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

Nisha Chandramoorthy

August 24, 2023

Last time

- Goal in this class: understand the foundations ("why"s and "how"s) of ML

Last time

- Goal in this class: understand the foundations ("why"s and "how"s) of ML
- Supervised, unsupervised, self-supervised, semi-supervised overview

Last time

- Goal in this class: understand the foundations ("why"s and "how"s) of ML
- Supervised, unsupervised, self-supervised, semi-supervised overview
- Empirical risk minimization, finite hypothesis classes

Last time

- Goal in this class: understand the foundations (“why”s and “how”s) of ML
- Supervised, unsupervised, self-supervised, semi-supervised overview
- Empirical risk minimization, finite hypothesis classes
- Overfitting, inductive bias, Intro to PAC learning

Supervised learning framework

- Distribution $\mathcal{D}$ over the joint distribution of random variables $Z = (X, Y)$, where X is an input and Y is a label/output.

Supervised learning framework

- Distribution $\mathcal{D}$ over the joint distribution of random variables $Z = (X, Y)$, where X is an input and Y is a label/output.
- Labeled training data, $S = \{z_i = (x_i, y_i)\}$, $1 \leqslant i \leqslant m$. Generally iid from $\mathcal{D}^m$.

Supervised learning framework

- Distribution $\mathcal{D}$ over the joint distribution of random variables $Z = (X, Y)$, where X is an input and Y is a label/output.
- Labeled training data, $S = \{z_i = (x_i, y_i)\}, 1 \leqslant i \leqslant m$. Generally iid from $\mathcal{D}^m$.
- Learner's output: predicted function or hypothesis, a transformation h from X to Y.

Supervised learning framework

- Distribution $\mathcal{D}$ over the joint distribution of random variables $Z = (X, Y)$, where X is an input and Y is a label/output.
- Labeled training data, $S = \{z_i = (x_i, y_i)\}, 1 \leqslant i \leqslant m$. Generally iid from $\mathcal{D}^m$.
- Learner's output: predicted function or hypothesis, a transformation h from X to Y.
- Loss function, measure of risk: $\ell(z, h) \in \mathbb{R}$. e.g., $\ell(z, h) = \mathbb{1}_{h(x) \neq y}$. (classification)

Supervised learning framework

- Distribution $\mathcal{D}$ over the joint distribution of random variables $Z = (X, Y)$, where X is an input and Y is a label/output.
- Labeled training data, $S = \{z_i = (x_i, y_i)\}, 1 \leqslant i \leqslant m$. Generally iid from $\mathcal{D}^m$.
- Learner's output: predicted function or hypothesis, a transformation h from X to Y.
- Loss function, measure of risk: $\ell(z, h) \in \mathbb{R}$. e.g., $\ell(z, h) = \mathbb{1}_{h(x) \neq y}$. (classification)
- Generalization error or risk:

$$R(h) = E_{z \sim \mathcal{D}} \ell(z, h)$$

Empirical Risk Minimization

- Take classifier h, $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$.

Empirical Risk Minimization

- Take classifier h, $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$.
- When you have finite amount of data, empirical risk or training loss
$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$

Empirical Risk Minimization

- Take classifier h, $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$.
- When you have finite amount of data, empirical risk or training loss
$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$
- ERM: find h that minimizes $\hat{R}_S(h)$.

Empirical Risk Minimization

- Take classifier h, $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$.
- When you have finite amount of data, empirical risk or training loss
$$\hat{R}_S(h) = \frac{1}{m} \sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$
- ERM: find h that minimizes $\hat{R}_S(h)$.
- Theory of supervised learning suggests that ERM leads to small generalization error with high probability

Empirical Risk Minimization

- Take classifier h, $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$.
- When you have finite amount of data, empirical risk or training loss

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

- ERM: find h that minimizes $\hat{R}_S(h)$.
- Theory of supervised learning suggests that ERM leads to small generalization error with high probability
- Now we will prove this result formally.

Empirical Risk Minimization

- Take classifier h, $R(h) = \mathcal{D}(\{z : h(x) \neq y\})$.
- When you have finite amount of data, empirical risk or training loss
$$\hat{R}_S(h) = \frac{1}{m}\sum_{z \in S} \ell(z) = \frac{|\{z \in S : h(x) \neq y\}|}{m}.$$
- ERM: find h that minimizes $\hat{R}_S(h)$.
- Theory of supervised learning suggests that ERM leads to small generalization error with high probability
- Now we will prove this result formally.
- After that: Linear models.

With probability $\geq 1-\delta$ over samples S, the generalization risk $R(h_S)$ of an ERM rule h_S

Probably

$$\longrightarrow \boxed{R(h_S) < \underbrace{\frac{1}{m} \log\left(\frac{|\mathcal{H}|}{\delta}\right)}_{\text{approximately correct}}}$$

Set $\delta := e^{-m\varepsilon}|\mathcal{H}|$

Proof:

Realizability assumption: With probability 1, there is some $h \in \mathcal{H}$ s.t.

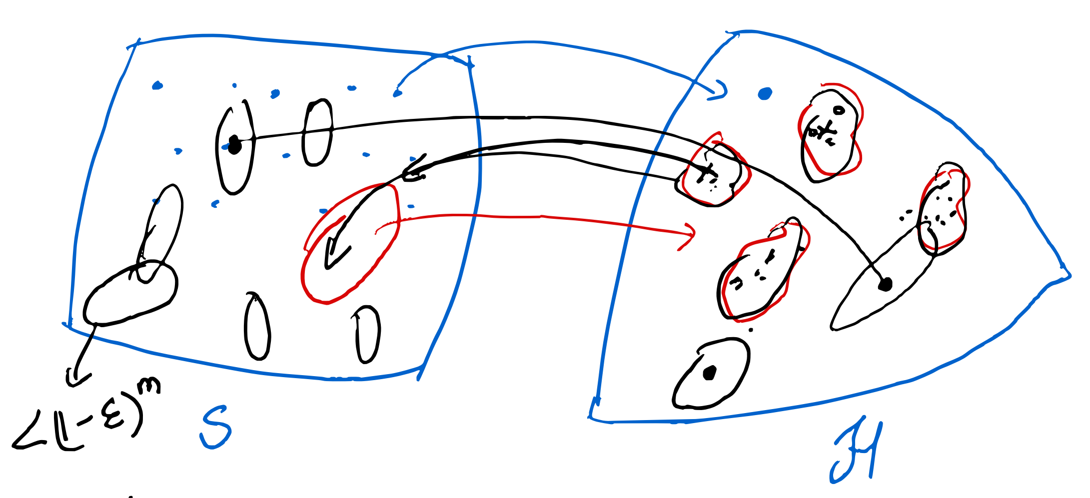
$$h(x) = y$$

$$\boxed{\text{Recall } R(h) = \mathop{\mathbb{E}}_{z \sim D} \ell(z, h)}$$

$$\boxed{h_S \in \underset{h\in\mathcal{H}}{\arg\min}\ \hat{R}_S(h) = \underset{h\in\mathcal{H}}{\arg\min} \frac{1}{m}\sum_{z\in S} \ell(z,h)}$$

For any S, ERM rule h_S is s.t.

$$\hat{R}_S(h_S) = 0$$



$$\mathcal{H}_B \subseteq \mathcal{H}$$

$$\boxed{\mathcal{H}_B := \{h \in \mathcal{H} : R(h) > \varepsilon\}}$$

$$R(h_S) > \varepsilon$$

$$D^m(\{S : h_S \in \mathcal{H}_B\})$$

$$\leq D^m(\{S : (\hat{R}_S(h) = 0) \text{ and } (h \in \mathcal{H}_B)\})$$

Prob review

$$D(A) = \Pr(X \in A) \qquad X \sim D$$

$$= \mathop{\mathbb{E}}_{X\sim D} \mathbb{1}_A(X)$$

$$\mathop{\mathbb{E}}_{X\sim D} f(X) = \sum_{x} f(x) \Pr(X = x) \quad \text{(Discrete)}$$

$$= \int f(x)\, dD(x) \quad \text{(cont)}$$

$$D^m \qquad S = (X_1, \dots, X_m)$$

$$D^m(\{S : (\hat{R}_S(h) = 0) \text{ and } (h \in \mathcal{H}_B)\})$$

$$h(x_i) = y_i \quad \forall\, i \in \{1, 2, \dots, m\}$$

$$D^m(\{S : (\hat{R}_S(h) = 0) \text{ and } (h \in \mathcal{H}_B)\})$$

$$\leq (1-\varepsilon)^m |\mathcal{H}_B| \quad \text{(Union bound)}$$

$$\leq e^{-m\varepsilon} |\mathcal{H}_B|$$

$$\leq e^{-m\varepsilon} |\mathcal{H}| =: \delta$$

Prob review

$$P(A \cup B) \leq P(A) + P(B)$$

$$(1-\varepsilon)^m$$

$$P(A) < \delta$$

$$P(A^c) > 1-\delta$$

$|A|$: size of set A.

$\mathcal{D}^m$: input distribution

$\mathcal{H}$: hypothesis class

"Complexity"

Linear models

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

$A(x)$ if $A(cx) = cA(x)$

$A(x_1 + x_2) = A(x_1) + A(x_2)$

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

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_d & b \end{bmatrix} \begin{bmatrix} \Phi(x)_1 \\ \vdots \\ \Phi(x)_d \\ 1 \end{bmatrix}$$

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

$$[w, b] = \underset{w,b}{\operatorname{argmin}} \frac{1}{m} \sum_{z \in S} \ell(z, h(w, b, \cdot))$$

$$= \underset{w,b}{\operatorname{argmin}} \frac{1}{m} \sum_{\substack{z \in S \\ \downarrow \\ (x,y)}} (w^T \Phi(x) + b - y)^2$$

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

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
- $$\operatorname{argmin}_{w,b} \frac{1}{m} (w^\top \Phi(x_i) + b - y_i)^2$$
- Equivalently, where X is $m \times (d + 1)$ matrix with rows $X_i = (\Phi(x_i)^\top, 1)$, $W = [w_1, \cdots, w_d, \cancel{1}\, b]^\top$, $Y = [y_1, \cdots, y_m]^\top$,
  $$\operatorname{argmin}_W \frac{1}{m} \|X\underline{\underline{W}} - Y\|^2$$
- 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
- $\nabla \frac{1}{m}\|XW - Y\|^2 = \frac{2}{m}X^T(XW - Y)$
- $X^T XW = X^T Y$. Can also get this by differentiating before writing in matrix form

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

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
- $\nabla \frac{1}{m}\|XW - Y\|^2 = \frac{2}{m}X^T(XW - Y)$
- $X^T XW = X^T Y$. 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 $\mathbb{R}^d$.

Least squares solutions: the linear algebraic way

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

$XW = Y$

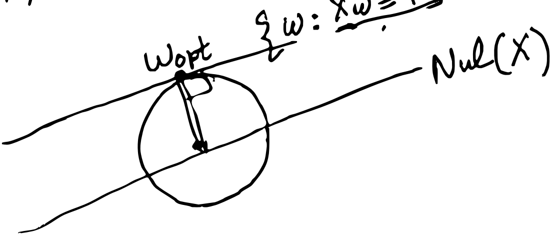
Least squares solutions: the linear algebraic way

- 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

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

$$\frac{1}{m}\|XW - Y\|^2$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$



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

$$W \in \mathbb{R}^{(d+1)}$$

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 $E_i = 0$ 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 $E_i = 0$ 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 $E_i = 0$ 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 $E_i = 0$ 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 $E_i = 0$ 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

- 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

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

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

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

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

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

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

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 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.

Generalization of regression

Hoeffding's inequality: $S_n = X_1 + X_2 + \cdots + X_n$

$X_i \perp\!\!\!\perp X_j \quad 0 \le X_i \le L$

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \leq e^{-\frac{2t^2}{nL^2}}$$

Thm: $\sup_{z,h} \ell(z,h) = L$. Let $\mathcal{H}$ be finite.

Then, for every $\delta > 0$, with probability at least $1-\delta$, $\forall\, h \in \mathcal{H}$,

$$R(h) \leq R_S(h) + L\sqrt{\frac{\log|\mathcal{H}| + \log 1/\delta}{2m}}$$