Regression

- In regression problems, label y is continuous (usually y is scalar)
- y also called response
- Supervised learning problem: given training data $\mathcal{T} = \{\underline{x}_i, y_i\}_{i=1}^{N}$, find a model f to predict $\hat{y} = f(\underline{x})$

Applications: prediction of house prices based on location attributes, prediction of income based on demographic features, etc.

Loss Function Is "0-1" loss useful? No.

Commonly used loss model for regression,

Square loss: $L(f(\underline{x}), y) = (y - f(\underline{x}))^2$

Linear Regression $f(\underline{x}) = \underline{\beta}^T \underline{x}$

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}$$ learned from training data

Example Prediction of home price

x_1 = crime rate, x_2 = pollution level

$f(\underline{x}) = \beta_1 x_1 + \beta_2 x_2$.

Linear regression can be used to do more general (non linear) modeling by augmenting $\underline{x}$

Example:

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longrightarrow \underline{x}' = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \end{bmatrix}$$

$d = 2$ $\qquad$ $d' = 4$

Predict response using $\underline{x}'$ as: $\beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1^2$

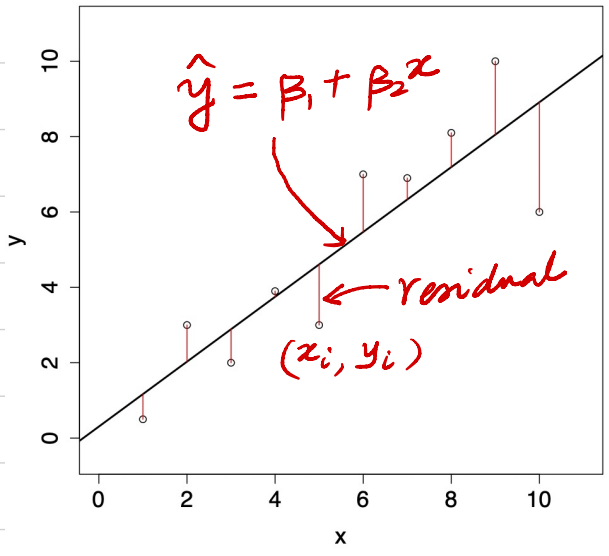
Learning $\underline{\beta}$ from $\tau = \{\underline{x}_i, y_i\}_{i=1}^N$

For squared loss, define Residual sum of squares:

$$RSS(\underline{\beta}) = \sum_{i=1}^{N} (y_i - \underbrace{\underline{x}_i^T \underline{\beta}}_{\hat{y}_i})^2$$

Choose $\underline{\beta}$ to minimize $RSS(\underline{\beta})$

$$\hat{\underline{\beta}}_{LS} = \arg\min_{\underline{\beta}} RSS(\underline{\beta})$$



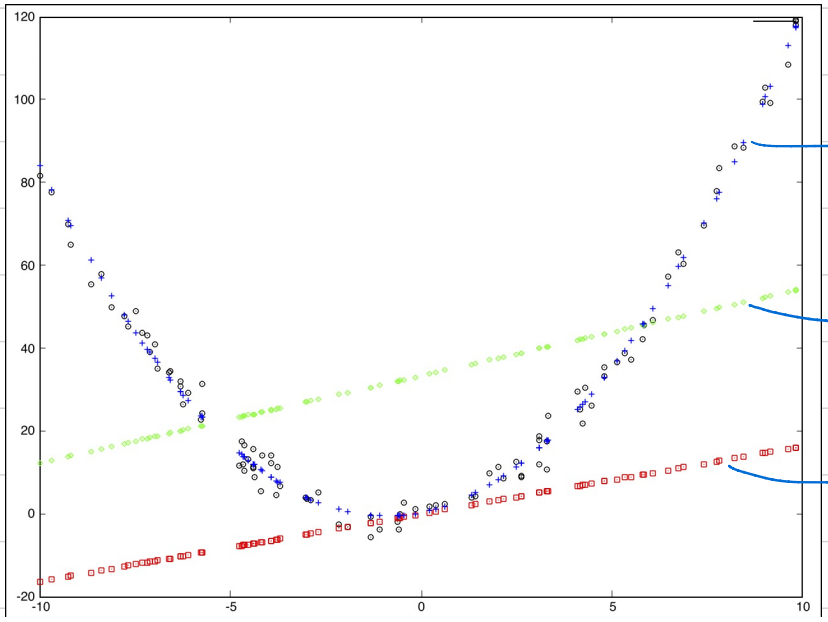
$$RSS(\underline{\beta}) = \| \underline{y} - X \underline{\beta} \|^2$$

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N \times 1}, \quad X = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}_{N \times d}$$

$$\hat{\underline{\beta}}_{LS} = (X^T X)^{-1} X^T \underline{y}$$

Model Selection

Train and validate (or cross-validate) several models and choose best one



Subset (Feature) Selection

- Often some features are irrelevant or of minor relevance to prediction $\hat{y}$, e.g. average crime rate in Illinois to home prices in Urbana.

- Goal of feature selection is to automatically give small (or zero) weight β_i to features that are irrelevant. This results in simpler models with higher accuracy

- Shrinkage Formulation:

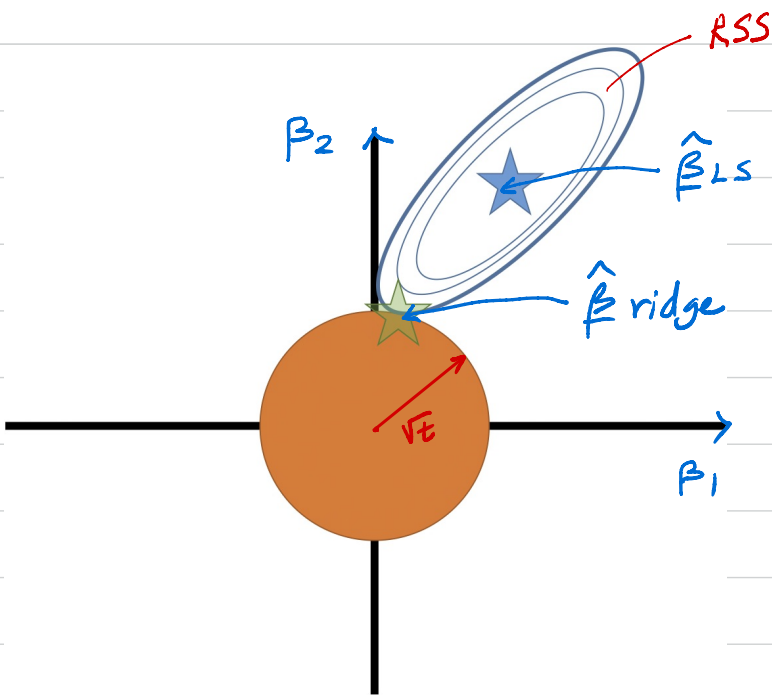
  Minimize $RSS(\beta)$ subject to $r(\beta) \leq t$

  ($r(\beta)$: regularizer)

Ridge Regression

$r(\beta) = \|\beta\|^2 = \sum_i \beta_i^2$

Minimize $RSS(\beta)$ subject to $\|\beta\|^2 \leq t$

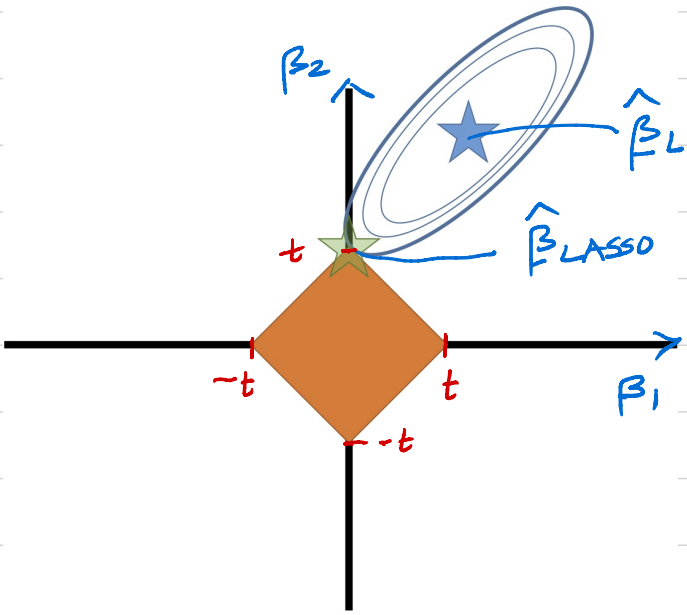


- Can solve for $\hat{\beta}_{ridge}$ in closed-form (see typed notes)
- $\hat{\beta}_j$'s are kept small but not forced to zero for less relevant features

LASSO (Least Absolute Shrinkage and Selection Operator)

$r(\beta) = \sum_i |\beta_i| \leftarrow L_1\text{-norm}$

Minimize $RSS(\beta)$ subject to $\sum_i |\beta_i| \leq t$



- LASSO constraint promotes sparsity
- Least relevant features can have β_j forced to zero