L19: Feature Selection

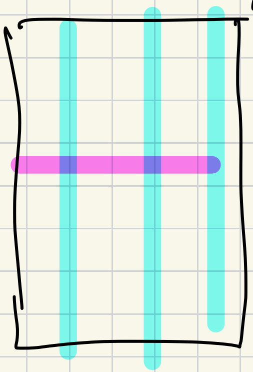
Nov 5, 2023

Data Mining

Jeff M. Phillips

Data $X = \{x_1, x_2 \dots x_n\} \subset \mathbb{R}^d$

$X \in \mathbb{R}^{n \times d}$

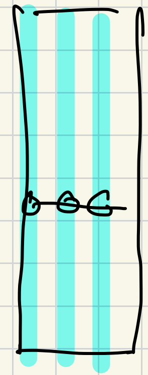


$x_{i:}$

$X_{:j}$

column = feature

Goal $\Longrightarrow$



$x'_{i:}$

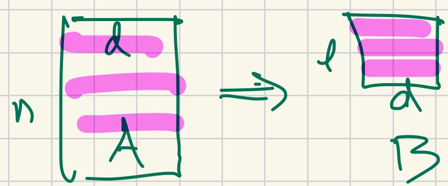
$X' \in \mathbb{R}^{n \times k}$

$k \ll d$

$X_{:j_1} X_{:j_2} X_{:j_3}$

Recall Matrix sketching

→ row sampling.



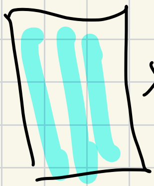
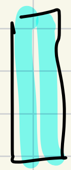
$\ell = \frac{d}{\epsilon^2}$

Transpose

→ column sampling

leverage score

Reduce

 $X \in \mathbb{R}^{n\times d}$ $\rightarrow$  $X' \in \mathbb{R}^{n\times p}$

but for what goal?

Input $(X, y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^{1\times n}$

point (x_i, y_i)

$\uparrow$ Label

goal $f(x_i) \approx y_i$

Regression

Linear Least Squares Reg

$$S_f(X, y) = \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2$$

$$f_\alpha(x_i) = \langle \alpha, x_i \rangle$$

solution $\alpha = (X^T X)^{-1} X^T y$

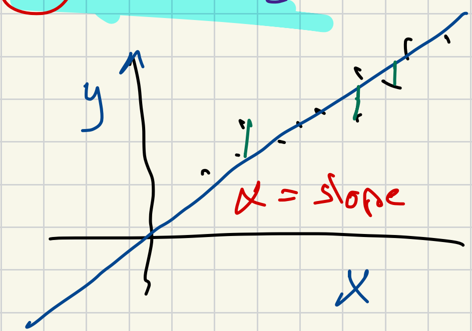
Regularized Lin Regression

$$S(\alpha; X, y) = \sum_{i=1}^{n} \left(\langle x_i, \alpha \rangle - y_i\right)^2 = \| X\alpha - y \|_2$$

parameter

regularization term

$$\underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \ \| X\alpha - y \|_2 + s \|\alpha\|_2$$



Ridge Regression

Tikhonov Regularization

$$\alpha_s = \left(X^T X + s^2 I\right)^{-1} X^T y$$

Lasso

parameter Lasso

$$\alpha_s^* = \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \|X\alpha - y\|_2 + s\|\alpha\|_1$$

- in high-dimensions $\Rightarrow$ leads to sparse solution

many $\alpha_j = 0$ $\Rightarrow$ f_α ignores these features

features not ignored $\alpha_j \neq 0$

$\hookrightarrow$ selected

larger s, fewer features selected

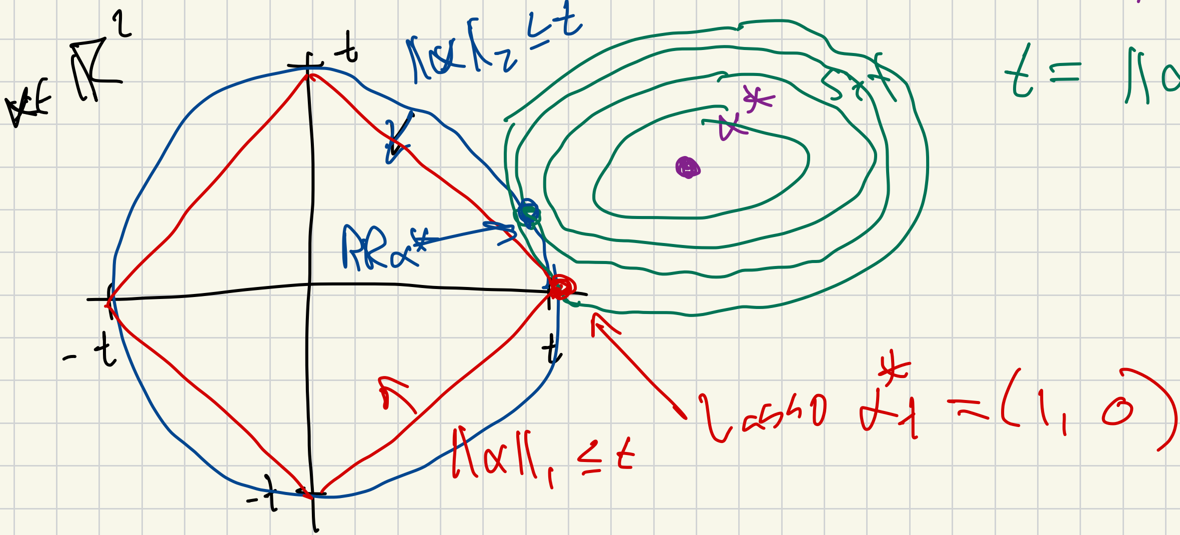
Why does Lasso give sparse solution?

Rewrite as

$$\alpha_t^* = \underset{\alpha \in \mathbb{R}^d}{\arg\min} \; \|X\alpha - y\|_2 \quad \text{s.t. } \|\alpha\|_1 \le t$$

for all s in Lasso, $\exists\, t$ s.t. $\alpha_s^* = \alpha_t^*$

$$t = \|\alpha_s^*\|_1$$



Orthogonal Matching Pursuit

Set residual $r = y$ $\qquad \alpha_j := 0 \quad \forall j$

for $i = 1$ to k

Set $x_j = \underset{x_j \in X}{\operatorname{argmax}} \, |\langle r, x_j \rangle|$

Set $\alpha_j = \underset{\gamma}{\operatorname{argmin}} \, \|r - x_j \gamma\|^2 + s|\gamma|$

Set $r = r - x_j \alpha_j$

Return α

Units

$x_i = (\text{height}_i, \text{weight}_i, \text{age}_i)$

inches, lbs, years

$$f_\alpha(x_i) = \langle x_i, \alpha \rangle = x_{i1}\alpha_1 + x_{i2}\alpha_2 + x_{i3}\alpha_3$$

not a problem

unit of $\alpha_1 = \frac{1}{\text{inches}}$

$\alpha_2 = \frac{1}{\text{lbs}}$

$\alpha_3 = \frac{1}{\text{years}}$

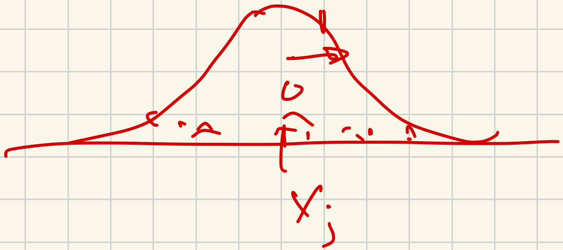
$$\alpha = \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \|X\alpha - y\| + s\|\alpha\|_1$$

$\hookrightarrow \|\alpha\|_1 = |\alpha_1| + |\alpha_2| + |\alpha_3|$

Normalize first

for each column / feature

- standardize $x_{ij} \in X_j$



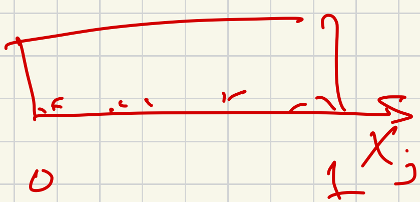
① subtract mean $\frac{1}{n}\sum_i x_{ij}$

② divide variance $\sigma^2 = \|X_j\|^2$

- 0-1 Normalization $x_{ij} \in X_j \implies \in [0,1]$

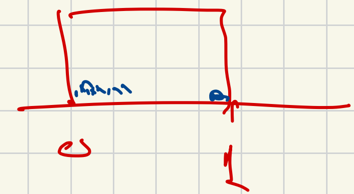
① subtract $\min_i x_{ij}$ from all x_{ij}

② divide all x_{ij} by $\max_i x_{ij}$



Problems w/ Normalization

- affected by outliers



- change 1 data point
  ↳ changes normalization.

- features may be colinear
  x_j & $x_{j'}$ have similar correlation w/ y.
  "fixed" using principal components

What can we learn from sparse

$$\alpha_s^* = \arg\min_{\alpha} \|X\alpha - y\|_2 - s\|\alpha\|_1$$

$$\alpha_s^* = (0.3, 0, 0, 0, 0.2, 0, 0, 7)$$

$$\begin{array}{ccccccccc} \bar{j} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array}$$

$\alpha_8^* \leftarrow$ largest

$\max_j \alpha_j^*$ might be most important feature

↳ maybe not!