giving the customer decisions for different line lengths encountered. Do this by writing a simple R program for simple gradient descent: taking a small step in the direction of the gradient until the gradient is sufficiently small in norm. You may find that simple gradient descent oscillates wildly if the step size is too large.

   (c) Implement gradient descent for the same data set using the Newton-Raphson method. You may find that that NR oscillates wildly for some starting parameters, so see if you can find an intial choice of the weight vector that leads to convergence.

4. Suppose we have a two-class problem where one class is far more likely than the other class — say a rare disease for instance. We want to model the probability of class 1 given our measured data x.

   (a) Write an equation for $p(c = 1|x)$ where c is the class according to the logistic regression model.

   (b) Logistic regression does not explicitly model the highly asymmetric prior distribution. If you believe that class 1 should be favored regardless of the value of x, how would you express this in the language of logistic regression?

5. Construct a random data set of $K = 4$ 2-d clusters where each cluster has $n = 50$ points. The kth cluster, $k = 1, \ldots, K$ is constructed by choosing a random 2x2 matrix T_k and a random 2x1 point b_k and generating the ith example from the kth cluster, x_{ki} by

$$x_{ki} = T_k z_i + b_k$$

   where the components of z_i are $N(0,1)$. For this problem you should choose the elements of T_k to be $N(0,1)$ and the elements of b_k to be $N(0,10)$

(a) Generate your points as above and plot them with different colors for each cluster.

(b) Implement the K-means algorithm, plotting the current clusters and prototypes for each iteration. Your implmentation should pause the process to allow the user to inspect the current clustering and prototypes. This can be done by requesting character input from the user through "readline()".

(c) Run your algorithm to convergence (no changes in clustering). Plot the number of steps to convergence using 100 random data sets.

6. As discussed in class, K-means tries to minimize the objective function

$$H = \sum_{k=1}^{K} \sum_{i:k(i)=k} ||x_i - m_k||^2 = \sum_{i=1}^{n} ||x_i - m_{k(i)}||^2$$

where $k(i) \in \{1, \ldots, K\}$ is the index of the cluster containing x_i and m_k is the prototype of the kth cluster. We have seen in class, and likely in your experments above, that the resulting clustering is occasionally not what we would hope for.

(a) Does K-means terminate in a local minimum of H?

(b) Does K-means terminate in a global minmum of H?

(c) Explain why K-means occaionsally produces results that are clearly undesirable.