Linear regression, quality of fit

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 15, Fall 2018

Monday March 10

Plan

- Linear regression
  - Geometric version
  - Normal equations
- Quality of fit and R^2

Linear regression (1/2)

- In many applications, we want to fit a curve to data to analyze the overall pattern.
- If we use **regression**, the curve is an analytically defined function. We can summarize a large dataset with one equation!
- Curve fitting better than joining dots with a piecewise curve. **We don't want to trust every data point and every digit of precision.**

Linear regression (2/2)

- We want the curve to get as close as possible to as many data points as possible.
- We want to pick a relatively simple function to fit the data.

Least squares

- We will use the **least-squares** method of determining the "best fit".
- The least-squares problem involves defining **basis functions** $f_j(x), j = 1, \ldots, n$ and unknown fit coefficients c_j that satisfy the following equation:

$$F(X) = c_1 f_1(x) + c_2 f_2(x) + \cdots + c_n f_n(x)$$

Least-squares (1/2)

- Given data $(x_i, y_i), i = 1 \ldots, m$ least-squares method returns us the c_j such that $F(x_i) \approx y_i$.
  - i.e. recover the coefficients so that our curve evaluated at x_i $(F(x_i))$ is close to the y data coordinate.
- If $f_1(x) = x$ and $f_2(x) = 1$, then we have the **simple linear regression** $F(x) = c_1 x + c_2$, fitting to a straight line.

Least-squares (2/2)

- Strictly speaking, **linear regression only requires $F(x)$ to be linearly dependent on c_j**, NOT that F(x) is a linear function. For example, we could do linear regression with $f_1(x) = x^2, f_2(x) = x, f_3(x) = 1, f_4(x) = sin(x)$.

Supplemental slides

Simple linear regression

Solving for the coefficients α, β with m data points:

$$\begin{bmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & m \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{bmatrix}$$

yields

$$\alpha = \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right) - m\left(\sum_i x_i y_i\right)}{\left(\sum_i x_i\right)^2 - m\sum_i x_i^2}$$

$$\beta = \frac{\left(\sum_i x_i\right)\left(\sum_i x_i y_i\right) - m\left(\sum_i x_i^2\right)\left(\sum_i y_i\right)}{\left(\sum_i x_i\right)^2 - m\sum_i x_i^2}$$

Normal equations

For matrix A with independent variable data inputs and y column vector of dependent variable data outputs, solving the **normal equations** gives us the c undetermined regression model coefficients.

$$\left(A^T A\right) c = A^T y$$

R^2 and quality of fit

How much is our model buying us, over a simpler model (data mean)?

$$R^2 = \frac{||r||_2^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\text{model error}}{\text{sum squared deviation from y-mean}}$$

where $||r||_2^2 = \sum_i (y_i - \hat{y}_i)^2$ and $\bar{y}$ is the y data mean.