10-702 Statistical Machine Learning: Assignment 3
Due Friday, February 22

1. (a) Generate data as follows.

\[
\begin{align*}
    n &= 100 \\
    p &= 1000 \\
    X &= \text{matrix}(\text{rnorm}(n*p),n,p) \\
    \beta &= \text{c}(\text{rep}(10,5),\text{rep}(5,5),\text{rep}(0,990)) \\
    y &= X \%\% \beta + \text{rnorm}(n)
\end{align*}
\]

(b) Write an R function to do forward stepwise regression. Apply it to your dataset and summarize the output. The output should include the cross-validation score for the sequence of models.

(c) Run the lasso on these data and compare to the results from stepwise.

(d) Write a function to do ridge regression. Run your code on these data and summarize the results.


   It is a binary classification problem, with 768 instances having eight features each.

   (a) Fit a maximum likelihood logistic regression model using Newton’s method (iteratively reweighted least squares). Summarize your findings.

   (b) Now use sparse logistic regression. Use coordinate-wise descent. Plot the estimates as a function of \( \lambda \). Also, plot the AIC score as a function of \( \lambda \).

3. Let \( x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d \) and let

\[
g(x) = ||x||_p = \left( \sum_{j=1}^{d} |x_j|^p \right)^{1/p}
\]

where \( 0 < p < 1 \). Find the subgradient of \( g \).
4. (a) Write an R program to fit a mixture of multivariate Normals using the EM algorithm (with the number of components $k$ known).

(b) Simulate 1000 observations from the model

$$\frac{1}{5}N(0, I) + \frac{4}{5}N(\mu, I)$$

where $\mu = (3, 3, 3, 3, 3)$ and $I$ is the 5 by 5 identity matrix. Use your program to find the MLE.

(c) Now repeat but take $k$ as unknown. Use AIC and BIC to choose $k$.

5. In this problem you will use mixture models for a classification task. The training data consist of 100 examples $(x_i, y_i)$, where $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ and each $x_{ij} \in \mathbb{R}^{100}$ is a 100-dimensional vector, and $y_i = \{c_{ij}\}$ is an unordered set of up to five labels, with each $c_{ij} \in \{1, 2, \ldots, 10\}$. Two examples are shown below (check them out in color in the electronic version):

In the first example, there are four classes $y = \{1, 5, 6, 9\}$. Each of the five $x_i$ was generated from exactly one of these four classes; thus, one of the four classes must have generated two of the $x_i$. Considering the alignment of each of the five images with one of the classes as a latent variable $z = (z_1, z_2, z_3, z_4, z_5)$, a possible alignment is $z = (9, 5, 9, 1, 6)$ where class 9 generates $x_1$, class 5 generates $x_2$, class 9 generates $x_3$, and so on.
The test data consist of 100 unlabeled examples $x_i$. The task is to label each of the test examples with an ordered set of labels $y_i = (c_{i1}, c_{i2}, c_{i3}, c_{i4}, c_{i5})$ where $c_{ij}$ is the predicted class of $x_{ij}$.

You may use any method and model you choose, but it should make use of mixture models in some way. Give a detailed explanation of your approach and methodology. The training data and test data are on the course website.

Describe your method clearly and report a summary of how well you do on the test data.