Semana 12

Wednesday, October 20, 2021 7:50 PM

Regression

Predicting a numerical variable

- Approach to prediction an outcome for an individual:
  - Find other who are like that individual
  - And whose outcomes you know
  - Use those outcomes as the basis of your prediction.

Nearest Neighbor Regression

A method for predicting a numerical variable y, given a value of x:

- Identify the group of points for which the values of x are close to the given value
- The prediction is the average of the y values for the group.

Graph of averages

- For each value of x, the predicted value of y is the average of the y values of the nearest neighbors.
- Graph these predictions for all the values of x. That's the graph of averages.
- If the association between the two variables is linear, then points on the graph of the average tend to fall on or near a straight line. That's the regression line.

Regression estimate

To estimate y based on a given value of x

- Convert the given x to standard units.
- Multiply the result by r. That's the estimate of y, in standard units.
  $Estimate\ of\ y_{(su)} = r\ x\ given\ x_{(su)}$
- Convert the estimate to the original units of y.

Regression to the mean

$Estimate\ of\ y_{(su)} = r\ x\ given\ x_{(su)}$

- The regression estimate of y is closer to the mean than the given value of x is.
- The regression estimate is an average. On average, the values of y at a fixed are closer to the mean than x is.
- "Regression to the mean" is a statement about averages. It is not true for all individuals.

Regression in the original units

The regression line

- Passes through the point of averages
  (average of x, average of y)
- Has slope r * (SD of y) / (SD of x)

Equation of Regression Line

$Estimate\ of\ y = slope\ x\ X + intercept$

Slope of the regression line = $r\ x\ \frac{SD\ of\ y}{SD\ of\ X}$

Intercept of the regression line = average of y - slope * average of x

Scatter plots of other shapes

- We have been able to come up with a good straight line to use for prediction when the scatter

diagram is football shaped.
- The same equation gives the best straight line to use for prediction no matter what the shape of the scatter diagram.

Least squares

Error in estimation

- Error = actual value - estimate
- Typically, some errors are positive and some negative
- To measure the rough size of the errors:
    - Square the errors to eliminate cancellation,
    - Take the mean of the squared error,
    - Take the square root to fix the units.
- The result is called root mean square error (rmse)

the regression line is the line that minimizes the root mean squared error over the data from which that regression line is computed.

Least Squares Line

- Minimizes the root mean squared error (rmse) among all line.
- Equivalently, minimizes the mean squared error (mse) among all lines.
- Names:
    - "Best fit" line
    - Least squares line
    - Regression line

Numerical optimization

- Numerical minimization is approximate but effective.
- Lots of machine learning uses numerical minimization.
- If the function `mse(a, b)` returns the mse of estimation using the line "estimate=ax+b"
    - Then `minimize (mse)` returns an array [a, b]
    - A is the slope and b the intercept of the line that minimizes the mse among all lines with any slope a and any intercept b (that is, among all lines)