Least Square Regression

CE282: Linear Algebra

Computer Engineering Department

Sharif University of Technology

Hamid R. Rabiee

Maryam Ramezani

Note

- Remember the regression model (affine function) :

$$\hat{f}(x) = x^T\beta + v$$

- The prediction error for example i is:

$$\begin{aligned} r^{(i)} &= y^{(i)} - \hat{f}(x^{(i)}) \\ &= y^{(i)} - \left(x^{(i)}\right)^T\beta - v \end{aligned}$$

- The MSE is :

$$\frac{1}{N}\sum_{i=1}^{N}(r^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \left(x^{(i)}\right)^T\beta - v)^2$$

Least Squares Regression

- choose the model parameters v, β that minimize the MSE

$$\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)}-\left(x^{(i)}\right)^{T}\beta-v\right)^{2}$$

this is the least square problem: minimize $\left\|A\theta-y^{d}\right\|^{2}$ with

$$A=\begin{bmatrix}1 & \left(x^{(1)}\right)^{T}\\ 1 & \left(x^{(2)}\right)^{T}\\ \vdots & \vdots\\ 1 & \left(x^{(N)}\right)^{T}\end{bmatrix},\qquad \theta=\begin{bmatrix}v\\ \beta\end{bmatrix},\qquad y^{d}=\begin{bmatrix}y^{(1)}\\ y^{(2)}\\ \vdots\\ y^{(N)}\end{bmatrix}$$

we write the solution as $\hat{\theta}=(\hat{v},\hat{\beta})$

Example

$$\hat{f}(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \ ... + \theta_p x^{p-1}$$

- a linear-in-parameters model with basis functions…….
- least squares model fitting in matrix notation?

Generalization And Validation

important

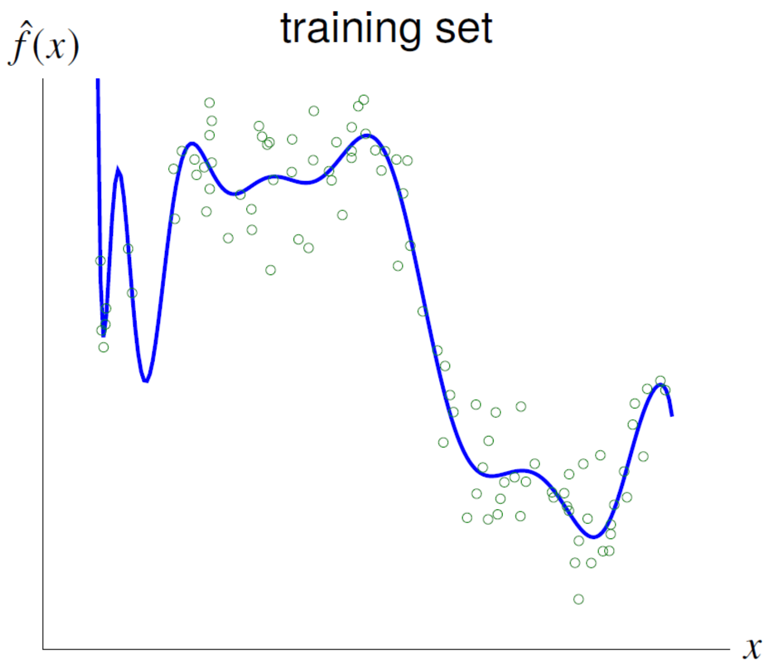
- Generalization ability: ability of model to predict outcomes for new, unseen data

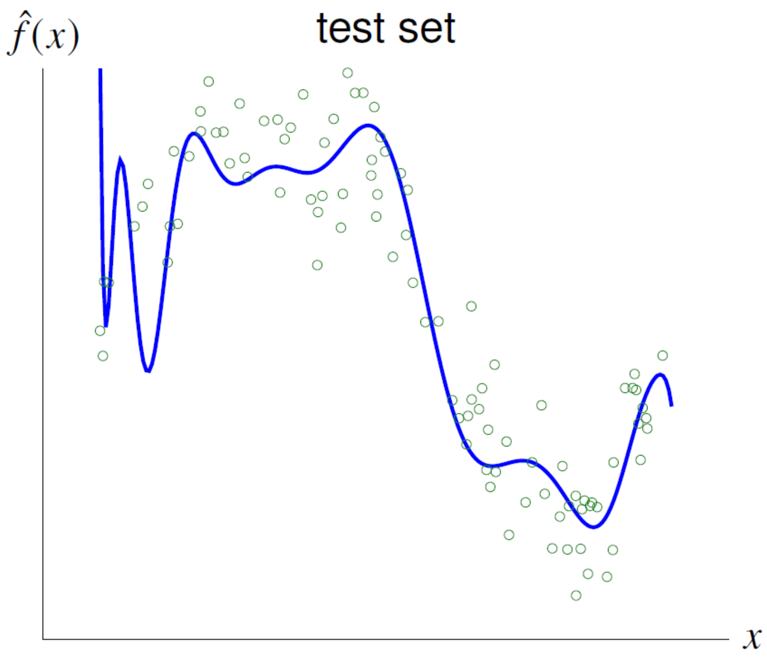
- Model validation: to access generalization ability,
  - divide data in two sets: training set and test (or validation) set
  - use training set to fit model
  - use test set to get an idea of generalization ability
  - this is also called out-of-sample validation

- Over-fit model
  - model with low prediction error on training set, bad generalization ability
  - prediction error on training set is much smaller than on test set

Over-fitting

- Polynomial of degree 20 on training and test set





over-fitting is evident at the left end of the interval

Cross-validating

important

- an extension of out-of-sample validation
  - divide data in K sets (*folds*); typical values are $K = 5, K = 10$
  - for $i = 1$ to K, fit model i using fold i as test and other data as training set
  - compare parameters and train/test RMS errors for the K models
- Remember the house price problem (data set of $N = 774$ house sales)

House price model with 5 folds (155 or 154 examples each)

	Model parameters								RMS error	
Fold	v	β_1	β_2	β_3	β_4	β_5	β_6	β_7	Train	Test
1	122.5	166.9	−39.3	−16.3	−24.0	−100.4	−106.7	−26.0	67.3	72.8
2	101.0	186.7	−55.8	−18.7	−14.8	−99.1	−109.6	−17.9	67.8	70.8
3	133.6	167.2	−23.6	−18.7	−14.7	−109.3	−114.4	−28.5	69.7	63.8
4	108.4	171.2	−41.3	−15.4	−17.7	−94.2	−103.6	−29.8	65.6	78.9
5	114.5	185.7	−52.7	−20.9	−23.3	−102.8	−110.5	−23.4	70.7	58.3

Boolean (two-way) classification

problem

- a data fitting problem where the outcome y can take 2 values +1, -1

  values of y represent two categories (true/false, spam/not spam, ….)

  Model $\hat{y} = \hat{f}(x)$ is called a *Boolean classification*

Least squares classifier

- use least squares to fit model $\tilde{f}(x)$ to training set $\left(x^{(1)},\ y^{(1)}\right), \dots, \left(x^{(N)},\ y^{(N)}\right)$
- $\tilde{f}(x)$ can be a regression model $\tilde{f}(x) = x^T\beta + v$ or linear in parameters

$$\tilde{f}(x) = \theta_1 f_1(x) + \cdots + \theta_p f_p(x)$$

- Take sign of $\tilde{f}(x)$ to get a Boolean classifier

$$\hat{f}(x) = sign(\tilde{f}(x)) = \begin{cases} +1, & if\ \ \tilde{f}(x) \geq 0 \\ -1, & if\ \ \tilde{f}(x) < 0 \end{cases}$$

Multi-class classification

problem

- a data fitting problem where the outcome y can takes values $1, \dots, K$
- values of y represent K labels or categories
- multi-class classifier $\hat{y} = \hat{f}(x)$ maps x to an element of $\{1, 2, \dots, K\}$

Least squares multi-class classifier

- for $k = 1, \dots, K$, compute Boolean classifier to distinguish class k from not k

$$\hat{f}_k(x) = \mathrm{sign}(\tilde{f}_k(x))$$

- define multi-class classifier as

$$\hat{f}_k(x) = \underset{k=1,\dots,K}{\mathrm{argmax}}\ \tilde{f}_k(x)$$

Multi-objective least squares

Important

we have several objectives

- $$J_1 = ||A_1 x - b_1||^2, \dots, J_k = ||A_k x - b_k||^2$$
  - A_i is an $m_i \times n$ matrix, b_i is an m_i-vector
  - we seek one x that makes all k objectives small
  - usually there is a trade-off: no single x minimizes all objectives simultaneously

  Weighted least squares formulation: find x that minimizes

- $$\lambda_1 ||A_1 x - b_1||^2 + \cdots + \lambda_k ||A_k x - b_k||^2$$
  - coefficients $\lambda_1, \dots, \lambda_k$ are positive weights
  - weights λ_i express relative importance of different objectives
  - without loss of generality, we can choose $\lambda_1 = 1$

Regularized data fitting

Theorem

- consider linear-in-parameters model

$$\hat{f}(x) = \theta_1 f_1(x) + \cdots + \theta_p f_p(x)$$

  we assume $f_1(x)$ is the constant function 1
- keeping $\theta_2, \dots, \theta_p$ small helps avoid over-fitting

$$J_1(\theta) = \sum_{k=1}^{N} \left(\hat{f}\left(x^{(k)}\right) - y^{(k)}\right)^2, \qquad J_2(\theta) = \sum_{j=2}^{p} \theta_j^2$$

$$\text{minimize} \quad J_1(\theta) + \lambda J_2(\theta) = \sum_{k=1}^{N}\left(\hat{f}\left(x^{(k)}\right) - y^{(k)}\right)^2 + \lambda \sum_{j=2}^{p} \theta_j^2$$

Example

- $$\text{minimize} \quad J_1(\theta) + \lambda J_2(\theta) = \sum_{k=1}^{N} \left(\hat{f}\left(x^{(k)}\right) - y^{(k)}\right)^2 + \lambda \sum_{j=2}^{p} \theta_j^2$$
  - λ is positive regularization parameter
  - equivalent to least squares problem: minimize

$$\left\| \begin{bmatrix} A_1 \\ \sqrt{\lambda} A_2 \end{bmatrix} \theta - \begin{bmatrix} y^d \\ 0 \end{bmatrix} \right\|^2$$

with $y^d = \left(y^{(1)}, \ldots, y^{(N)}\right)$,

$$A_1 = \begin{bmatrix} 1 & f_2(x^{(1)}) & \cdots & f_p(x^{(1)}) \\ 1 & f_2(x^{(2)}) & \cdots & f_p(x^{(2)}) \\ \vdots & \vdots & & \vdots \\ 1 & f_2(x^{(N)}) & \ldots & f_p(x^{(N)}) \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

  - stacked matrix has linearly independent columns (for positive λ)
  - value of λ can be chosen by out-of-sample validation or cross- validation

Nonlinear least squares

note

- find $\hat{x}$ that minimizes

$$\|f(x)\|^2 = f_1(x)^2 + \cdots + f_m(x)^2$$

- optimality condition: $\nabla \left\|f(\hat{X})\right\|^2 = 0$

  any optimal point satisfies this

  points can satisfy this and not be optimal

  can be expressed as $2Df(\hat{X})^T f(\hat{X}) = 0$

  $Df(\hat{X})$ is the $m \times n$ derivative or Jacobian matrix,

$$Df(\hat{X})_{\mathrm{ij}} = \frac{\partial f_i}{\partial x_j}(\hat{x}), \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, n$$

optimality condition reduces to normal equations when f is affine

SVD and Least Squares

- Solving $Ax = b$ by least squares
- x=pseudoinverse(A) times **b**
- Compute pseudoinverse using SVD
  - Lets you see if data is singular
  - Even if not singular, ratio of max to min singular values tells you how stable the solution will be
  - Set $1/\sum_i$ to 0 if $\sum_i$ is small (even if not exactly 0)

SVD and Least Squares

Theorem

- If $\boldsymbol{A}$ is a $n \times n$ square matrix and we want to solve $\boldsymbol{A}\,\boldsymbol{X} \;=\; \boldsymbol{b}$, we can use the SVD for $\boldsymbol{A}$ such that

$$\boldsymbol{U}\sum\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{x} = \boldsymbol{b}$$

$$\sum\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{x} = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{b}$$

solve

$\sum\boldsymbol{y} = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{b}$(diagonal matrix, easy to solve!)

Evaluate: $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y}$

Cost of solve: $\boldsymbol{O}(\boldsymbol{n}^{2})$

Cost of decomposition $\boldsymbol{O}(\boldsymbol{n}^{3})$ (recall that SVD and LU have the same cost asymptotic behavior, however the number of operations – constant factor before $\boldsymbol{n}^{3}$ - for the SVD is larger than LU)

Class activity

Class Activity

Given the actual values [2, 4, 6, 8] and the predicted values [3, 5, 7, 9], what is the Mean Squared Error (MSE)?

References

References

- Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares, Stephen Boyd Lieven Vandenberghe
- Linear Algebra and Its Applications, David C. Lay