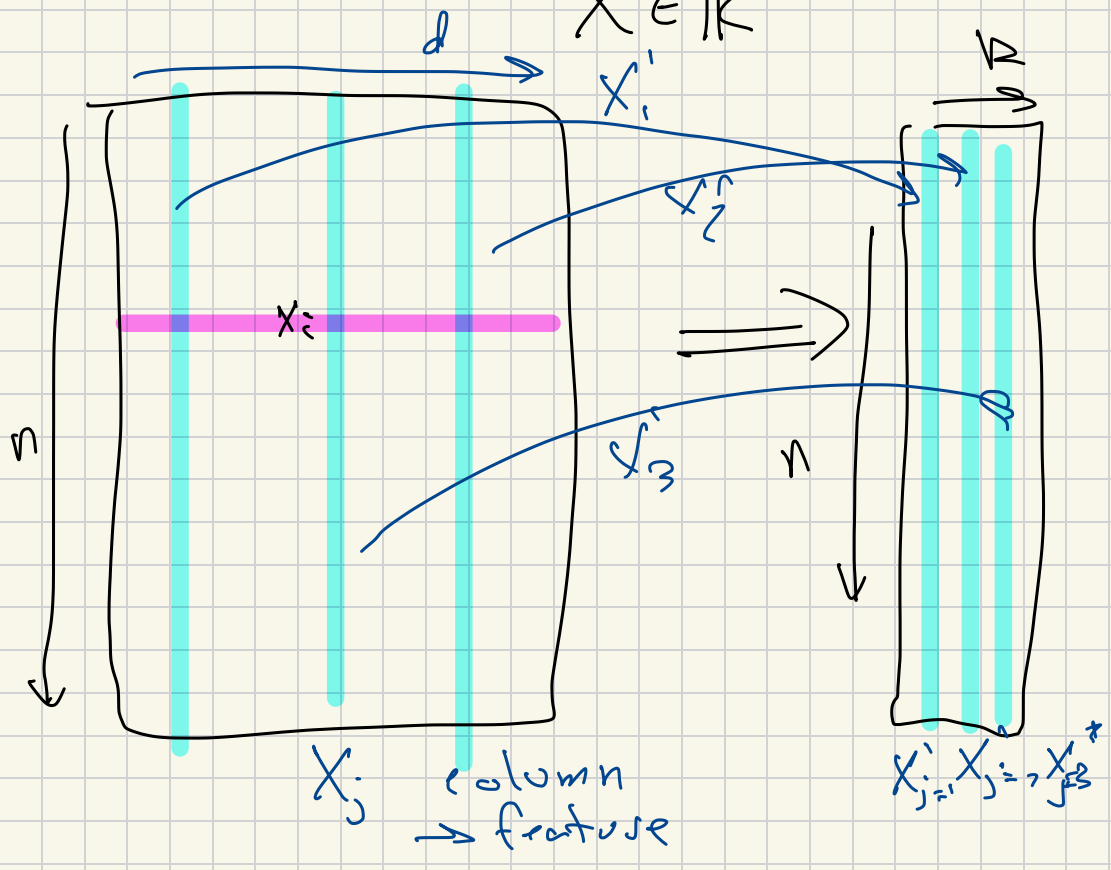
L19: Feature Selection

Data Mining
Nov 5, 2026
Jeff M. Phillips

Data Input

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

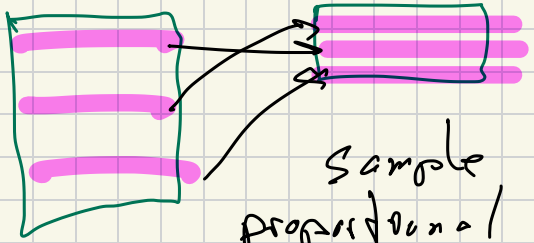
$X \in \mathbb{R}^{n \times d}$



$k < d$

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

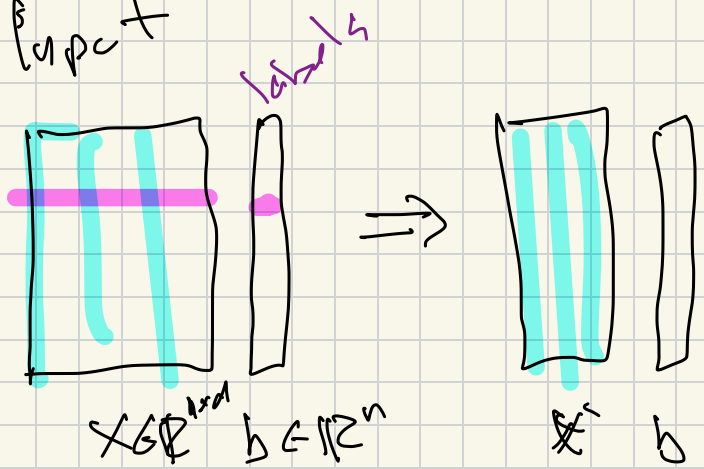
Row Sampling Matrix Sketching



Sample proportional to leverage scores, $\|a_i\|^2$

Run on columns $k = \frac{d}{\varepsilon^2}$

What is the goal of feature selection?



input $(X, b) \in \mathbb{R}^{n \times d}, \mathbb{R}^n$

point (x_i, b_i)

goal $f(x_i) \approx b_i$

OLS Regression

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

$S_f(X, b) = \sum_{i=1}^{n} (f(x_i) - b_i)^2$

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

Regularized Regression

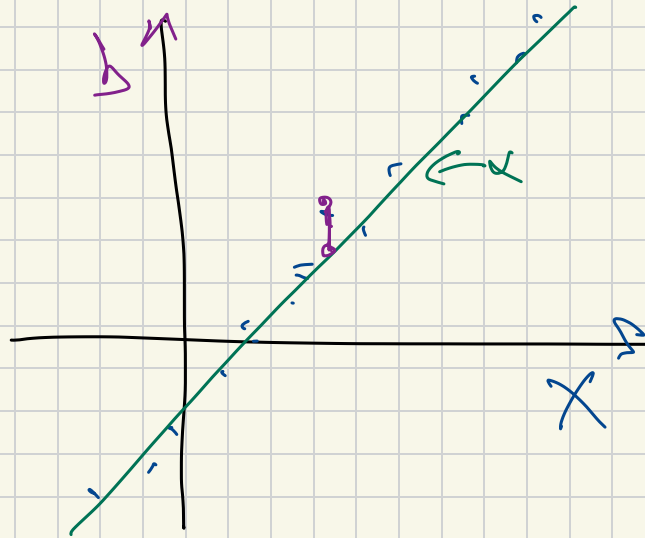
$$S(\alpha; X, b) = \|X\alpha - b\|_2^2 + s\|\alpha\|_2^2$$

Regularization term

slope small

Ridge Regression

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



Lasso

$$\alpha_S^* = \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \; \| X\alpha - b \|_2 + S \, \|\alpha\|_1$$

lasso

If d is high $\Longrightarrow$ sparse solution

many $(d-k)$ $\alpha_j = 0$ exactly

features $\alpha_j \neq 0 \Rightarrow j$ selected

larger set S, fewere features selected.

Why does Lasso give sparse α?

Rewrite

$$\alpha_t^* = \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \|X\alpha - b\|_2 \quad \text{s.t. } \|\alpha\|_1 \le t$$

$\alpha \in \mathbb{R}^{2\,(=d)}$

$\|\alpha\|_1 \le t$

$\|\alpha\|_2 \le t$

$\alpha_{t,2}^*$ radial

for every α_s^* exist t s.t. $\alpha_s^* = \alpha_t^*$

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

$(t, 0)$

α_t^* Lasso

opt

$s = 0$

$t = \infty$

Orthogonal Matching Pursuit

Set residual $r = b$ $\quad \alpha_j = 0 \ \forall j$

for $i = 1$ to k (select features incrementally)

Set $\chi_j = \underset{\chi_j \in X}{\arg\max} \ |\langle r, \chi_j \rangle|$

Set $\alpha_j = \underset{\gamma}{\arg\min} \ \|r - \chi_j \gamma\|^2 + s|\gamma|$

or solve small Lasso where $\alpha_j \neq 0$

Set $r = r - \chi_j \alpha_j$

Return α ← by definition, k non-zeros

Challenges w/ Feature Selection

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

x_{i1} inches $\quad x_{i2}$ lbs $\quad x_{i3}$ years

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

A-OK if unit $\alpha_1 = \frac{1}{\text{inches}} \quad \alpha_2 = \frac{1}{\text{lbs}} \quad \alpha_3 = \frac{1}{\text{years}}$

Lasso $\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|X\alpha - b\| + s\|\alpha\|_1$

$\frac{1}{\text{inches}} + \frac{1}{\text{lbs}} + \frac{1}{\text{years}}$

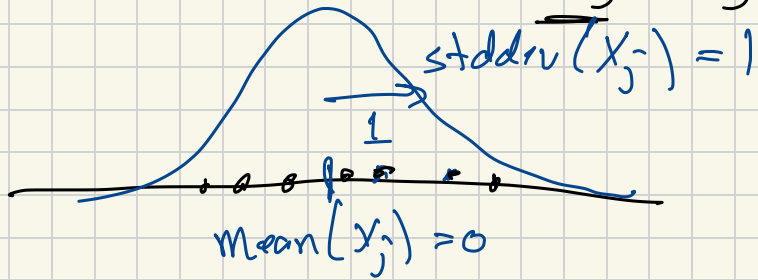
$\hookrightarrow \sum_{j=1}^{d} |\alpha_j| = |\alpha_1| + |\alpha_2| + |\alpha_3|$

problem!

Normalization first

for each column / feature

- standardize $X_{ij} \in X_j$

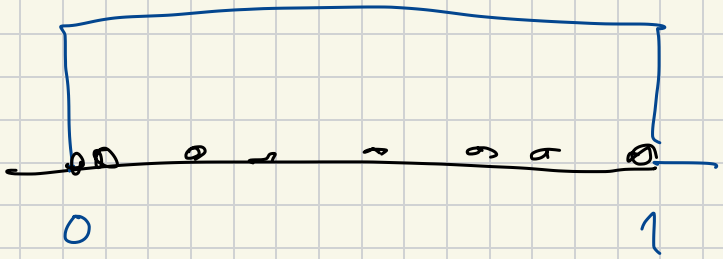


(1) subtract $\frac{1}{n}\sum_{i=1}^{n} X_{ij} = \mu_j$ (mean)

(2) divide by $\sqrt{\text{variance}}$

$\sigma^2 = \|X_j - \mu_j\|^2$

- (0-1) normalization



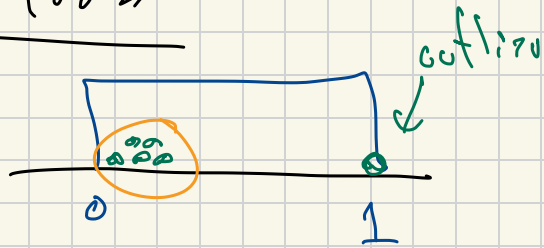
$X_{ij} \Rightarrow [0,1]$

(1) subtract $\min_i X_{ij}$ $\forall X_j$

(2) divide $\max(X_{ij} - \min)$

Problems w/ Normalization

1. affected by outliers



2. change 1 data point
   ↳ change normalization

3. features may be co-linear
   X_j, $X_{j'}$ are very correlated.
   ↳ don't know which to use.

PC linear combos
↳ new features
↳ "fix" mapping
principal components ⇒ regression
↳ orthogonal
PCR

What can we learn from sparse models

$$\alpha_s^* = \underset{\alpha}{\operatorname{argmin}} \|X\alpha - b\|_2 - s\|\alpha\|_1$$

$$\alpha_s^* = (0.3,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0.2)$$

j : 0 1 2 3 4 5 6 7

highest coefficient

select $j=4$ → most important

SHAP, LIME → sometimes work → sometimes are fooled.

Maybe correct.