CMSC 828G:Spring 2002

Sketch of Solutions to Homework 1

- Problem 1
  - Part A
    - Single real valued variable
      - Check the point value to see whether or not it lies within the valid range.
      - Run statistics tests like chi squares.
      - Visually by using histograms and boxplots.
      - Compute the sample mean μ and choose those points within the range $\mu \pm \lambda$ for some λ.
    - Set of 20 real valued variables
      - Check range of each variable.
      - Check for invalid combinations of values.
      - Visually by using Scatter plots, Trellis plots.
      - Use PCA and Clustering Algorithms.
    - Categorical variables
      - Partition the data based on possible values of the categorical variable and analyze each set using the previous techniques.
      - Visually by using plots for the observations within each category.
      - Use Clustering Algorithms
  - Part B
    Mean is more sensitive to outliers than Median.
  - Part C
    - Use PCA for dimension reduction then apply the trimmed α technique over each component.
    - Transform each point to a scalar by using techniques like reduced sub-ordering based on distance measure then apply the trimmed α technique over the scalars.
    - For each value of a specific variable, apply the trimmed α technique over the other variables. That's for a simple two dimension case X and Y, for each value of X check the possible Y values. Unusual Y values are outliers that can be removed by using the trimmed α technique.
  - Part D
    The problem can not be completely automated, and it is not usually possible to detect outliers without having a prior model for the data.
- Problem 2
  - Part A This part should be solved in two steps.

- * Step 1

    Take the partial derivatives $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$ and set both to 0 and simplify to obtain two equations in two unknowns that can be solved simultaneously.

    $$S(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2 = \sum_{i=1}^{n}(a^2x_i^2 - 2ax_iy_i + 2abx_i + b^2 - 2by_i + y_i^2)$$

    $$\frac{\partial S(a,b)}{\partial a} = \sum_{i=1}^{n} \frac{\partial}{\partial a}(a^2x_i^2 - 2ax_iy_i + 2abx_i + b^2 - 2by_i + y_i^2) = \sum_{i=1}^{n}(2ax_i^2 - 2x_iy_i + 2bx_i) = 0$$

    Similarly

    $$\frac{\partial S(a,b)}{\partial b} = \sum_{i=1}^{n} \frac{\partial}{\partial b}(a^2x_i^2 - 2ax_iy_i + 2abx_i + b^2 - 2by_i + y_i^2) = \sum_{i=1}^{n}(2ax_i + 2b - 2y_i) = 0$$

    Solving the two equations yields

    $$a = \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) - n\sum_{i=1}^{n} x_iy_i}{(\sum_{i=1}^{n} x_i)^2 - n\sum_{i=1}^{n} x_i^2}$$

    and

    $$b = \frac{(\sum_{i=1}^{n} x_iy_i)(\sum_{i=1}^{n} x_i) - (\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i)}{(\sum_{i=1}^{n} x_i)^2 - n\sum_{i=1}^{n} x_i^2}$$

  * Step 2

    Check to see that the values obtained for a and b in Step 1 actually minimizes S by taking the second derivative and check to see if it is positive.

    $$\frac{\partial^2 S(a,b)}{\partial a^2} = \sum_{i=1}^{n}(2x_i^2)$$

    and

    $$\frac{\partial^2 S(a,b)}{\partial b^2} = \sum_{i=1}^{n}(2)$$

    Clearly these values are positive.

- Part B

  * Use the trimmed α technique.
  * Replace sum by the median in least squares formula, that's median$(y(i) - (ax(i) + b))^2$
  * Select a random set of points and get its fit. Repeat several times and choose the set with the smallest error.