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
  - l2 regularization
  - l1 regularization
- sparsity

Recommend reading:

-

warm up:

1. $w \in \mathbb{R}^d$, $\|w\|_2^2 =$ _____, $\|w\|_1 =$ _____

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

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

   $\nabla R(w) =$ _____

3. X^TX is _____ and therefore real-diagonalizable

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

   X^TX is invertible iff _____

   $(X^TX)^{-1} =$ _____ if X^TX 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 $R(f) := \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i)$ is small

($f(x_i)$: prediction, y_i: data)

(loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$

e.g. $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, $\ell(\hat{y}, y) = |\hat{y} - y|$)

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

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

issues with (*) no restriction on f $\Rightarrow$ { overfitting; impractical to solve

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

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

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

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

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

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 centerized : $\sum_{i=1}^{n} x_i = 0$.

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

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

$$\min_{w \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \qquad (*_1)$$

Q: Is this method good enough ? No

explanation from [linear algebra / statistics] p.o.v

2.2. Linear algebra p.o.v

Denote $X := \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$, $\underline{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Then $R(w) = \frac{1}{2n} \| Xw - \underline{y} \|_2^2$

and $(*_1) \iff \min_{w \in \mathbb{R}^d} \frac{1}{2n} \| Xw - \underline{y} \|_2^2 \qquad (*_2)$

$\nabla R(w) = \frac{1}{n} X^T (Xw - \underline{y})$, $\nabla^2 R(w) = X^T X \succeq 0$

Let $\nabla R(w) = 0$

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

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

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

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

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

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

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, but can "highly" correlated → X^TX is close to singular

⇕

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

⇓

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

⇓

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

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

Statement:

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

w_j are deterministic but unobservable

X_{ij} are deterministic and observable

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

($\leftrightarrow$ 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. $\hat{w} = Ay$, $\mathbb{E}\hat{w} = w$.

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

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

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

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

↑ inherited from true model, can not avoid

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

appears b/c we choose unbiased estimator (choose to stick to data)

"overfitting"

Proof of theorem (optional)

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

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

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

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

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

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

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

Then $\mathbb{E}\hat{w} = (C+D) X w + (C+D)\mathbb{E}\varepsilon$

$= (I + DX) w$

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

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

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

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

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

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

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

i.e. $\mathrm{Var}(\hat{w}) - \mathrm{Var}(\hat{w}_{OLS}) \geq 0$. □

3. Ridge Regression (LR with l_2 regularization)

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

Lesson from linear regression:

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

Formulation of Ridge Regression:

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

↑ regularization

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

Q: Why the regularization helps?

① solution to $(*_r)$: $\hat{w} = (X^TX + \lambda I_d)^{-1} X^T \underline{y}$ (HW)

λ can improve the smallest eigenvalue of X^TX.

② stability of prediction depends on the magnitude of w

larger magnitude → worse stability

so we penalize the magnitude of w when fitting the data.

③ data $\sim \| Xw - \underline{y} \|^2$, knowledge $\sim \| w \|^2$

regularization includes our knowledge / preference (bias) and can help avoid overfitting.

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.

$\leftrightarrow$ only a small portion of features contribute.

$\leftrightarrow$ want w to be sparse

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

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

↑ 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 $(*_\ell)$?

Answer in Optimization Module