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

Nisha Chandramoorthy

August 23, 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.

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

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

- Linear regression seeks ERM solution for square loss

Linear models

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

- 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

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

- 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 $n \times (d+1)$ matrix with rows $X_i = (\Phi(x_i)^\top, 1)$, $W = [w_1, \cdots, w_d, 1]^\top$, $Y = [y_1, \cdots, y_n]^\top$,

Linear models

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

- 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 $n \times (d+1)$ matrix with rows $X_i = (\Phi(x_i)^\top, 1)$, $W = [w_1, \cdots, w_d, 1]^\top$, $Y = [y_1, \cdots, y_n]^\top$,
$$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2$$
- Features may be defined by kernels

Least squares solutions

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

Least squares solutions

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

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

Least squares solutions

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

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

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

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

Gauss Markov theorem

- Take y_i to be random with variance σ^2, x_i to be non-random.

Gauss Markov theorem

- Take y_i to be random with variance σ^2, x_i to be non-random.
- Statement: the OLS estimator is the best linear unbiased estimator (blue). It has the lowest variance.

Gauss Markov theorem

- Take y_i to be random with variance σ^2, x_i to be 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 $C = (X^\top X)^{-1} X^\top Y + D$.

Gauss Markov theorem

- Take y_i to be random with variance σ^2, x_i to be 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 $C = (X^\top X)^{-1} X^\top Y + D$.
- For W' to be unbiased, show $DX = 0$. Then show, $\mathrm{Var}(W') = \mathrm{Var}(W) + \sigma^2 DD'^\top$.

Gauss Markov theorem

- Take y_i to be random with variance σ^2, x_i to be 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 $C = (X^\top X)^{-1} X^\top Y + D$.
- 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
- $$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$

Ridge regression

- Motivation: unbiased estimation does not mean least mean-squared error
- $$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$
- penalizes l^2 norm of W. Still convex problem.

Ridge regression

- Motivation: unbiased estimation does not mean least mean-squared error
- $$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$
- 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
-
$$\mathrm{argmin}_W \frac{1}{m}\|XW - Y\|^2 + \lambda\|W\|^2.$$
- 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.