I3344-Digital Modeling & Simulation

problem: given a set of "x"s and their respectful "y"s. Estimate y_n for a given X_n.
$0 \to i$

Solved by interpolation & extrapolation.
$0 \le n \le i$ — n outside range

Old way: polynomial regression (fitting a curve that goes through all points)

not suitable for cases of high noise/error

Better way: Least-Square Approximation by Linear Regression

	x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e_i^2
n	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Σx	Σy	Σx^2	Σxy			

$(\Sigma x)^2$

$$\hat{y}_i = a_1 x_i + a_0$$

$$e_i = y_i - \hat{y}_i$$

$$a_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = r\frac{\sigma_y}{\sigma_x}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

(Sum of the Squares of the Residual) $SSE = \Sigma (e_i)^2$

Normalize the error: $MSE = \frac{SSE}{n}$

$RMSE = \sqrt{MSE}$

$$NRMSE = \frac{RMSE}{\bar{y}}$$

close to 0 ⇒ Perfect model

close to 1 ⇒ Null model

- K-Nearest Neighbors:

  for a given X_n, we can approximate $\hat{y}_n$ by:
  - take K (usually $K = \sqrt{n}$)
  - for every X_i, calculate $d_i = \sqrt{(X_n - X_i)^2}$
  - take the "y"s of the K least "d"s and $\hat{y}_n = \frac{y + y + \dots + y}{K}$

- K-fold cross validation: taking the set as K subsets. Each iteration we set one of them as the validation set and the rest as a training set. We fit the function on the latter (training) and test it on the former (validation)

  * KNN: per iteration, calculate $\hat{y}$ for each point in the validation set using the K nearest neighbors from the training set only and get NRMSE.

  * Linear Regression: get a_1 & a_0 using the training set and calculate NRMSE in the validation set using these values per iteration

Cross validation: $CV = \frac{\Sigma NRMSE}{K}$

the method with the lowest CV is better.

Multiple Linear Regressions: one y and multiple xs.

$$\hat{Y} = X * \beta \quad \text{with} \quad \beta = (X^t X)^{-1} X^t Y$$

$$R^2 = 1 - \frac{SSE}{TSS} \quad \begin{matrix} \nearrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ \searrow \sum_{i=1}^{n} (y_i - \bar{y})^2 \end{matrix}$$

~ 1 Perfect Model

~ 0 Null Model.

problems: R^2 always $\nearrow$ as we add variables (xs) but that doesn't mean a better model.

fix: adjusted R^2: $R^2_a = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

where p = number of predictors

ex)

A	B	C	D
1	2	14	3
3	5	23	4
7	9	17	8
4	5	19	6

$Y = A = [1, 3, 7, 4]$

for $X = B = \begin{pmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 9 \\ 1 & 5 \end{pmatrix}$, $R^2_a = 0.973$

$X = C$, $R^2_a = -0.4$

$X = D$, $R^2_a = 0.95$

pick B for second iteration:

$X = BC = \begin{pmatrix} 1 & 2 & 14 \\ 1 & 5 & 23 \\ 1 & 9 & 17 \\ 1 & 5 & 19 \end{pmatrix}$, $R^2_a = 0.94$

$X = BD$, $R^2_a = 1$ ✓

Notes for solving by hand:

- $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$, $A^t = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}$

- $A = \begin{pmatrix} 4 & 10 & 10 \\ 10 & 30 & 26 \\ 10 & 26 & 26 \end{pmatrix}$, $\begin{pmatrix} 4 & 10 & 10 \\ 10 & 30 & 26 \\ 10 & 26 & 26 \end{pmatrix} \begin{matrix} \nearrow \\ \nearrow \\ \nearrow \end{matrix} \begin{pmatrix} 4 & 10 & 10 \\ 10 & 30 & 26 \\ 10 & 26 & 26 \end{pmatrix}$

3000 + 2704 + 2600 = top

3120 + 2600 + 2600 = bottom

$|A|$ = top − bottom = 120 − 104 = 16

- $A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$, $cof(A) = \begin{pmatrix} \begin{vmatrix} e & f \\ h & i \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i \end{vmatrix} & \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ -\begin{vmatrix} b & c \\ h & i \end{vmatrix} & \begin{vmatrix} a & c \\ g & i \end{vmatrix} & -\begin{vmatrix} a & b \\ g & h \end{vmatrix} \\ \begin{vmatrix} b & c \\ e & f \end{vmatrix} & -\begin{vmatrix} a & c \\ d & f \end{vmatrix} & \begin{vmatrix} a & b \\ d & e \end{vmatrix} \end{pmatrix}$

$$A^{-1} = \frac{1}{|A|} cof(A)$$