Least squares

Linear Algebra

Department of Computer Engineering

Sharif University of Technology

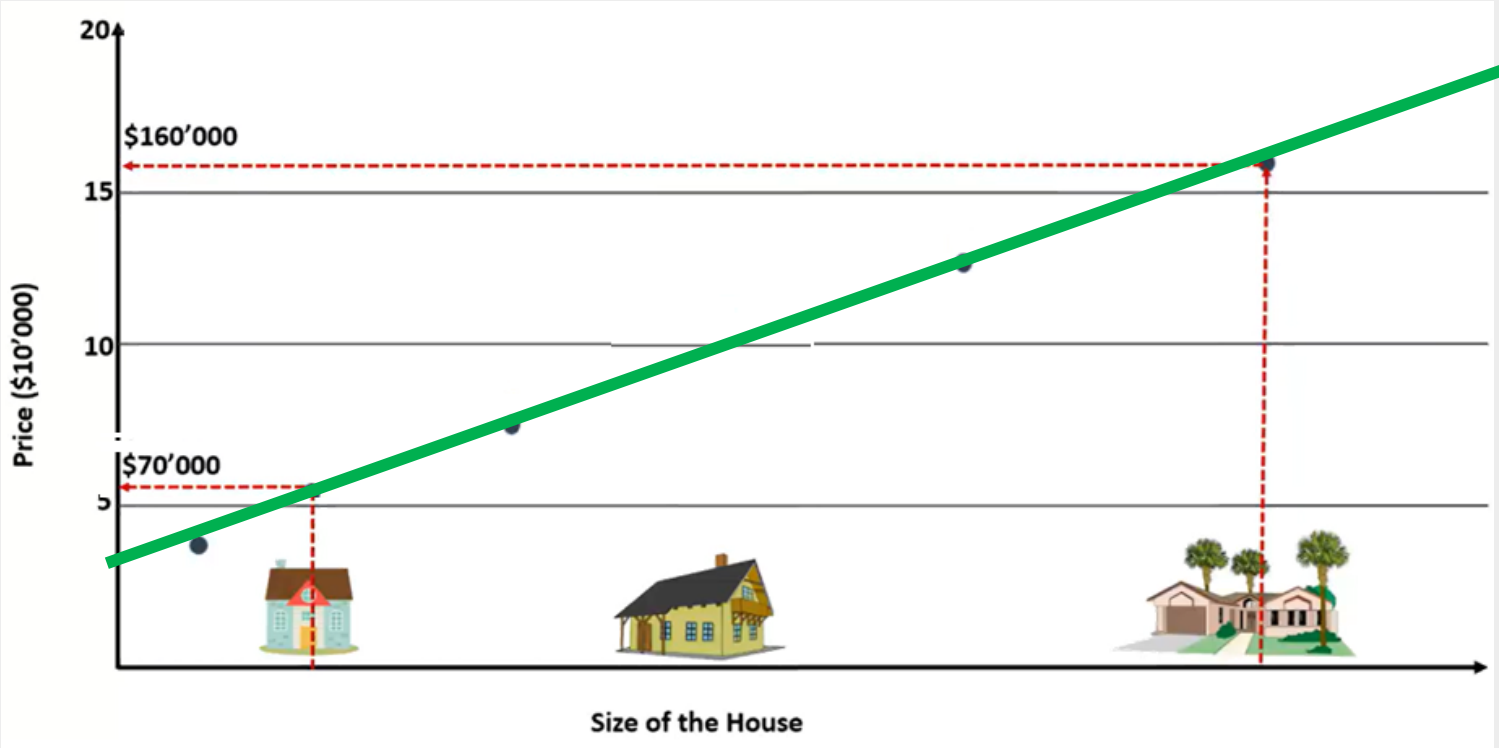
Hamid R. Rabiee rabiee@sharif.edu

Maryam Ramezani maryam.ramezani@sharif.edu

Introduction

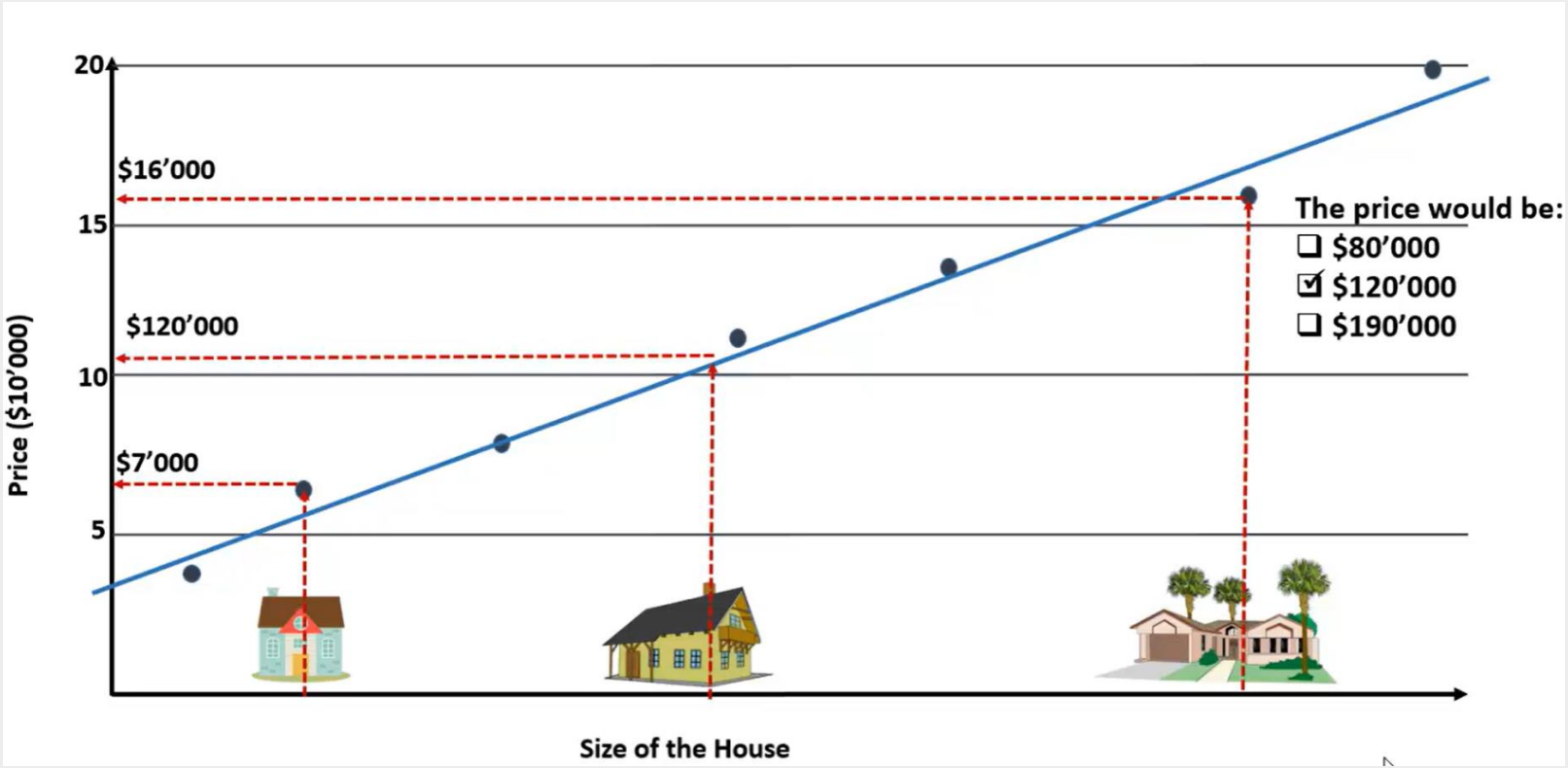
Linear Equation

$Ax = b$ has solution.



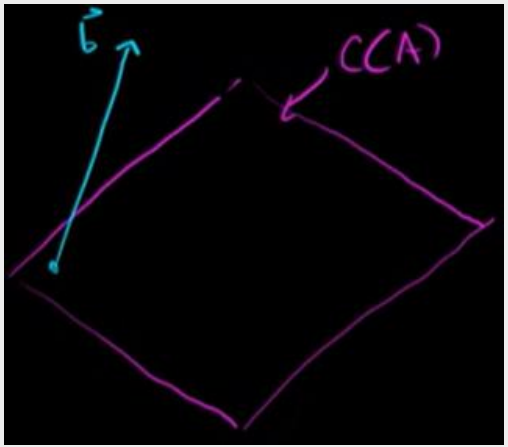
Least Squares Error Correction

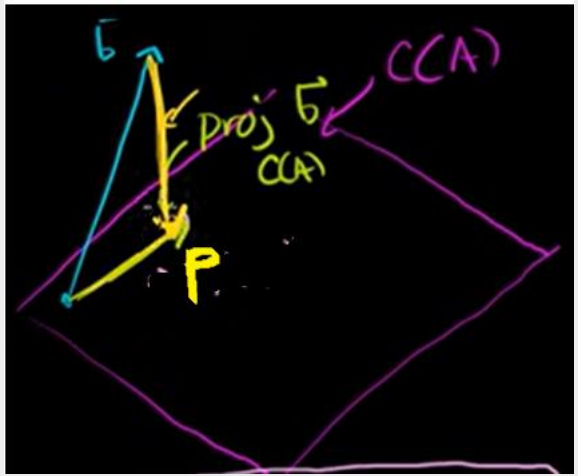
$Ax = b$ has no solution.



What is the problem?

- A is $m \times n$ matrix
- $Ax = b$ has no solution -> b is not in the $C(A)$ why?





How to solve the problem?

- Bad News: $\boldsymbol{Ax = b}$ has no solution
- Good News: $\boldsymbol{A\hat{x} = p}$ has solution

Review

- 4 Subspaces:
  - Column Space $C(A)$
  - Null Space $N(A)$
  - Row Space $C(A^T)$
  - Null Space of A^T= Left Null Space of A = $\mathrm{N}(A^T)$

Theorem

- Orthogonality of the Row Space and the Null Space
- Orthogonality of the Column Space and the Left Null Space

Review

- Projection a vector on a vector
  - Column space of matrix?
  - Rank of matrix?
  - Is the matrix symmetric?
  - Power two o this matrix?
- Projection a vector on a plane

Fill this page with my notes on the board 😄

Lets think!!!

$$P = A(A^T A)^{-1} A^T$$

- Think about Ps when:
  - s is in the column space of A
  - s is in the orthogonal complement space of A

  - Geometry?
  - Math?

Fill this page with my notes on the board 😄

Good News!

- Fill this page with my notes on the board 😄
  - Least square in $\mathbb{R}^2$ and regression!!!
  - Error
  - Outlier

Look another way!!

- $(A^TA)^{-1}A^T$ is the left inverse of A
- $A(A^TA)^{-1}A^T$ is the projection matrix on C(A)

$$\hat{x} = (A^TA)^{-1}A^Tb$$

What will happen when A is an invertible matrix?

Another problem?

$$\hat{x} = (A^T A)^{-1} A^T b$$

Theorem

- If A has linearly independent columns, then $A^T A$ is invertible.

What will happen when $A^T A$ is not an invertible matrix?

Good News!!!

$$\hat{x} = (A^T A)^{-1} A^T b$$

In fact, given any real $m \times n$-matrix A, there is always a unique x^+ of minimum norm that minimizes $\|Ax - b\|^2$, even when the columns of A are linearly dependent.

$$\hat{x} = (A^T A)^{-1} A^T b$$

$$= A^{\dagger} b$$

pseudo-inverse of a left-invertible matrix

SVD

Least squares problem

Theorem

- given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, find vector $x \in \mathbb{R}^n$ that minimizes

$$||Ax - b||^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - b_i \right)^2$$

- ''least squares'' because we minimize a sum of squares of affine functions:

$$||Ax - b||^2 = \sum_{i=1}^{m} r_i(x)^2, \qquad r_i(x) = \sum_{j=1}^{n} A_{ij} x_j - b_i$$

- the problem is also called the linear least squares problem

Least squares and linear equations

Important

$$\text{minimize } ||Ax - b||^2$$

solution of the least squares problem: any $\hat{x}$ that satisfies

- $$||A\,\hat{x} - b|| \leq ||Ax - b|| \quad \text{for all } x$$

Note

$\hat{r} = A\hat{x} - b$ is the residual vector

if $\hat{r} = 0$, then $\hat{x}$ solves the linear equation $Ax = b$

if $\hat{r} \neq 0$, then $\hat{x}$ is a least squares approximate solution of the equation

in most least squares applications, $m > n$ and $Ax = b$ has no solution

Solving least squares problems (Method 1)

Example

- Normal equations of the least squares problem $A^T Ax = A^T b$
  - Coefficient matrix $A^T A$ is the .....
  - Equivalent to $\nabla f(x) = 0$ where $f(x) =$
  - All solutions of the least squares problem satisfy the normal equations

$$\hat{x} = (A^T A)^{-1} A^T b$$

Look at board I am writing in vector and matrix form with derivation

Solving least squares problems (Method 2): QR factorization

Example

- Rewrite least squares solution using QR factorization $A = QR$

- Complexity: $2mn^2$

Algorithm: Least squares via QR factorization

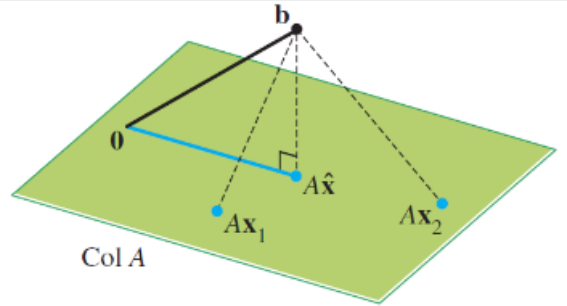
Input: A : $m \times n$ left-invertible
Input: b : $m \times 1$
output: x_{LS} : $n \times 1$
Find QR factorization $A = QR$
Compute $Q^T b$
Solve $Rx_{LS} = Q^T b$ using back substitution

- Identical to algorithm for solving $Ax = b$ for square invertible A, but when A is tall, gives least squares approximate solution

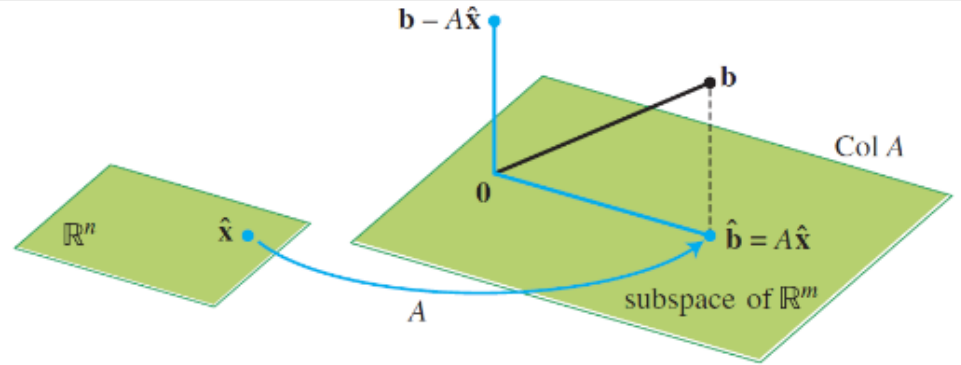
Normal equation

Note

The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ coincides with the nonempty set of solutions of the normal equations $A^T A\mathbf{x} = A^T\mathbf{b}$.



The vector $\mathbf{b}$ is closer to $A\hat{\mathbf{x}}$ than to $A\mathbf{x}$ for other $\mathbf{x}$.



The least-squares solution $\hat{\mathbf{x}}$ is in $\mathbb{R}^n$.

Solution of a least squares problem

Theorem

- A has linearly independent columns, then below vector is the unique solution of the least squares problem

$$\text{minimize } ||Ax - b||^2$$

$$\hat{x} = (A^T A)^{-1} A^T b$$

$$= A^{\dagger} b$$

pseudo-inverse of a left-invertible matrix

- Proof?

Solving least squares problems

Example

a 3×2 matrix with “almost linearly dependent” columns

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 10^{-5} \\ 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 10^{-5} \\ 1 \end{bmatrix},$$

round intermediate results to 8 significant decimal digits

- Solve using both methods
  - Which one is more stable? Why?

Review: Linear-in-parameters model

Note

- we choose the model $\hat{f}(x)$ from a family models

$$\hat{f}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_p f_p(x)$$

model parameters

scalar valued basis functions (chosen by us)

Solution of weighted least squares

Example

weighted least squares is equivalent to a standard least squares problem

$$\text{minimize} \quad \left\| \begin{bmatrix} \sqrt{\lambda_1} A_1 \\ \sqrt{\lambda_2} A_2 \\ \vdots \\ \sqrt{\lambda_k} A_k \end{bmatrix} x - \begin{bmatrix} \sqrt{\lambda_1} b_1 \\ \sqrt{\lambda_2} b_2 \\ \vdots \\ \sqrt{\lambda_k} b_k \end{bmatrix} \right\|^2$$

- Solution is unique if the $stacked\ matrix$ has linearly independent columns
- Each matrix A_i may have linearly dependent columns (or be a wide matrix)
- if the stacked matrix has linearly independent columns, the solution is

$$\hat{x} = \left(\lambda_1 A_1^T A_1 + \cdots + \lambda_k A_k^T A_k\right)^{-1} (\lambda_1 A_1^T b_1 + \cdots + \lambda_k A_k^T b_k)$$

Lagrange multiplier

Example

$$f(x) = \min(x_1 x_2)$$

$$g(x) = 1 - \mathrm{x}_1 - \mathrm{x}_2$$

$$g(x) = 0$$

$$L(x, \lambda) = f(x) + \lambda g(x)$$

$$\nabla f(x)$$

Constrained Least Square

Example

- $\begin{cases} \min\limits_{x} ||Ax-b||^2 & A: m \times n \\ s.t. \quad Cx = d & C: p \times n \end{cases}$

$$L(x,\lambda) = ||Ax-b||^2 + \lambda^T(Cx-d)$$

$$\begin{cases} \nabla_x L = 2A^TAx - 2A^Tb + C^T\lambda = 0 \\ \nabla_\lambda L = Cx - d = 0 \end{cases} \rightarrow \begin{bmatrix} 2A^TA & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} 2A^Tb \\ d \end{bmatrix}$$

Note

- #equations: $n+p$ #Unkowns: $n+p$
- KKT equations
- Least Square problem is a KKT problem with $A = I, b = 0$

Note

- Remember the regression model (affine function) :

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

- The prediction error for example i is:

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

- 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 $\hat{\theta} = (\hat{v}, \hat{\beta})$

Least Squares Regression

Example

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

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

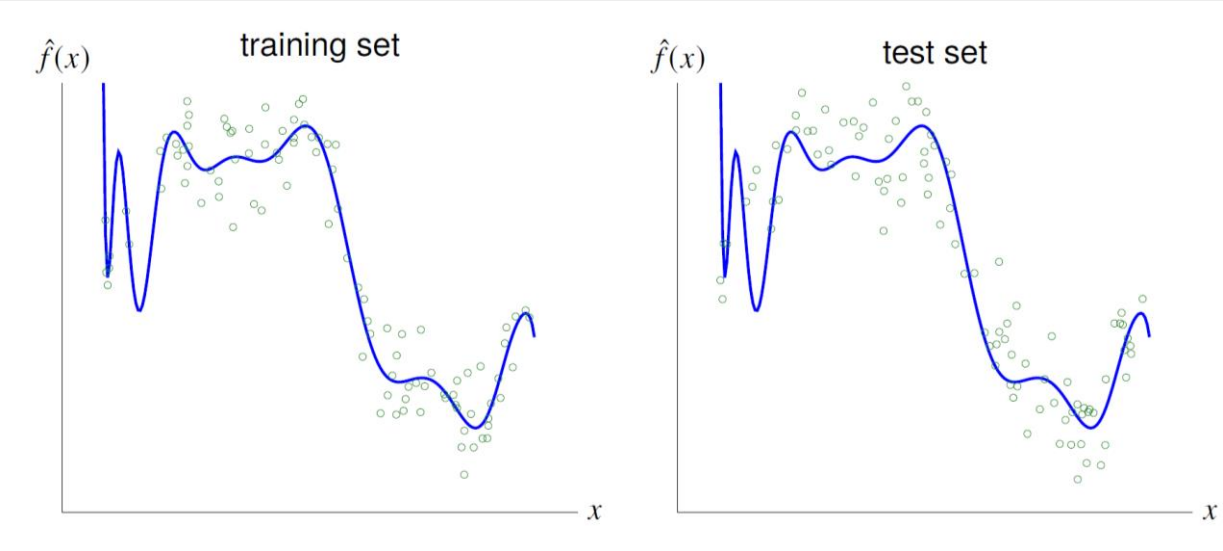
Generalization And Validation

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 $(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$
- $\tilde{f}(x)$ 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

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

- define multi-class classifier as

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

Multi-objective least squares

Important

we have several objectives

$$J_1 = \left||A_1x - b_1|\right|^2, \dots, J_k = \left||A_kx - b_k|\right|^2$$

- 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

$$\lambda_1\left||A_1x - b_1|\right|^2 + \cdots + \lambda_k\left||A_kx - b_k|\right|^2$$

- 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 $\theta_2, \ldots, \theta_p$ small helps avoid over-fitting

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

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

Solution for Weighted least squares

Example

- minimize $J_1(\theta) + \lambda J_2(\theta) = \sum_{k=1}^{N} \left(\hat{f}\left(x^{(k)}\right) - y^{(k)}\right)^2 + \lambda \sum_{j=2}^{p} \theta_j^2$
  - 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: $\nabla \left\|f\left(\hat{X}\right)\right\|^2 = 0$

  any optimal point satisfies this

  points can satisfy this and not be optimal

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

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

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

optimality condition reduces to normal equations when f is affine