B555: Homework 5

1. We want to classify documents that are drawn independently from K classes with probabilities $\pi = (\pi_1, \ldots, \pi_K)$. For each document we measure D different binary features, $x_1, \ldots, x_D$, assumed to be conditionally independent given the class, but having different probabilities of "success" for each feature and class. For the kth class these probabilities are $q_{k1}, \ldots, q_{kD}$ where

   $$q_{kd} = p(x_d = 1|\text{class } k)$$

   These probabilities are summarized in the $K \times D$ matrix of probabilities, $Q = \{q_{kd}\}$. Thus

   $$p(x_1, \ldots, x_D|\text{class } k) = \prod_{d=1}^{D} q_{kd}^{x_d}(1-q_{kd})^{1-x_d}$$

   We observe N such documents and write $X = \{x_{nd}\}$ for the $N \times D$ binary data matrix of feature observations whose nth observation is $x_{n1}, \ldots, x_{nD}$.

   (a) Write out the log likelihood $\log p(X|\pi, Q)$

   (b) For a document observation, $x_{n1}, \ldots, x_{nD}$, what is the "responsibility," $\gamma_{nk} = p(\text{class} = k|x_{n1}, \ldots, x_{nD})$?

   (c) Given our current values π^{old} and Q^{old} and using the intuition of the EM algorithm, give new estimates π^{new} and Q^{new} using the $\{\gamma_{nk}\}$ as a "soft labeling" of the documents into classes.

   (d) Writing $z = (z_1, \ldots, z_n)$ for the true class of the n documents, write out the complete data log likelihood: $\log p(z, X|\pi, Q)$

   (e) Letting π^{old} and Q^{old} be our current estimates, derive

   $$E_{\pi^{\text{old}}, Q^{\text{old}}} \log p(z, X|\pi, Q)$$

   where the expectation is taken over the posterior distribution on z given $\pi^{\text{old}}, Q^{\text{old}}$

   (f) Show that $\pi_k^{\text{new}} = \frac{\sum_n \gamma_{nk}}{N}$ by maximizing your expected log likelihood from the previous part with respect to π subect to the sum constraint on π.

2. Continuing with the formulation above, now with computation, suppose we have $K = 2$ classes with $\pi = (.3, .7)$ and $D = 3$ features with

   $$Q = \begin{pmatrix} .8 & .5 & .2 \\ .2 & .5 & .8 \end{pmatrix}$$

   (a) In R, sample $N = 1000$ binary D-tuples according to the model described to form the data matrix X.

   (b) Run 10 iterations of the EM algorithm starting from $\hat{\pi} = (.5, .5)$

   $$\hat{Q} = \begin{pmatrix} .6 & .5 & .5 \\ .4 & .5 & .5 \end{pmatrix}$$

   and report your ending estimates of Q and π. For each iteration of the algorithm write out the data log likelihood under the current parameters. This should be monotoncally non-decreasing.

   (c) What happens if we start this process from constant Q matrix and why?

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

   $$y(x) = \frac{\exp\{w_0 + w_1 x + w_2 x^2\}}{1 + \exp\{w_0 + w_1 x + w_2 x^2\}}$$

   (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) Chipotle has a never-ending stream of customers, so it makes sense to consider on-line learning as a possible strategy. The data set "chipotle.dat" on the class web site represents a sequence of observations (x, t) where x is the length of the line and t is the customer's decision to either join (1) or not join (0) the line. Implement on-line learning to develop a sequential stragegy for estimating w_0, w_1, w_2.

(c) For different values of the step size parameter plot the resulting probability estimate on the same range as before. What happens when the step size is too small? Too large?

4. In the multiclass logistic regression problem we have

$$P(C = k|x, w_1, \ldots, w_K) = \frac{e^{w_k^t x}}{\sum_j e^{w_j^t x}}$$

where the weight vectors, w_k, and x have dimension D.

(a) Consider a classification problem where the class variable, $C \in \{1, \ldots, K\}$ has prior distribution $\pi = (\pi_1, \ldots, \pi_K)$ and

$$X|C = k \sim N(\mu_k, \Sigma_k)$$

Show that the Bayes classifier can be expressed as multi-class logistic regression for appropriate choice of parameters $w_1, \ldots, w_K$.

(b) Suppose we add a constant, α, to the dth component of each of the weight vectors. Show that this leaves the resulting class probabilities unchanged. Argue that, with no loss of generality, we can fix the last weight vector, w_K, to have $w_K = 0$ and fit the model using $w_1, \ldots, w_{K-1}$.

5. Download the "Balance+Scale" dataset from https://archive.ics.uci.edu/ml/datasets/Balance+Scale. The last 4 numbers in each row of the dataset are the weight and distance on the left side of a balance and the weight and distance on the right side of a balance. The first number in each row tells us if the configuration will tip to the left (L), tip to the right (R), or balance (B). From the leverage principle we get L/B/R as the product of the first two numbers is greater than/equal to/less than the product of the last two numbers. The data set enumerates all possibilities for values 1 through 5. Obviously these are not random data from an experiment following the assumptions of logistic regression. Rather, this problem explores the fundamental limitations of logistic regression.

(a) Write an R program to perform multiclass logistic regression using these data, estimating the logistic parameters by gradient descent, using only the 2 weights and distances as your features. It is not necessary to involve the Hessian matrix. Simply write a loop that takes a step in the direction of greatest improvement, terminating when the gradient is sufficiently small. Compute the number of errors your resulting classifier makes on this dataset after training by classifying each observation according to the most likely class having observed the observation.

(b) It isn't possible to get good performance on these data due to a mismatch between the strict functional form of logistic regression and this particular problem. Add a new feature that will aid the classifier in detecting balanced configurations and retrain your classifier. Report the number of errors in this case too. You should show a significant improvement.