Cody Morgan

CS398 Machine Learning

Ola Amayri

* Result Analysis
  + Data plots
    - Many of these variable relations are not that unexpected. For example, these have a positive correlation: ZN, CHAS, RM, DIS, B. While these have a negative correlation: CRIM, NOX, AGE, TAX, PTRA, LSTAT, RAD and INDUS.
    - We can say that property value goes up with an increase in:
      * (ZN) other big houses / big plots
      * (CHAS) being near a river
      * (RM) having more rooms
      * (B) I’m not really sure what to make of this variable. Its actual definition is a parabola using the number of black people as it’s independent variable (B(x) = 1000(x-0.63)2). So, it seems that there is some “ideal” amount of black people according to statistician(s) who created this document. Using linear regression, it looks like the trend is positive, so it would seem that property value goes up as the so called “black concentration” goes up. The thing that I really don’t understand is why 63%? I would understand measuring the actual concentration of a demographic (which we still do to this day), but why the 63%? This would suggest that low density black areas have high property value, and high density black areas *also* have high property values. What would cause this?
        + A little more light is shown on this if we reverse the formula and determine an average Bk value – i.e how many black people are there on average?
        + B = 1000(bk – 0.63)2 => bk =
        + Using 356 as an average I find that the average bk value is ~1.2 and ranges from 1 to 1.3
        + So, I think this small range has something to do with the potentially misleading nature of this variable
    - Property value goes down as these variables increase:
      * (AGE) architecture from <=1940
      * (TAX) high property tax value
      * (PTRATIO) high student to teacher ratio. This is another surprising statistic. I would think that high student to teacher ratios would indicate higher property values. Perhaps it indicates a good school – but it seems to not positively impact the surrounding properties.
      * (LSTAT) the poorer your neighbors are – the less your house is worth
      * (RAD) There isn’t a great relationship here – the range of home values for each x value goes almost top to bottom – which isn’t too unexpected. Proximity to highway(s) just isn’t a valuable commodity
      * (INDUS) having more industrial property nearby lowers your home value.
  + Residual plots
    - I’ll call out a few examples here, as I don’t think it’s necessary to talk about each one.
    - (CHAS) and (RAD) both demonstrate my point in that they have high and low residual values – indicating that there is a lot of error on a lot of points in both directions.
    - (RM) the residual here is mostly near the horizontal zero line – indicating that there is less error than the aforementioned two.
  + Heat map
    - The first thing I notice is the black line along the (CHAS) variable. This seems to indicate that proximity to a river has little correlation with any other variable. This is reasonable as there isn’t a lot of property along river bends compared to the rest of a city / state. Though there is a weak correlation to low student-teacher ratios along rivers (which is reinforced by the negative correlation between PTRATIO and MEDV)
    - (AGE) and (DIS) also have a negative correlation and it seems reasonable that older houses are farther away from employment centers.
  + Multi variable
    - I chose (LSTAT), PTRATIO) and (RM) as my three variables as I feel that they will make the greatest impact on property value. The overall R2 is negative, which is fine, but my RMSE value is > 5 which leads me to believe that a linear model may not be the best approach. This is expected as very few human attributes can be measured linearly.
  + Overview
    - This is an *okay* predictor for house value if certain values are used and others are ignored. We want to ignore those with low correlations ( for example CHAS) and prioritize those variables with higher correlation (like RM and LSTAT). There is a lot of error inherent in a linear model – so this may be the best we can do without other models (such as quadratic)
  + Code overview
    - BestFit.cs
      * Single Variable Regression drawing
        + Y=mx+b
        + M = (n \* SumXY() - (SumX() \* SumY())) / (n \* SumXX()) - ((SumX() \* SumX()))
        + B = meanY - m \* meanX
        + See Update() line 30
    - DrawPoints.cs
      * Draws points on the graph
    - DropGeneration.cs
      * Drop menu ui management
    - HeatMap.cs
      * Generates heat map
      * Uses generated r values to color squares on a range (red)[-1,1] (green)
      * See GenerateMap() line 28
    - InputManager.cs
      * Allows you to escape
    - LineSet.cs
      * Draws line between two points (for single variable regression)
    - LinearRegressionToggle.cs
      * Toggle residual plotting
    - LoadFile.cs
      * Load the data file – do not move or otherwise change the data file
    - MultiVarGen.cs
      * Generates the multivariable linear regression equation and associated values
      * Using normal theta equation
      * See BuildMatrix() line82 and 102 for xmatrix and ymatrix generation and EvalEquationAt() line 143 for equation evaluation
    - OptionSlider.cs
      * Manage multi variable regression parameters
    - ToggleHeatmap.cs
      * Toggle heap map
    - ToggleMV.cs
      * Toggle mutli variable screen
    - ToggleResidual.cs
      * Toggle residual plots
    - ToolTip.cs
      * Display parameter data
      * See OnMouseOver() line 19 and [LoadFile.cs].FindRValues() line 134 for rValue generation