

MATH466/MATH766

Math of machine learning

01/13 Lecture 2 regression models

References:

- Ch2, 3 of The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman

Todays contents:

- linear regression
- ridge regression
- LASSO regression (Least Absolute Shrinkage and Selection Operator)

Important concepts:

- ERM (Empirical Risk Minimization)
 - MSE (Mean Square Error), least square
- regularization
 - ℓ_2 regularization
 - ℓ_1 regularization
- sparsity

Recommend reading:

- .

warm up : 1. $w \in \mathbb{R}^d$, $\|w\|_2^2 = \underline{\hspace{2cm}}$, $\|w\|_1 = \underline{\hspace{2cm}}$

2. given $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\lambda \geq 0$.

for $w \in \mathbb{R}^d$, let $R(w) = \frac{1}{2} \|Xw - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2$,

$\nabla R(w) = \underline{\hspace{2cm}}$

3. $X^T X$ is and therefore real-diagonalizable

Let eigen-decomp of $X^T X$ be $U \Lambda U^T$

$X^T X$ is invertible iff

$(X^T X)^{-1} = \underline{\hspace{2cm}}$ if $X^T X$ is invertible.

1. Empirical Risk Minimization (ERM)

Recall: Data. $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ (assume $n \geq d$)

Regression: y_i takes value in some continuous space.

Goal: learn to predict y given x .

e.g. learn a function $f(x)$

such that $y = f(x)$ is a good prediction.

f is a good prediction

\Leftarrow empirical risk $R(f) := \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$ is small

(
 loss function: $l: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$
 e.g. $l(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$, $l(\hat{y}, y) = |\hat{y} - y|$)

The idea of ERM: find f such that the empirical risk is minimized

$$\min_f R(f) \quad (*)$$

issues with (*) no restriction on $f \Rightarrow \begin{cases} \text{overfitting} \\ \text{impractical to solve} \end{cases}$

① add restriction $\min_{f \in \mathcal{F}} R(f) \quad (*_{\mathcal{F}})$

e.g. $\mathcal{F} = \{f: f \text{ is continuous}\}$

② parameterize functions in \mathcal{F} with $\theta \in \mathbb{H}$, i.e. $f(x; \theta)$

$$\min_{\theta \in \mathbb{H}} R(\theta) := \frac{1}{n} \sum_{i=1}^n l(f(x_i; \theta), y_i) \quad (*_{\mathbb{H}})$$

e.g. $\mathcal{F} = \{f: x \mapsto w^T x + b\}$, $\theta = (w, b) \in \mathbb{R}^{d+1}$, $\mathbb{H} = \mathbb{R}^{d+1}$

Different choices of \mathcal{F} and l leads to different ERM models.

2. (Least Square) Linear Regression

2.1 Problem formulation

WLOG, assume the data are centered: $\sum_{i=1}^n \underline{x}_i = 0$.

consider all linear models $\mathcal{F} = \{ f: \underline{x} \mapsto \underline{w}^\top \underline{x} \}$

consider squared error loss $l(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$

$$\min_{\underline{w} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^n (\underline{w}^\top \underline{x}_i - y_i)^2 \quad (*)$$

Q: Is this method good enough? NO

explanation from $\begin{bmatrix} \text{linear algebra} \\ \text{statistics} \\ \text{geometric (optional)} \end{bmatrix}$ P.O.V

2.2. Linear algebra P.O.V

$$\text{Denote } \underline{X} := \begin{bmatrix} \underline{x}_1^\top \\ \vdots \\ \underline{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \underline{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\Rightarrow \text{optimal } \hat{\underline{w}} \text{ satisfies } \underline{X}^\top \underline{X} \hat{\underline{w}} = \underline{X}^\top \underline{y}$$

① If $\underline{X}^\top \underline{X}$ is invertible, $\hat{\underline{w}} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{y}$

predict label/response of \underline{x} is $\underline{x}^\top \hat{\underline{w}}$

Q: if the input is perturbed by some noise,
how does the prediction change?

② $X^T X$ is not invertible $\Leftrightarrow X$ does not have full-rank

$X^T X \in \mathbb{R}^{d \times d}$, we assume $n \geq d$ \Updownarrow

columns of X are linearly dependent

columns of X represents
different features of input x



one or more quantitative features
are coded redundantly.

in practice, features

may not be "perfectly" correlated, \rightarrow
but can "highly" correlated

$X^T X$ is close to
singular
 \Updownarrow

the smallest eigen-value of $X^T X$ is close to 0.



some entries of \hat{w} can be very large



the prediction $X^T \hat{w}$ can be sensitive
to perturbation on x

2.3. Statistics p.o.v : Gauss - Markov Theorem

Statement:

Suppose that $\underline{y} = \underline{X}\underline{w} + \underline{\varepsilon}$. ($\underline{y}, \underline{\varepsilon} \in \mathbb{R}^n$, $\underline{w} \in \mathbb{R}^d$, $\underline{X} \in \mathbb{R}^{n \times d}$)

w_j are deterministic but unobservable

X_{ij} are deterministic and observable

$E\varepsilon_i = 0$, $\text{Var}(\varepsilon_i) = \sigma^2 < \infty$, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ when $i \neq j$.

(\hookrightarrow Assume ground truth is linear model,

observation of label / response is perturbed by white noise)

i.e. $y = x^T w + \varepsilon$ for any x

Consider all linear unbiased estimator \hat{w} of w ,

i.e. $\hat{w} = Ay$, $E\hat{w} = w$.

For any input $x \in \mathbb{R}^d$, let $x^T \hat{w}$ be the estimated response.

Define the Mean Square Error (MSE) of the estimation as

$$E[(x^T \hat{w} - x^T w)^2]$$

Then the ordinary least square estimator

$$\hat{w}_{OLS} := (X^T X)^{-1} X^T y$$

is the one with the smallest MSE for every x .

and is called the best linear unbiased estimator (BLUE)

and the smallest MSE is $\sigma^2 x^T (X^T X)^{-1} x$

Interpretation:

If we stick to linear **unbiased** estimator of w

the smallest error variance we can achieve is

$$\sigma^2 x^T (X^T X)^{-1} x$$

\uparrow

can be very bad when $X^T X$ is close to singular.

inherited from true model,

can not avoid

appears b/c we choose unbiased estimator

(choose to stick to data)

"overfitting"

Proof of theorem (optional)

$$\textcircled{1} \quad E(\hat{w}_{OLS}) = (X^T X)^{-1} X^T E y = (X^T X)^{-1} X^T X w = w$$

i.e. \hat{w}_{OLS} is a linear unbiased estimator

$$\textcircled{2} \quad \text{Recall } \text{Var}(\lambda^T \hat{w}) = \lambda^T \text{Var}(\hat{w}) \lambda.$$

It is sufficient to show for any linear unbiased estimator $\hat{w} \neq \hat{w}_{OLS}$

$\text{Var}(\hat{w}) - \text{Var}(\hat{w}_{OLS})$ is p.s.d.

Denote $C = (X^T X)^{-1} X$, let another linear unbiased estimator be

$$\hat{w} = (C+D)y, \text{ where } D \text{ is non-zero.}$$

$$\text{Then } E\hat{w} = (C+D)Xw + (C+D)E\varepsilon$$

$$= (I + DX)w$$

since \hat{w} is unbiased and w is unobservable, $DX = 0$.

$$\text{Then } \text{Var}(\hat{w}) = \text{Var}((C+D)y)$$

$$= (C+D) \text{Var}(Xw + \varepsilon) (C+D)^T = \sigma^2 (C+D)(C+D)^T$$

$$= \sigma^2 (CC^T + DC^T + CD^T + DD^T)$$

$$= \underbrace{\sigma^2 CC^T}_{\text{Var}(\hat{w}_{OLS})} + \underbrace{\sigma^2 DD^T}_{\text{p.s.d.}} + \sigma^2 (DC^T + CD^T)$$

$$DC^T = DX(X^T X)^{-1} = 0, \quad CD^T = (X^T X)^{-1}(DX)^T = 0$$

$$= \text{Var}(\hat{w}_{OLS}) + \sigma^2 DD^T$$

$$\text{i.e. } \text{Var}(\hat{w}) - \text{Var}(\hat{w}_{OLS}) \succcurlyeq 0. \quad \square.$$

3. Ridge Regression (LR with ℓ_2 regularization)

$$\text{LR: } \min_{w \in \mathbb{R}^d} \frac{1}{2n} \|Xw - y\|^2$$

Lesson from linear regression:

$(X^T X)^{-1}$ can be very bad when $X^T X$ is close to singular

Formulation of Ridge Regression:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2 \quad (*_r)$$

\uparrow
regularization

$\lambda > 0$ is a regularization coefficient we pick

Q: Why the regularization helps?

①

②

③

Regularization is a common strategy in machine learning and other ill-posed inverse problems.

4. LASSO (LR with ℓ_1 regularization)

Motivation: want the prediction model to be simple.

↔ only a small portion of features contribute.

↔ want w to be sparse

mathematically: let $\|w\|_0 = \sum_{i=1}^d \mathbb{1}_{\{w_i \neq 0\}}$

counts the number of non-zero entries in w .

($\|\cdot\|_0$ is called ℓ_0 -norm but is not a norm)

want $\|w\|_0$ to be small.

Formulation: $\min_{w \in \mathbb{R}^d} \frac{1}{2n} \|Xw - y\|_2^2 + \lambda \|w\|_1$ (*)_1

\uparrow
regularization

Q: Why not using ℓ_0 -norm as a regularization?

What's the benefit of ℓ_1 -regularization?

Answer in Optimization Module

Q: Does ℓ_1 -regularization encourages sparsity? Why?

Discuss later (01/22 lecture)

Q: How to solve $(*)_1$?

Answer in Optimization Module