Noise in Data

- 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

- Clustering
- PCA + principal components
- Regression
  - Degree polynomial
  - Regularization parameter

\[ \text{Error vs. Parameter} \]

\[ L_p(A, \alpha) = \| P_y - P_x A \|_2 + \alpha \| A \|_2 \]
\[ P_i - M(\alpha) \]

\[ \text{M} \in \text{model} \]

\[ \text{Err} (P, M) = L_p (M, \alpha) \]

\[ M \in \text{model} (P_{\text{train}}, \alpha) = L_p (M, \alpha) \]

\[ \text{Learn } \alpha = \arg \min_{\alpha} L_p (M^\alpha, P_{\text{test}}) \]

\[ \text{Evaluate } \text{Err} (P_{\text{test}}, M) \]

\[ P = [P_x, P_y] \]

\[ \text{goal is find lineare regression } \]

\[ \text{Ridge Regression} \]

\[ L_p (\hat{\alpha}, \alpha) = (|P_y - P_x A|^2 + \alpha A^2) \]

\[ \Rightarrow \hat{\alpha} = \left( P_x P_x^T + \alpha^2 I \right)^{-1} P_x P_y \]

\[ \text{Goal: Find } \alpha \]

\[ \text{minimizes } L_{P_{\text{test}}} (\hat{\alpha}, (P_{\text{train}}, \alpha)) \]

\[ L_{P_{\text{test}}} (\ldots ; \alpha) = |P_y_{\text{test}} - P_x \hat{\alpha}| \]

\[ \text{C-V} \]

\[ \text{0-folds C-V} \]

\[ \text{Leave-out one fold C-V} \]

\[ \text{Train on all sets of k-1 points} \]

\[ \text{Test on kth set} \]
Bootstrapping

- Sample with replacement
- \( T \in n \) items from \( \mathcal{R} \)
- \( M_i \in \text{model}(P_i) \)
- \( \{M_1, M_2, \ldots, M_k\} \)

Outliers

- Find + Remove Approach
- Density-based Approach

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

Normal Distribution

- 95% largest 1/50 data points
- \( P(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)
Density-based Approach

Reverse Nearest Neighbor

Heavy-Tail Distributions

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

the 75% of 3.55° and 7.8°