Noise

Data Science Club

Tomorrow Tue, Apr 3 @5pm

Abode + pizza

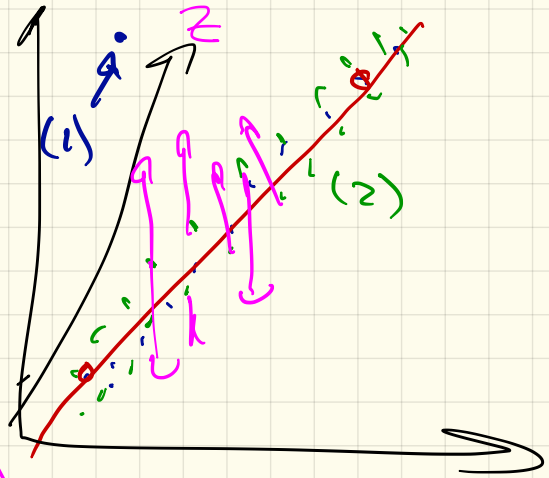
WEB 230

Noise in Data

(1) Spurious Readings (outliers)

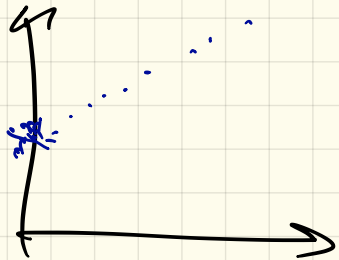
(2) Measurement Error
hopefully unbiased

(3) Background / Missing Data



- Cross-Validation
- Explicitly handling Outliers

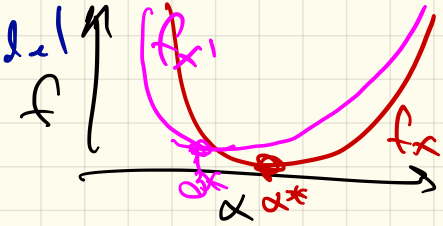
- Regularization
- Robust Statistics



Most Modeling

cost $f(X, \alpha)$ — data X, model α

goal $\alpha^* = \arg\min_{\alpha} f(X, \alpha)$

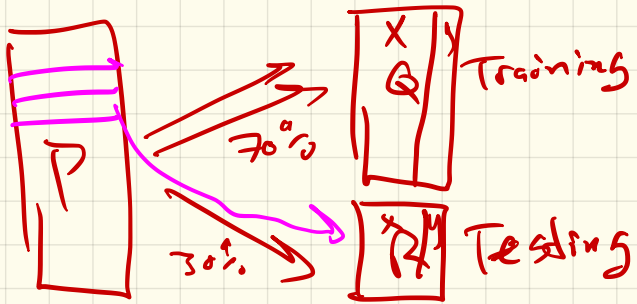


$$f(P_1, (\alpha, \gamma)) = \|P_y - P_x \alpha\|_2^2 + \gamma \|\alpha\|_2^2 \quad \text{(Ridge)}$$

if γ fixed → solv α^* closed form

$$\alpha_{X,\gamma}^* = g(X, \gamma)$$ — data X

$\mathcal{U}$ all data observe i.i.d



for all γ

$$\alpha_{Q,\gamma}^* = g(Q, \gamma)$$

Evaluate $\gamma^* = \arg\min_{\gamma} f(R_1, (\alpha_{Q,\gamma}^*, \gamma))$

use α_{Q,γ^*}^*

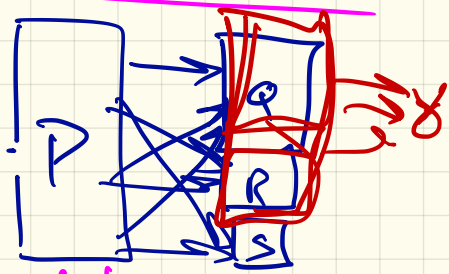
Evaluate $\gamma^* = \underset{\gamma}{\operatorname{argmin}} f(R, (\alpha^*_{(Q,\gamma)}, \gamma))$

$g(Q, \gamma)$

Cross-Validation

(1) Choose parameters

(2) Predict Generalization



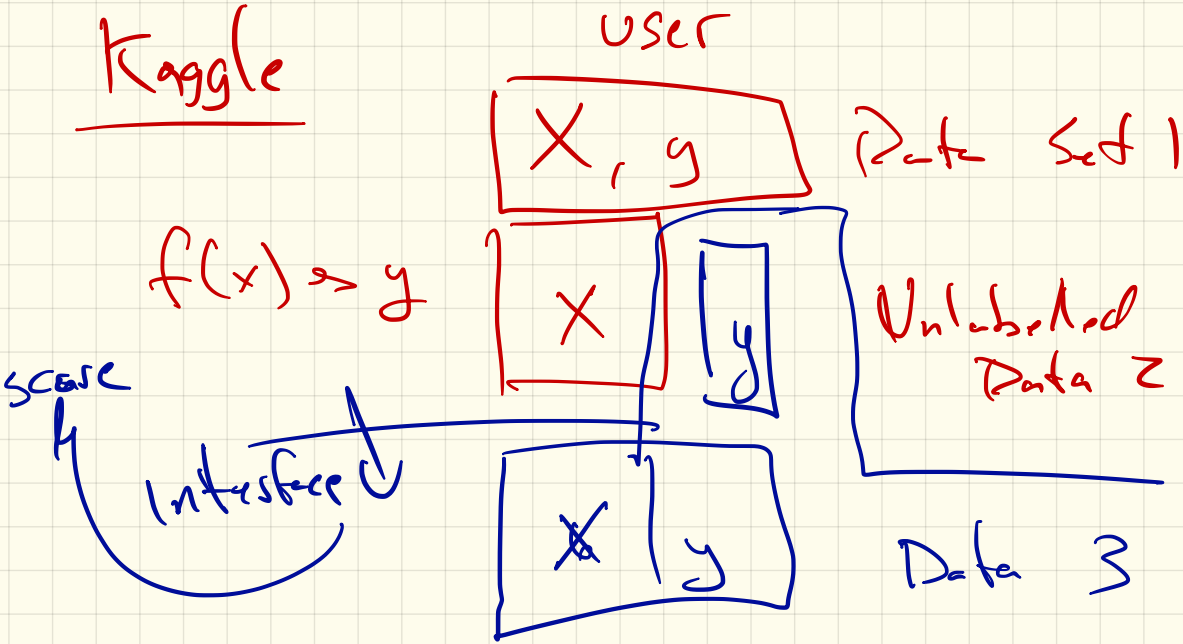
Train on Q

Chose Param on R

Generalize on S.

if we saw new data?

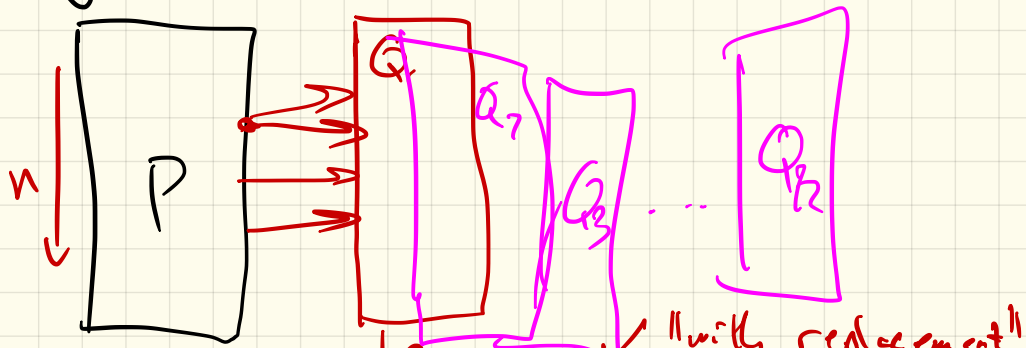
how well will our model do?


Kaggle
user
X, y
Data Set 1
f(x) -> y
X
y
Unlabeled
Data 2
score
Interface
X
y
Data 3

Bootstrapping

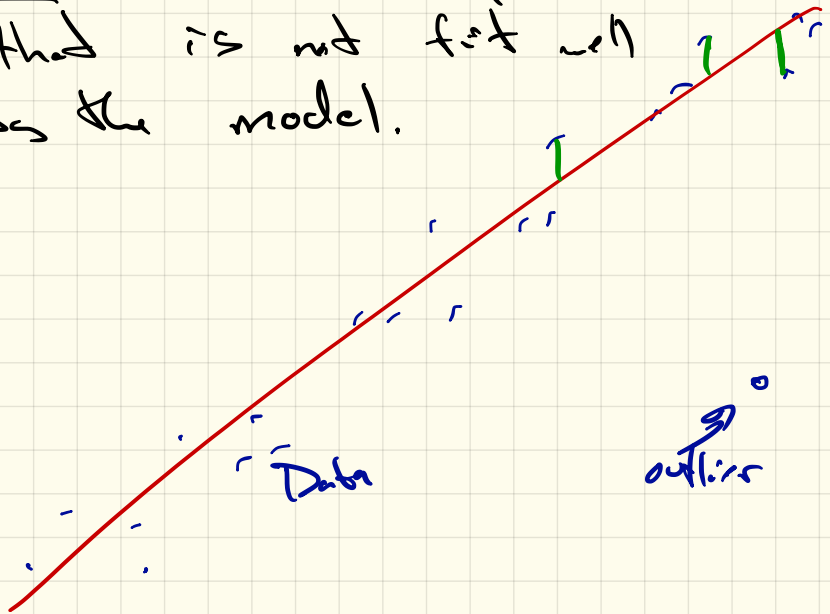
(Efron 1979)

$\mathcal{U}$ $\int_{iid}$



Outliers

Outlier is a datapoint that is not fit well by the model.



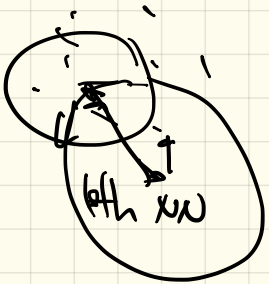
Algs for outlier removal

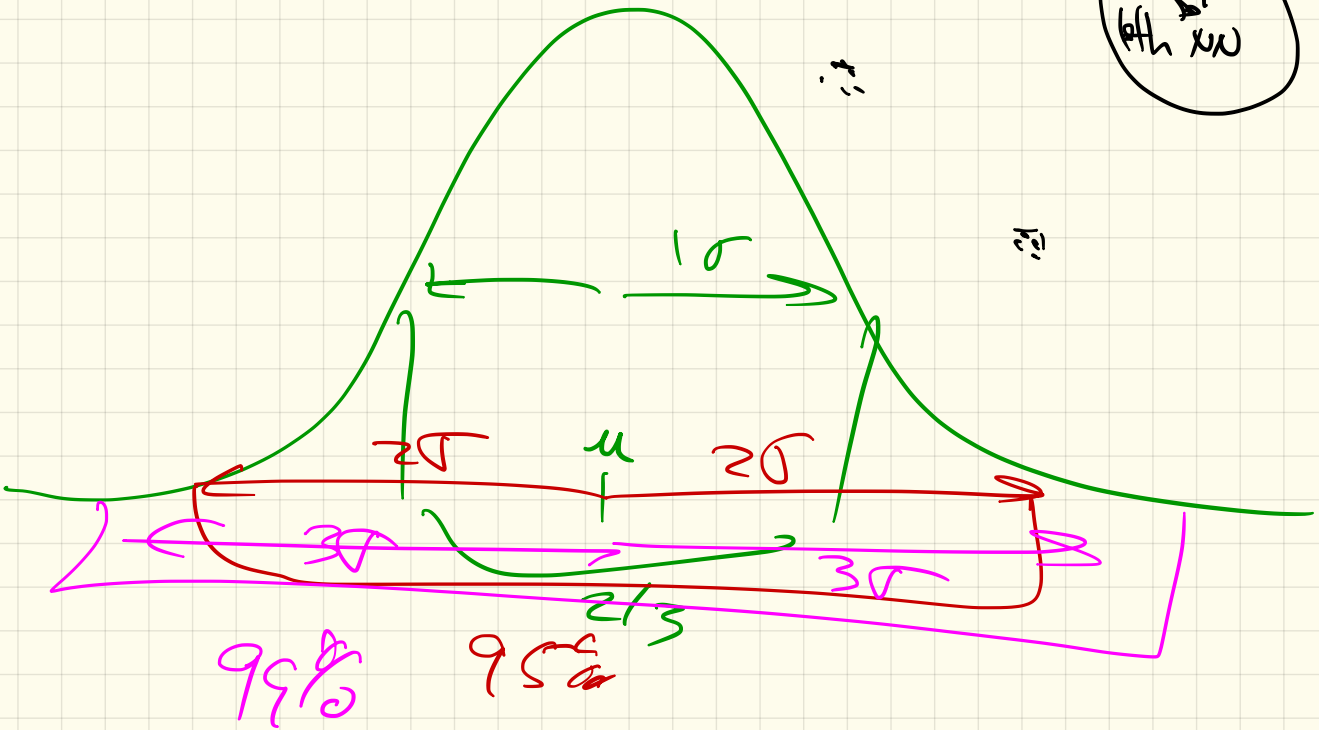
1. Fit Model
2. Calc Resid
3. Remove $x \in X$ if resid too big.

- Density Rule

Reverse NN

$e^{-\|x-\mu\|^2}$





Heavy-Tailed Distr.

Zipf's $f_i = c \cdot \left(\frac{1}{i}\right)^{\alpha}$

the 7%
of 3.5%
and 2.8%