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

Nisha Chandramoorthy

August 29, 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

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

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,

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

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
- $$\mathrm{argmin}_{w,b} \frac{1}{m} (w^\top \Phi(x_i) + b - y_i)^2$$

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
- $$\operatorname{argmin}_{w,b} \frac{1}{m} (w^\top \Phi(x_i) + b - y_i)^2$$
- Equivalently, where X is $m \times (d+1)$ matrix with rows $x_i = (\Phi(x_i)^\top, 1)$ (abuse of notation), $w = [w_1, \cdots, w_d, b]^\top$, $Y = [y_1, \cdots, y_m]^\top$,

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
- $$\mathrm{argmin}_{w,b} \frac{1}{m}(w^\top \Phi(x_i) + b - y_i)^2$$
- Equivalently, where X is $m \times (d+1)$ matrix with rows $x_i = (\Phi(x_i)^\top, 1)$ (abuse of notation), $w = [w_1, \cdots, w_d, b]^\top$, $Y = [y_1, \cdots, y_m]^\top$,
$$\mathrm{argmin}_w \frac{1}{m}\|Xw - Y\|^2$$
- Features may be defined by kernels

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

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
- $\nabla \frac{1}{m} \|Xw - Y\|^2 = \frac{2}{m} X^T (Xw - Y)$
- $X^T X w = X^T Y$.

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
- $\nabla \frac{1}{m} \|Xw - Y\|^2 = \frac{2}{m} X^T (Xw - Y)$
- $X^T X w = X^T Y$.
- When is $X^T X = \sum_{i=1}^{m} \Phi(x_i) \Phi(x_i)^T$ invertible? When the training features span $\mathbb{R}^d$.

Least squares solutions: the linear algebraic way

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

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

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

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

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

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

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

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

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

- 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

- 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 $C = (X^\top X)^{-1} X^\top + D$.

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, $\mathrm{Var}(w') = \mathrm{Var}(w) + \sigma^2 DD^\top$.

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.

Bias-variance decomposition

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

Bias-variance decomposition

- 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

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

- $$\mathrm{argmin}_w \frac{1}{m}\|Xw - Y\|^2 + \lambda\|w\|^2.$$

Bias-variance decomposition

- 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.

Bias-variance decomposition

- 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.

Bias-variance decomposition

- 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 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.