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

Nisha Chandramoorthy

September 5, 2023

Last time

- Recap of shrinkage by ridge regression

Last time

- Recap of shrinkage by ridge regression
- Geometric view of compression by LASSO

Last time

- Recap of shrinkage by ridge regression
- Geometric view of compression by LASSO
- Generalization of LASSO

Last time

- Recap of shrinkage by ridge regression
- Geometric view of compression by LASSO
- Generalization of LASSO
- ℓ^0 regularization and compressed sensing

Today

- Classification ERM

Today

- Classification ERM
- Logistic regression, Bayesian view

Today

- Classification ERM
- Logistic regression, Bayesian view
- Perceptron algorithm and convergence proof

Today

- Classification ERM
- Logistic regression, Bayesian view
- Perceptron algorithm and convergence proof
- Support vector regression or maximum margin classification

Today

- Classification ERM
- Logistic regression, Bayesian view
- Perceptron algorithm and convergence proof
- Support vector regression or maximum margin classification
- Convex optimization

Today

- Classification ERM
- Logistic regression, Bayesian view
- Perceptron algorithm and convergence proof
- Support vector regression or maximum margin classification
- Convex optimization
- Bias-complexity tradeoff

Linear predictors

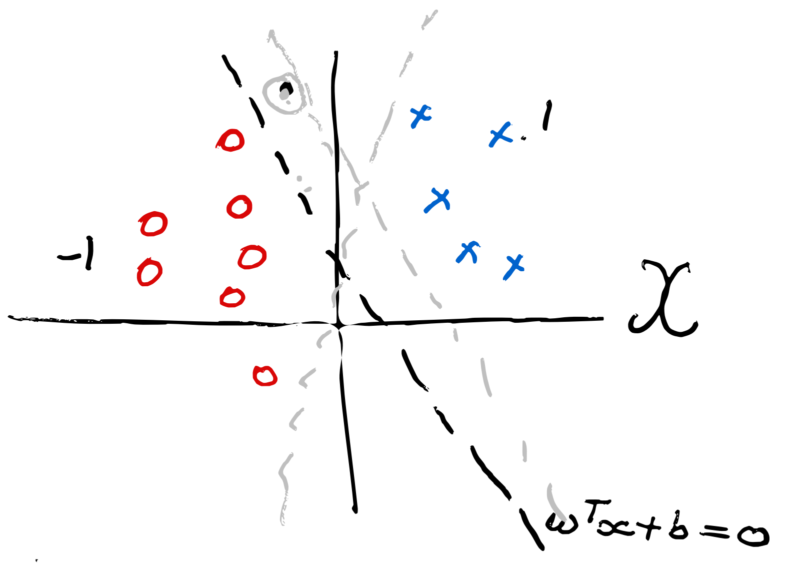
Classification ERM

$$\mathcal{H} = \{ h^{lin}(\cdot\,; w, b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$

$HS =$

$$\{ \underbrace{\operatorname{sgn} \circ h^{lin}}(\cdot, w, b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$



$$R(h) = \mathop{\mathbb{E}}_{z \sim D} \mathbb{1}_{\{(x,y) : h(x) \neq y\}}$$

<u>ERM</u> $\quad S : \{(x_i, y_i) : 1 \leq i \leq m\}$ iid from $\mathcal{D}$.

$$R_S(h) = \frac{1}{m} \sum_{z \in S} \mathbb{1}_{\{(x,y) \in S : \underline{h(x) \neq y}\}}$$

$$h(\cdot, w, b) \in HS$$

Realizability : $\exists\ h(\cdot, w, b) \in \underline{HS}$

s.t. $\quad h(x, w, b) = y$ for almost every $(x, y) \sim \underline{D}$.

$$h^{ERM} = \underset{\underline{h \in HS}}{\operatorname{argmin}} \frac{1}{m} \sum_{z \in S} \mathbb{1}_{\{(x,y) \in S : h(x) \neq y\}}$$

any h^{ERM} will be s.t.

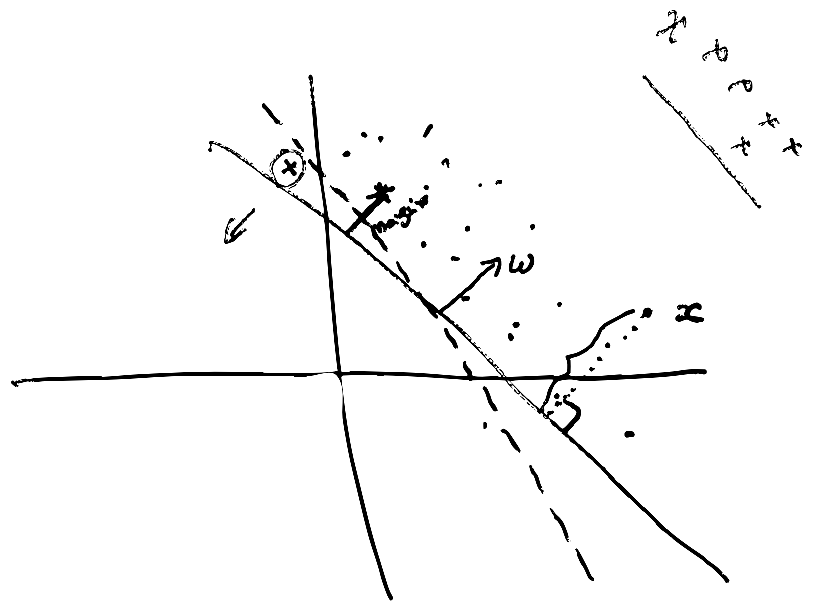
$$h^{ERM}(x_i) = y_i \quad \forall\ i \in [m]$$

$(1, 2, \ldots, m)$

<u>Margin</u>

Given a sample set S and a classifier $h \in HS$,

$$\operatorname{Margin}(h, S) = \min_{x_i \in S} d_h(x_i)$$



$$\max_{w, b} \operatorname{Margin}(h(\cdot, w, b), S)$$

subject to

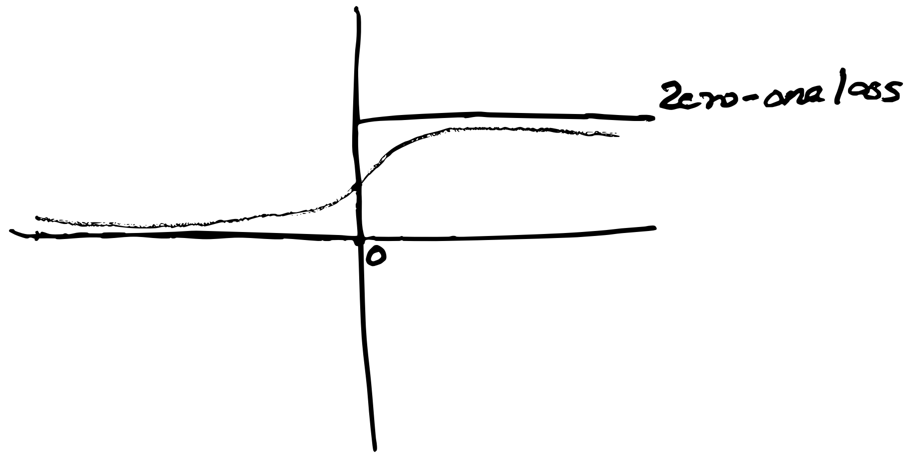
SVM $\qquad \to h(x_i, w, b) = y_i \quad \forall\ i \in [m]$

Logistic regression

$$p(x) = \frac{1}{1+e^{-x}}$$

ERM :

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \ \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-y_i \langle w, x_i \rangle}\right)$$



Likelihood function:

$$P((x,y) \mid w, b) = \frac{1}{1 + e^{-y(\langle w, x\rangle + b)}}$$

Bayes rule

$$P(\underbrace{w, b} \mid S) \ \propto \ P(S) \ \underline{P(S \mid w, b)}$$

MLE :

$$\underline{P(S/w,b)} = \prod_{i=1}^{m} \frac{1}{1+ e^{-y_i(\langle w, x_i\rangle + b)}}$$

$$\log P(S/w,b) = -\sum_{i=1}^{m} \log\left(1 + e^{-y_i(\langle w, x_i\rangle + b)}\right)$$

$$\max_{w,b} \ \underline{\log} \text{ likelihood } (w, b, S)$$

Convex optimization

Halfspaces

Linear programming :

$$\max_{w \in \mathbb{R}^d} \langle u, w \rangle$$

subject to $Aw \geq v$

$$y_i \langle w, x_i \rangle > 0 \quad \forall \ i \in [m]$$

$$y_i (\langle w, x_i \rangle + b) > 0$$

Recall realizability. $\exists (w^*, b^*)$

Define $s := \min_i y_i (\langle w^*, x_i \rangle + b^*)$

$$\frac{y_i}{s} (\langle w^*, x_i \rangle + b^*) \geq 1$$

$$y_i (\langle w, x_i \rangle + b) \geq 1$$

Perceptron algorithm

$$w^{(t+1)} = w^{(t)} + y_i x_i \text{ , where}$$

i is such that x_i is mis-classified by $h_{w^{(t)}}$

$$h(x_i, w, b) = y_i \quad \forall \; i \in [m]$$

$$w^{(0)} = 0 \in \mathbb{R}^{d+1}$$

$$w^{(t+1)} = w^{(t)} + y_i x_i$$

$$w^{(T)}$$

Convergence proof

Assumptions: $\to$ realizability

Given: $B = \min\{ \|w\| \cdot \forall i \in [m]$
$y_i \langle w . x_i \rangle \geq 1 \}$

$\to R = \max_i \|x_i\|$

Thm: The perceptron algorithm stops after at most $(RB)^2$ steps

Given:
$\|w^*\| = B$

$w^{(1)} = 0$

$w^{(t+1)} = w^{(t)} + y_i x_i$

$w^{(t+1)} - w^{(t)} = y_i x_i$

$y_i \langle w^{(t+1)}, x_i \rangle > y_i \langle w^t, x_i \rangle$

$\parallel$

$y_i \langle w^{(t)} + x_i y_i, x_i \rangle$

$y_i \langle w^{(t)}, x_i \rangle + \|x_i\|^2$

Proof:

$w^{(t+1)} = w^{(t)} + y_i x_i$

$\|w^{(t+1)}\|^2 = \|w^{(t)}\|^2 + \|x_i\|^2 +$
$2y_i \langle w^{(t)}, x_i \rangle$

Since $y_i \langle w^{(t)}, x_i \rangle \leq 0$

$\|w^{(t+1)}\|^2 \leq \|w^{(t)}\|^2 + \|x_i\|^2$
$\leq \|w^{(t)}\|^2 + R^2$

Telescoping,

$\|w^{(T+1)}\|^2 \leq \|w^{(T)}\|^2 + R^2$
$\leq \|w^{(T-1)}\|^2 + 2R^2$
$\vdots$
$(+) \leq TR^2 \quad (w^{(1)} = 0)$

$\langle w^{(t+1)}, w^* \rangle = \langle w^{(t)} + y_i x_i, w^* \rangle$
$= \langle w^{(t)}, w^* \rangle + y_i \langle x_i, w^* \rangle$

$\Rightarrow \langle w^{t+1} - w^{(t)}, w^* \rangle = y_i \langle x_i, w^* \rangle$

($\because y_i \langle x_i, w^* \rangle \geq 1$ since w^* is a correct classifier)

$\Rightarrow \langle w^{(T+1)}, w^* \rangle = \sum_{t=1}^{T} \langle w^{(t+1)} - w^{(t)}, w^* \rangle$
$(\because w^{(1)} = 0)$
$\geq T \quad (*)$

Using Cauchy-Schwarz,

$|\langle w^{(T+1)}, w^* \rangle| \leq \|w^{(T+1)}\| \|w^*\|$
$\leq \|w^{(T+1)}\| B$

$\overset{(*)}{\Rightarrow}$

$T \leq \|w^{(T+1)}\| B$

Applying $(+)$,

$T \leq \|w^{(T+1)}\| B$
$\leq \sqrt{T} RB$

$\Rightarrow T^2 \leq RB.$