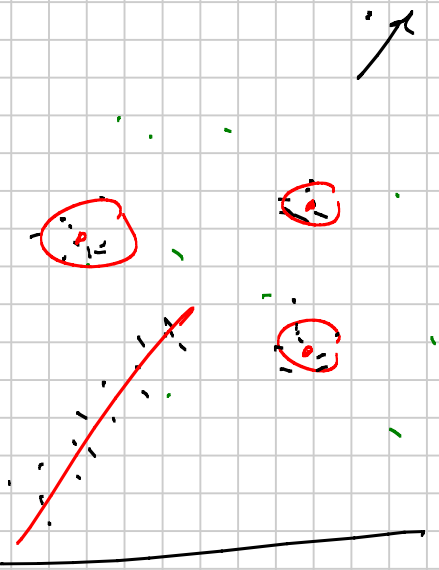
Noise in Data

Note Title 3/30/2016

- Spurious Readings (outliers)
- Measurement Error
  small full-dimensional noise
- Background Data
  missing data



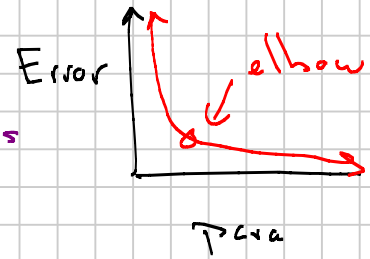
- Cross-Validation
- Outliers
- Uncertain Data
- Robust Estimators

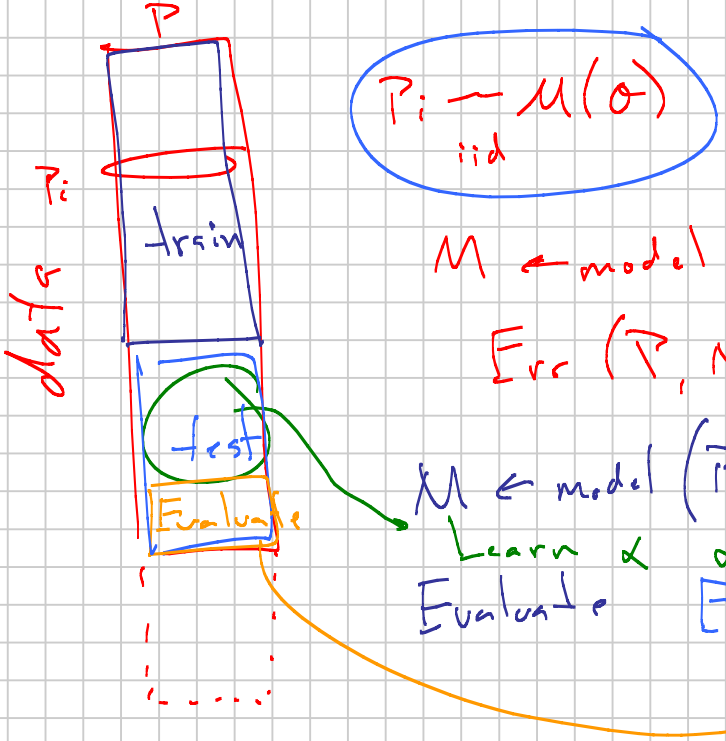
Cross-Validation

Algo deals with noise, has a parameter

- Clustering → k
- PCA → # principal components
- Regression
  - degree polynomial
  - Regularization parameter

$$L_p(A,\alpha) = \| P_y - P_x A \|_2 + \alpha \|A\|_2$$





$P_i \sim_{iid} M(\alpha)$

$M \leftarrow$ model

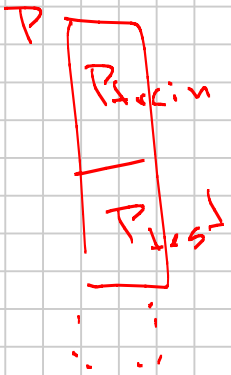
$Err(P, M) = L_P(M, \alpha)$

$M \leftarrow \text{model}(P_{train}, \alpha) = L_P(M, \alpha)$

Learn α $\quad \alpha = \underset{\alpha}{\text{argmin}}\ L_T(M_\alpha, P_{test})$

Evaluate $\quad Err(P_{test}, M)$

P_{eval}



$P = [P_x, P_y]$ goal is find linear regression

Ridge Regression

$L_P(A, \alpha) = (\|P_y - P_x A\|_2 + \alpha \|A\|_2$

$\hookrightarrow \hat{A}_\alpha = \left(P_{x\,train}^T P_{x\,train} + \alpha^2 I\right)^{-1} P_{x\,train} P_{y\,train}$

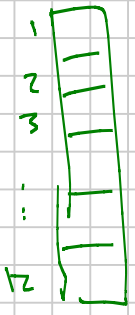
Goal: Find α

minimizes $L_{P_{test}}(\hat{A}_\alpha(P_{train}), 0)$

$L_{P_{test}}(\ldots, 0) = \|P_{y,test} - P_{x,test} \hat{A}_\alpha\|_2$

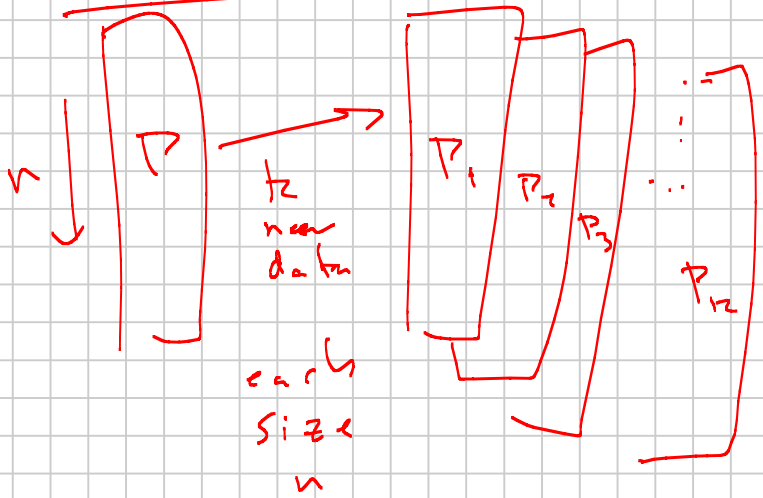
C-V

- k-fold C-V → Leave-one-out C-V



- train on all sets of k-1 parts
- test on other

Bootstrapping



- Sample with replacement

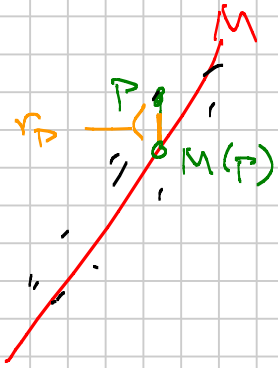
$P_i \leftarrow n$ items from P i.i.d.

$M_i \leftarrow \text{model}(P_i)$

$\{M_1, M_2, \ldots, M_k\}$

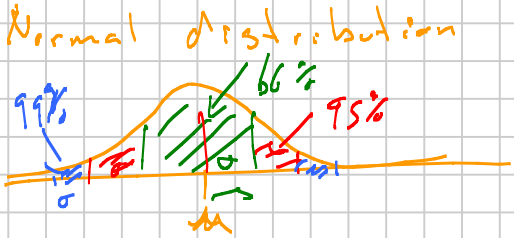
Outliers

- Find + Remove Approach
- Density-based Approach



1. Build model M from input P.
2. For all $p \in P$, calc $r_p = d(M(p), p)$
3. If r_p is "too large" it's an outlier!
4. Throw outliers away

iterate

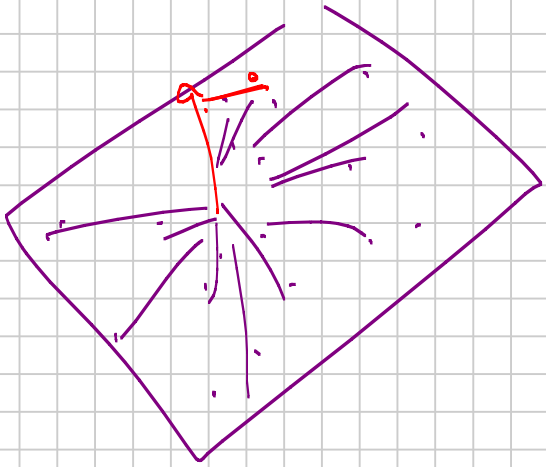


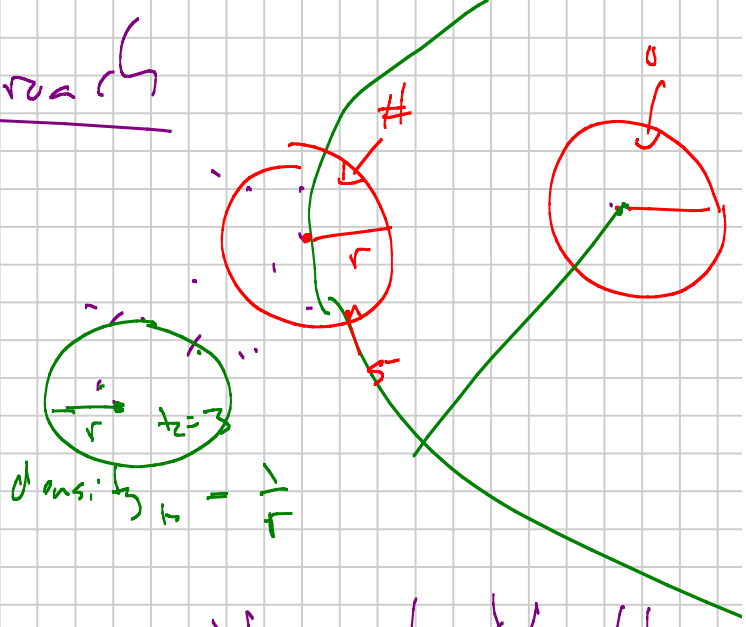
$P(x) = e^{-\|\mu - x\|^2}$

- 5% largest / 50 datapoints
  - ↳ run out
- Pick threshold on r_p
  - "scale"

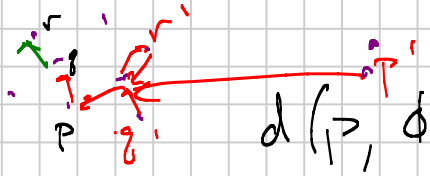
Density-based Approach

P





Reverse Nearest Neighbor



$d(p, \phi_X(p) = q)$

$\approx d(q, \phi_X(q) = ?)$

Heavy-Tail Distributions

Zipf $P(i) = c \cdot \left(\frac{1}{i}\right)$

the 7%

of 3.5%

and 2.8%