CSE 6740 A/ISyE 6740: Computational Data Analysis: Introductory lecture

Nisha Chandramoorthy

August 28, 2023

Quick notes

- Review session tonight 7:30-9:30pm on Zoom. See Piazza/Canvas for details.
- Homework 1 will be out soon. Due on 9/13/2023.

Last time

- Empirical risk minimization, finite hypothesis classes

Last time

- Empirical risk minimization, finite hypothesis classes
- Overfitting, inductive bias, Intro to PAC learning

Last time

- Empirical risk minimization, finite hypothesis classes
- Overfitting, inductive bias, Intro to PAC learning
- Let ERM rule

$$h_S := \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_S(h),$$

where empirical error,

$$\hat{R}_S(h) := \frac{1}{m} \sum_{z \in S} \ell(z, h).$$

Let the realizability assumption be satisfied $\implies$ ERM rule h_S has zero empirical error. Then, with probability at least $1 - \delta$, the generalization error,

$$R(h_S) := E_{z \underset{\sim}{\cancel{\in}} \mathcal{D}} \ell(z, h_S) \leqslant \frac{1}{m} \log \frac{|\mathcal{H}|}{\delta}.$$

Linear models

$$\mathcal{H} = \{h(\cdot, w, b) : h(x, w, b) = w \cdot \Phi(x) + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \quad (1)$$

Linear models

$$\mathcal{H} = \{h(\cdot, w, b) : h(x, w, b) = w \cdot \Phi(x) + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \quad (1)$$

- $\Phi : \mathbb{X} \to \mathbb{R}^d$ is a set of *features*.

Linear models

$$\mathcal{H} = \{h(\cdot, w, b) : h(x, w, b) = w \cdot \Phi(x) + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \quad (1)$$

- $\Phi : \mathbb{X} \to \mathbb{R}^d$ is a set of *features*.
- Linear regression seeks ERM solution for square loss

Linear models

$$\mathcal{H} = \{h(\cdot, w, b) : h(x, w, b) = w \cdot \Phi(x) + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \quad (1)$$

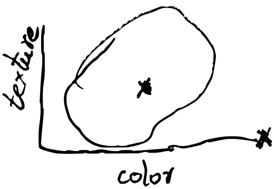
- $\Phi : \mathbb{X} \to \mathbb{R}^d$ is a set of *features*.
- Linear regression seeks ERM solution for square loss
- $$\longrightarrow \operatorname{argmin}_{w,b} \frac{1}{m} \sum_{i=1}^{m} (w^\top \Phi(x_i) + b - y_i)^2$$

$\mathrm{Supp}(\mathcal{D}) \subseteq \mathbb{X}$

$x \in \mathbb{X} \equiv \mathbb{R}^d$

$\mathcal{D}$

S



$$\operatorname*{argmin}_{\substack{w,\ b \\ \in \mathbb{R}^d\ \ \in \mathbb{R}}} \frac{1}{m} \sum_{i=1}^{m} \underbrace{\ell(z_i, h(z_i, w, b))}_{(x_i, y_i) \in S}$$

$$\ell(z, h(z, w, b)) = |h(z, w, b) - y|^2$$

Linear models

$$\mathcal{H} = \{h(\cdot, w, b) : h(x, w, b) = w \cdot \Phi(x) + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \quad (1)$$

- $\Phi : \mathbb{X} \to \mathbb{R}^d$ is a set of *features*.
- Linear regression seeks ERM solution for square loss
- $$\boxed{\operatorname{argmin}_{w,b} \frac{1}{m} (\underline{w^\top \Phi(x_i) + b - y_i})^2}$$ convex
- Equivalently, where X is $m \times (d+1)$ matrix with rows $X_i = (\Phi(x_i)^\top, 1)$, $W = [w_1, \cdots, w_d, b]^\top$, $Y = [y_1, \cdots, y_m]^\top$,

$f(x) = x^2$

$f \circ h$ is convex

Linear models ✓

$w \cdot \Phi(x)$

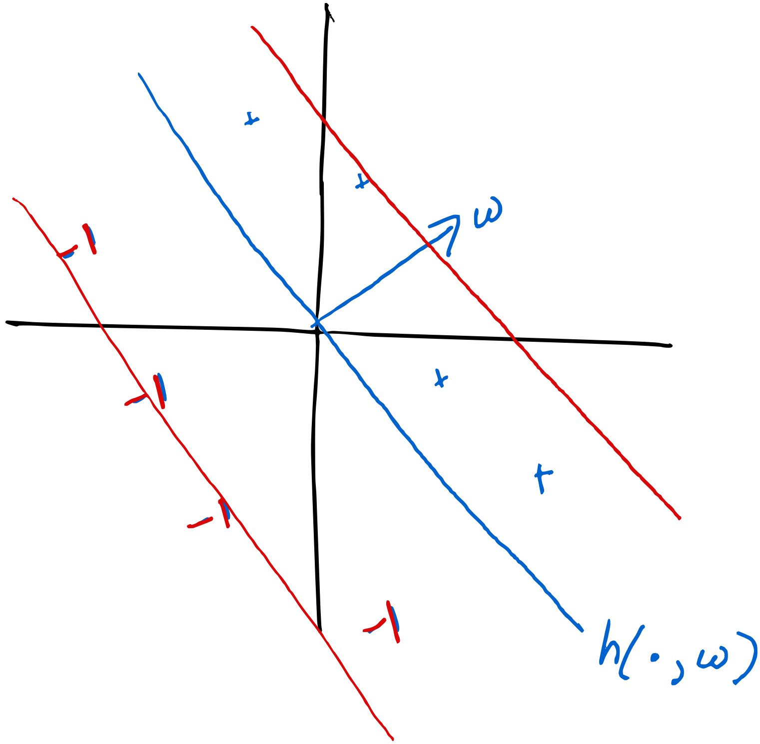
$$\mathcal{H} = \{h(\cdot, w, b) : h(x, w, b) = w \cdot \Phi(x) + b, w \in \mathbb{R}^d, b \in \mathbb{R}\} \quad (1)$$

- $\Phi : \mathbb{X} \to \mathbb{R}^d$ is a set of *features*.
- Linear regression seeks ERM solution for square loss
- $$\rightarrow \mathrm{argmin}_{w,b} \frac{1}{m}(w^\top \Phi(x_i) + b - y_i)^2 \quad ✓$$
- Equivalently, where X is $m \times (d+1)$ matrix with rows $X_i = (\Phi(x_i)^\top, 1)$, $W = [w_1, \cdots, w_d, b]^\top$, $Y = [y_1, \cdots, y_m]^\top$,
$$\rightarrow \mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2$$
- Features may be defined by kernels

$$\underset{m \times (d+1)}{X} = \begin{bmatrix} \Phi(x_1)^T & 1 \\ \vdots & \\ \Phi(x_m)^T & 1 \end{bmatrix}$$

$$X W, \quad W = \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ b \end{bmatrix}$$

$$\|X w - Y\|^2$$



$x^T w$

$\mathcal{H}^{lin}_{d+1}$ $\mathcal{H}^{aff}_{d}$

Least squares solutions: the optimization way

- Convex, differentiable function of W – composition of convex, differentiable functions

Least squares solutions: the optimization way

- Convex, differentiable function of W – composition of convex, differentiable functions
- Global minimum is the extremal point where derivative vanishes

Least squares solutions: the optimization way

- Convex, differentiable function of W – composition of convex, differentiable functions
- Global minimum is the extremal point where derivative vanishes

Matrix calculus

- $\nabla \frac{1}{m} \underbrace{\|XW - Y\|^2} = \frac{2}{m} \underbrace{X^T(XW - Y)}$

Least squares solutions: the optimization way

- Convex, differentiable function of W – composition of convex, differentiable functions
- Global minimum is the extremal point where derivative vanishes
- $\nabla \frac{1}{m}\|XW - Y\|^2 = \frac{2}{m}X^T(XW - Y)$
- $X^T XW = X^T Y$. Can also get this by differentiating before writing in matrix form

$X^T(X\overset{W}{w} - Y) = 0$

w is the minimizer

$X^T X w^{OLS} = X^T Y$

$w^{OLS} = (X^T X)^{-1} X^T Y$

Ordinary least squares

Least squares solutions: the optimization way

- Convex, differentiable function of W – composition of convex, differentiable functions
- Global minimum is the extremal point where derivative vanishes
- $\nabla \frac{1}{m} \|XW - Y\|^2 = \frac{2}{m} X^T (XW - Y)$
- $X^T X W = X^T Y$. Can also get this by differentiating before writing in matrix form
- When is $X^T X = \sum_{i=1}^{m} \Phi(x_i)\Phi(x_i)^T$ invertible? When the training features span $\mathbb{R}^d$.

$$X = \begin{bmatrix} \Phi^T(x_1) \\ \vdots \\ \Phi^T(x_m) \end{bmatrix}$$

$\Phi(x) = x$
$\in \mathbb{R}^d$

$x_i \sim D$

$$X X^T = \sum_{i=1}^{m} \underbrace{x_i x_i^T}$$

Least squares solutions: the linear algebraic way

- Case 1: $d + 1 = m$, X is invertible. $W = X^{-1}Y$ ✓

$XW = Y$

$m = d$

Least squares solutions: the linear algebraic way

- Case 1: $d + 1 = m$, X is invertible. $W = X^{-1}Y$
- Case 2: $d + 1 > m$, underdetermined/overparameterized. If X has full row rank, then, min norm solution

$$W = X^{\top}(XX^{\top})^{-1}Y$$

Least squares solutions: the linear algebraic way

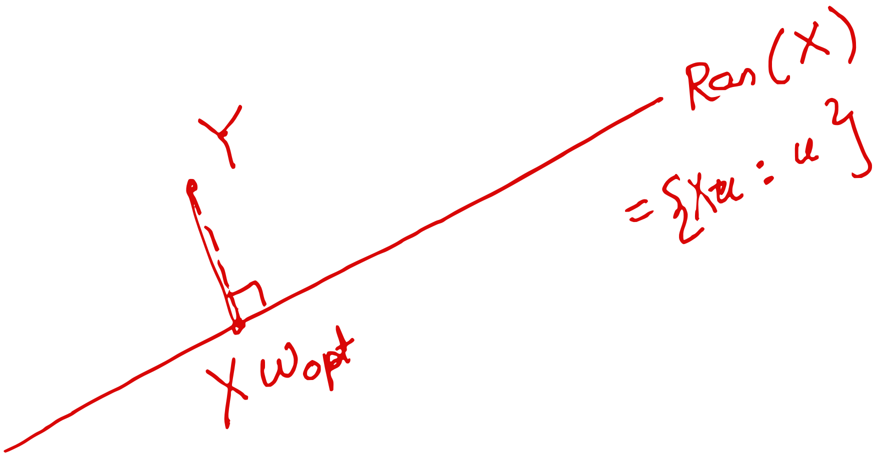
- Case 1: $d + 1 = m$, X is invertible. $W = X^{-1}Y$
- Case 2: $d + 1 > m$, underdetermined/overparameterized. If X has full row rank, then, min norm solution

$$W = X^\top (XX^\top)^{-1} Y$$

- Case 3: $m > d + 1$, overdetermined. If X has full col rank, then, many solutions. Min norm solution

$$W = (X^\top X)^{-1} X^\top Y$$

$$\underset{W}{\operatorname{argmin}} \|XW - Y\|^2$$



$$\langle Y - Xw_{opt}, Xu \rangle = 0 \quad \forall u$$

$$\langle X^T(Y - Xw_{opt}), u \rangle = 0 \quad \forall u$$

$$X^T Y = X^T X w_{opt}$$

$$w_{opt} = (X^T X)^{-1} X^T Y$$

Recall

When is $X^T X$ invertible?

$$u^T X^T X u = \|Xu\|^2 \geq 0$$

SPSD

$m < d$ $\qquad$ $\operatorname{rank}(X^T X)$

$$X^T X = \sum_{i=1}^{m} x_i x_i^T$$

if x_i are linearly independent

Least squares solutions: the linear algebraic way

- Case 1: $d+1 = m$, X is invertible. $W = X^{-1}Y$
- Case 2: $d+1 > m$, underdetermined/overparameterized. If X has full row rank, then, min norm solution

$$W = X^\top (XX^\top)^{-1} Y$$

- Case 3: $m > d+1$, overdetermined. If X has full col rank, then, many solutions. Min norm solution

$$W = (\underline{X^\top X})^{-1} X^\top Y$$

- Can solve normal equations above directly, or use iterative methods for linear systems. Cost $\mathcal{O}(\underline{d}^3)$

$$\underset{m \times d}{X}$$

$$\underline{X^T X W} = \underline{X^T Y}$$

$$W = (X^T X)^{\dagger} X^T Y$$

d: input dim/
feature dim

Gauss Markov theorem

- Take noisy $y_i = x_i^\top W + \epsilon_i$, with $E\epsilon_i = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$; x_i is non-random.

Gauss Markov theorem

$w^{OLS} = (X^T X)^{-1} X^T Y$

- Take noisy $y_i = x_i^\top W + \epsilon_i$, with $E\epsilon_i = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$; x_i is non-random.
- Statement: the OLS estimator is the best linear unbiased estimator (blue). It has the lowest variance.

Gauss Markov theorem

$$W^{OLS} = (X^T X)^{-1} X^T Y$$

$$= \mathbb{E}[\varepsilon] = 0$$

- Take noisy $y_i = x_i^\top W + \epsilon_i$, with $E\epsilon_i = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$; x_i is non-random.
- Statement: the OLS estimator is the best linear unbiased estimator (blue). It has the lowest variance.
- Proof: consider another linear estimator $W' = CY$, for some $(d+1) \times n$ matrix $C = (X^\top X)^{-1} X^\top + D$.

OLS is an unbiased estimator

Let W be true value that makes

$$y_i = x_i^T W + \varepsilon_i$$

$$\mathbb{E}[W^{OLS}] = \mathbb{E}[(X^T X)^{-1} X^T Y]$$

$$= \mathbb{E}[(X^T X)^{-1} X^T (XW + \varepsilon)]$$

$$= \underset{\text{true value}}{W} + \underset{0}{\mathbb{E}[\varepsilon]}$$

$$= W$$

Bias of $W^{OLS} = \mathbb{E}[W^{OLS}] - W = 0$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Gauss Markov theorem

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix}$$

$$\mathbb{E}[Y_i] = 0$$

- Take noisy $y_i = x_i^\top W + \epsilon_i$, with $E\epsilon_i = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$; x_i is non-random.
  true
- Statement: the OLS estimator is the best linear unbiased estimator (blue). It has the lowest variance.
- Proof: consider another linear estimator $W' = CY$, for some $(d+1) \times n$ matrix $C = \underline{(X^\top X)^{-1} X^\top} + \underline{D}$.
- For W' to be unbiased, show $DX = 0$. Then show, $\mathrm{Var}(W') = \mathrm{Var}(W) + \boxed{\sigma^2 DD^\top}$

CY

$$\mathbb{E}[W'] = \mathbb{E}[CY]$$

$$\mathrm{Var}(W_i') \geq \mathrm{Var}(W_i^{OLS}) \quad 1 \leq i \leq m$$

$$\mathrm{Var}(W) = \mathbb{E}\left(\left(W - \mathbb{E}[W]\right)^2\right)$$

$$\mathrm{Var}(W') = \mathrm{Var}(W^{OLS}) + \sigma^2 DD^\top$$

Gauss Markov theorem

- Take noisy $y_i = x_i^\top W + \epsilon_i$, with $E\epsilon_i = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$; x_i is non-random.
- Statement: the OLS estimator is the best linear unbiased estimator (blue). It has the lowest variance.
- Proof: consider another linear estimator $W' = CY$, for some $(d+1) \times n$ matrix $C = (X^\top X)^{-1} X^\top + D$.
- For W' to be unbiased, show $DX = 0$. Then show, $\mathrm{Var}(W') = \mathrm{Var}(W) + \sigma^2 DD^\top$.
- Since $DD^\top$ is positive semi-definite, qed.

Ridge regression

- Motivation: unbiased estimation does not mean least mean-squared error

Ridge regression

Regularization

- Motivation: unbiased estimation does not mean least mean-squared error
- Let true $h_W \in \mathcal{H}$. Mean-squared error of statistical estimator of W, $\hat{W}$, is its variance + bias-squared

$$E[(\hat{W}_i - W_i)^2) = \mathrm{Var}(\hat{W}_i) + (E[\hat{W}_i] - W_i)^2.$$

- $$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$

Ex:

$+ E[\hat{W}_i] - E[\hat{W}_i]$

$$E((\hat{W}_i - W_i)^2) =$$

$$\underbrace{\mathrm{Var}(\hat{W}_i)}_{\downarrow} + (\underbrace{E[\hat{W}_i] - W_i}_{\uparrow})^2$$

Ridge regression

- Motivation: unbiased estimation does not mean least mean-squared error
- Let true $h_W \in \mathcal{H}$. Mean-squared error of statistical estimator of W, $\hat{W}$, is its variance + bias-squared

$$E[(\hat{W}_i - W_i)^2) = \mathrm{Var}(\hat{W}_i) + (E[\hat{W}_i] - W_i)^2.$$

- $$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$

- penalizes l^2 norm of W. Still convex problem.

Ridge regression

- Motivation: unbiased estimation does not mean least mean-squared error
- Let true $h_W \in \mathcal{H}$. Mean-squared error of statistical estimator of W, $\hat{W}$, is its variance + bias-squared

$$E[(\hat{W}_i - W_i)^2) = \mathrm{Var}(\hat{W}_i) + (E[\hat{W}_i] - W_i)^2.$$

- Ridge regression

$$\underline{\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2}.$$

- penalizes l^2 norm of W. Still convex problem.
- to derive OLS, also can take derivative and set it to zero. Similarly here.

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$$\|W\|^2 = \sum_{i=1}^{d} |w_i|^2$$

Ridge regression

- Motivation: unbiased estimation does not mean least mean-squared error
- Let true $h_W \in \mathcal{H}$. Mean-squared error of statistical estimator of W, $\hat{W}$, is its variance + bias-squared

$$E[(\hat{W}_i - W_i)^2) = \mathrm{Var}(\hat{W}_i) + (E[\hat{W}_i] - W_i)^2.$$

- $$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$
- penalizes l^2 norm of W. Still convex problem.
- to derive OLS, also can take derivative and set it to zero. Similarly here.
- Equivalent formulation: $\min_w \sum_{i=1}^m (w^\top \Phi(x_i) - y_i)^2$ subject to $\|w\|^2 \leqslant \Lambda^2$

Ridge regression optimization

$d \gg m$

Primal:

$$\min_{w \in \mathbb{R}^d} \left(\|Xw - Y\|^2 + \lambda \|w\|^2 \right)$$

Lagrangian

$\downarrow$

Dual:

$$\max_{\alpha \in \mathbb{R}^m} \underbrace{-\alpha^T (XX^T + \lambda I)\alpha + 2\alpha^T Y}_{\text{Concave}}$$

XX^T $\qquad\qquad$ $X^T X$

Closed form solution

$$W^{Ridge} = (\underbrace{X^T X + \lambda I}_{\text{always invertible}})^{-1} X^T Y$$

Shrinkage

$$W^{Ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

$$Y^{Ridge}_{pred} = X W^{Ridge}$$

$$= X (X^T X + \lambda I)^{-1} X^T Y$$

$$\underset{m \times d}{X} = U \Sigma V^T \qquad \underset{u_i \in \mathbb{R}^m}{U} = [u_1 | u_2 | \cdots | u_d]$$

Ex: $Y^{Ridge}_{pred} = U\Sigma V^T (V^T \Sigma V + \lambda I)^{-1} V \Sigma U^T Y$

$$= \sum_{i=1}^{d} u_i \left(\frac{d_i^2}{\lambda + d_i^2} \right) \underline{u_i^T Y}$$

$$\Sigma = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_d \end{bmatrix}$$

Ridge regression solution and interpretation

- Revisit convex optimization. Derive solution using KKT conditions.

Ridge regression solution and interpretation

- Revisit convex optimization. Derive solution using KKT conditions.
- Now simply use convexity and differentiability to obtain global minimum:

$$W = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Ridge regression solution and interpretation

- Revisit convex optimization. Derive solution using KKT conditions.
- Now simply use convexity and differentiability to obtain global minimum:
$$W = (X^\top X + \lambda I)^{-1} X^\top Y.$$
- LASSO: with l^1 regularization.

Ridge regression solution and interpretation

- Revisit convex optimization. Derive solution using KKT conditions.
- Now simply use convexity and differentiability to obtain global minimum:
$$W = (X^\top X + \lambda I)^{-1} X^\top Y.$$

- LASSO: with l^1 regularization.
- Generalization bounds for bounded regression problems.

Ridge regression solution and interpretation

- Revisit convex optimization. Derive solution using KKT conditions.
- Now simply use convexity and differentiability to obtain global minimum:
$$W = (X^\top X + \lambda I)^{-1} X^\top Y.$$
- LASSO: with l^1 regularization.
- Generalization bounds for bounded regression problems.
- Shrinkage by l^2 regularization.