Regression

© Malay K. Das, 210 Southern Lab, ph-7359, mkdas@iitk.ac.in

Office hours: W 1030-1130, SL-210

Previously:

Linear Regression

Matrix-vector notation

Today:

1. Linear regression

2. Validation and regularization

HW due:

August 16, 2024

1

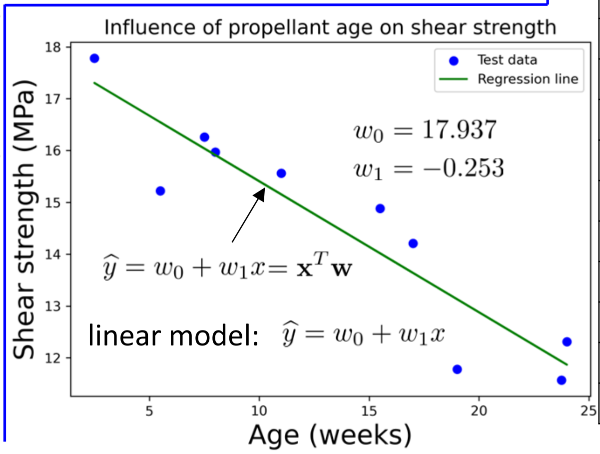
Malay K. Das, mkdas@iitk.ac.in

$$\mathbf{x} = \begin{bmatrix} 1 & x \end{bmatrix}^T \quad \mathbf{w} = \begin{bmatrix} w_0 & w_1 \end{bmatrix}^T \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix} = \begin{bmatrix} 14.88 \\ 11.57 \\ \vdots \\ 15.56 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 15.5 \\ 1 & 23.75 \\ \vdots & \vdots \\ 1 & 11 \end{bmatrix} \quad \mathbf{X}^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 15.5 & 23.75 & \dots & 11 \end{bmatrix}$$

Linear regression solves

$$\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

$$\Rightarrow \mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 17.937 \\ -0.253 \end{bmatrix}$$



Shear strength of rocket propellant

Test	Propellant age (Weeks) x_i	Shear strength (MPa) y_i
1	15.5	14.88
2	23.75	11.57
3	8	15.97
4	17	14.21
5	5.5	15.22
6	19	11.78
7	24	12.31
8	2.5	17.78
9	7.5	16.26
10	11	15.56

2

Mathematically

Least square problems intends to solve linear equations with no solution!!

Consider linear equations with no solution $\mathbf{Au} = \mathbf{b}$

Least square solution minimizes $\|\mathbf{Au} - \mathbf{b}\|$

in regression problems, usually

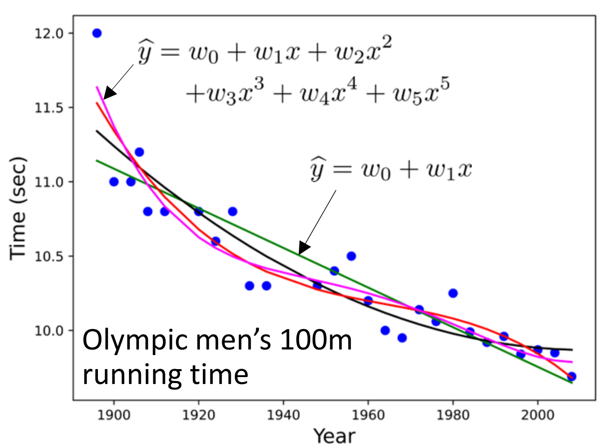
$\mathbf{A} \in \mathbb{R}^{m \times n} \quad m > n \quad \mathrm{rank}\,(\mathbf{A}) = n$

Two approaches

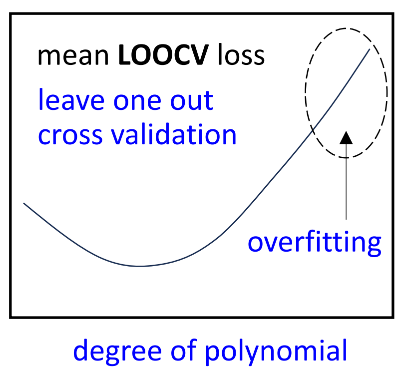
- direct minimization of $\|\mathbf{Au} - \mathbf{b}\|$
- solving normal equations $\mathbf{A}^T\mathbf{Au} = \mathbf{A}^T\mathbf{b}$
  - use of linear solvers (such as Gaussian elimination etc.)
    - $\mathbf{A}$ is often ill-conditioned, linear solvers accumulate errors
  - use of matrix decomposition (factorization)
    - $\mathbf{QR}$ decomposition
    - singular value decomposition $(\mathbf{SVD})$

Malay K. Das, mkdas@iitk.ac.in

Validation for hypothesis selection



Calculating R^2 (or similar parameters) are necessary but not enough



LOOCV: n-1 data are used for training, 1 for test; calculated n times changing the test data and the losses are averaged; hypothesis with minimum loss is selected

For large number of data, K-fold cross validation is used

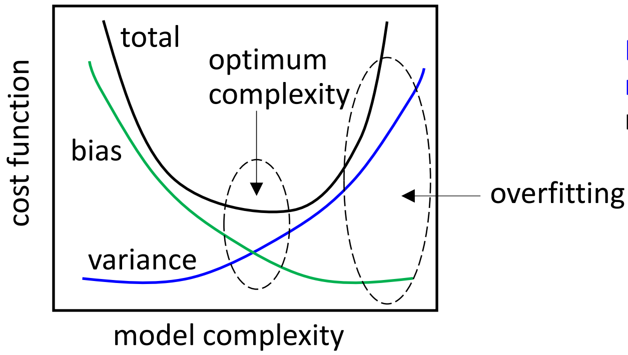
Bias-variance trade-off

Fitting a complex function (such as higher degree polynomial)

– less cost in training, high cost in testing (bias error)

Fitting a simpler function (such as a lower degree polynomial)

– higher cost in training, lower cost in testing (variance error)



Bias-variance trade-off is not limited to regression alone, applicable to various machine learning algorithms

End goal is to generalize, but not too much

Malay K. Das, mkdas@iitk.ac.in

Overfitting is the outcome of noise creeping into the signal

difficult to avoid with noisy data

Regularization is a procedure to control overfitting

consider fitting a linear hypothesis: $\widehat{y} = \mathbf{x}^T \mathbf{w}$

In **regularized regression,** we define a cost function $\underbrace{E = \frac{1}{n} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) + \overset{\text{penalty term}}{\lambda \mathbf{w}^T \mathbf{w}}}_{\text{Ridge regression (Tikonov regularization)}}$

$\lambda :$ penalty parameter

$\lambda \to 0 :$ classical least square regression

$\lambda \to \infty : \quad \widehat{y} \to 0$

minimization of E requires

$$\nabla E(\mathbf{w}) = \mathbf{0} \Rightarrow \left(\mathbf{X}^T \mathbf{X} + n\lambda \mathbf{I} \right) \mathbf{w} = \mathbf{X}^T \mathbf{y}$$

Thus regularization tends to reduce the model complexity by reducing $\mathbf{w}$

Optimum value of λ is decided based on cross-validation