CS 5135/6035 Learning Probabilistic Models

Exercise Questions for Lecture 9: MV Gaussian MLE, Logistic Regression

Gowtham Atluri

2/13/2020

Questions

1. For the dataset 'Davis' from the package RDatasets, **[1+2+1+2+2+2 points]**
   a. Load the data using the code:
   ```
   using RDatasets;
   data = dataset("car","Davis");
   ```
   b. Write the equations for estimating the mean and variance of the two variables 'Height' and 'Weight', treating them as univariate variables, assuming they follow Gaussian distribution. Estimate the mean and variance.

   c. Plot a two-dimensional histogram of variables 'Height' and 'Weight'. State your observations pertaining to the dependence or independence between the two variables. Use the code:
   ```
   using Plots;
   gr()
   histogram2d(data[:Weight],data[:Height],nbins=40,xlabel="Weight",ylabel="Height")
   ```
   d. Write the equations for estimating the mean vector and covariance matrix assuming they follow a multivariate Gaussian distribution. Estimate the mean vector and the co-variance matrix.
   e. Write Julia code to draw 1000 samples from this multivariate distribution and plot the 2d histogram of the resultant data. Comparing this histogram with that generated in (c), does the estimated multivariate distribution appear to be a good fit to the data?
   f. From the covariance matrix, determine if the variables 'Height' and "Weight" are independent.

2. Logistic regression using two independent variables. Use the 'iris' dataset from the package RDatasets to answer the following questions. **[1+1+2+2+2+2+2+3 points]**
   a. Using Julia, select the samples for species 'versicolor' and 'virginica'. *Hint:* Load data and select relevant samples using the code:
   ```
   using RDatasets;
   iris = dataset("datasets", "iris");
   groups = groupby(iris,:Species);
   new_iris = vcat(groups[2],groups[3]);
   ```
   b. Generate a scatter plot between variables 'PetalLength' and 'PetalWidth', coloring each point based on 'Species'. (Use Gadfly.plot() with Geom.point option)
   ```
   using Gadfly;
   Gadfly.plot(new_iris, x=:PetalLength, y=:PetalWidth,color=:Species,Geom.point);
   ```
   c. Considering the goal of performing logistic regression using 'PetalLength' and 'PetalWidth' as independent variables and 'Species' as a dependent variable, write the functional form for the conditional probability $f(y|\boldsymbol{x})$.
   d. Write equations for conditional likelihood and log conditional likelihood.
   e. Write equations for the first gradient with respect to parameters $\beta_0, \beta_1, \beta_2$.
   f. Write update equations for second derivatives with respect to parameters $\beta_0, \beta_1, \beta_2$.
   g. Write Newton's algorithm for parameter estimation (with update equations for $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$.
   h. Implement Newton's algorithm in Julia and report the estimated parameters $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.

Bonus Questions

1. Logistic regression when x is univariate.

   a. Using Julia, select variables x and y using the following code. x captures heart weight in grams. $y = 1$ indicates male and $y = 0$ indicates female.

   ```
   data = dropmissing(dataset("MASS","cats"));
   x = data[:,2];
   x = [ones(length(x)) x];
   y = Int.(data[:,1].=="M");
   ```

   b. Visualize the shape of the log conditional log likelihood function using the code

   ```
   b0 = collect(-40:40); #select the points along b0
   b1 = collect(-40:40); #select the points along b1
   heatm = zeros(81,81);
   for i=1:length(b0)
     for j=1:length(b1)
       heatm[i,j]= compute_l(x,y,[b0[i] b1[j]]');
                   #computing l at each b0, b1 combination
     end
   end
   #plot surface
   using Plots;
   Plots.surface(b0, b1,heatm)
   ```

   c. Estimate the parameters β and visualize the fit using the following code and comment (i) if the data was amenable to fit a sigmoid function (ii) if the estimation was the best possible estimation.

   ```
   # estimate b using newton method
   b = newtons_lr(x,y,[-4 1]')
   # compute the sigmoid function
   x1  = collect(-5:0.01:5);
   p_fit = 1./(1.+e.^(-(b[1].+b[2].*x1)));
   # plot the points and overlay the learned sigmoid function
   Gadfly.plot(layer(x=x[:,2],y=y,Geom.point),layer(x=x1,y=p_fit, Geom.line))
   ```

   d. Generate the data (x, y) using the following code, Estimate β and visualize the fit using the code in (c). Comment (i) if the data was amenable to fit a sigmoid function (ii) if the estimation was the best possible estimation.

   ```
   data = dropmissing(dataset("MASS","cats"));
   x = data[:,2];
   x = [ones(length(x)) x];
   y = Int.(data[:,1].=="M");
   x[y.==0,2] = x[y.==0,2]-2;
   ```

   e. Generate the data (x, y) using the following code, Estimate β and visualize the fit using the code in (c). Comment (i) if the data was amenable to fit a sigmoid function (ii) if the estimation was the best possible estimation.

   ```
   data = dropmissing(dataset("MASS","cats"));
   x = data[:,2];
   x = [ones(length(x)) x];
   y = Int.(data[:,1].=="M");
   x[y.==0,2] = x[y.==0,2]-1;
   ```

   f. What differences did you notice in the data samples (x, y) that are used in (c), (d), and (e).

2. Exploring Newton's method.

   a. Plot the shape of the log conditional likelihood for the logistic regression in Q2 above.
   b. Based on your plot, comment if your estimate is a global optimum.
   c. Plot the path taken by the Newton's method.
   d. Implement gradient descent algorithm for estimating parameters for the above logistic regression.

e. Plot the path taken by the gradient descent algorithm. Do you notice any significant differences in the paths taken my the two methods? Which of the two methods is more efficient?