Discussion 1

Review

Least squares & nearest neighbours

RSS & MSE & EPE

Variable Types and Terminology

Input: a variable X. If X is a vector, its j-th element is X_j

an observation x_i
(scalar or vector)

Typically, we use i to denote the index of observations,
while use j to denote the index of variables.

Model

Scalar

$$N \text{ observations} \left\{ \begin{array}{c} X \\ \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_N \end{bmatrix} \end{array} \right. \quad \mathbf{x} \in \mathbb{R}^N$$

Vector

$$N \text{ observations} \left\{ \begin{array}{c} X_1 \quad \cdots \quad X_j \quad \cdots \quad X_p \\ \underbrace{\begin{bmatrix} \text{---} \; x_1^T \; \text{---} \\ \vdots \\ \text{---} \; x_i^T \; \text{---} \\ \vdots \\ \text{---} \; x_N^T \; \text{---} \end{bmatrix}}_{p \text{ variables}} \end{array} \right. \quad \mathbf{X} \in \mathbb{R}^{N \times p}$$

Simple Approach 1: Least Squares

- Training procedure:
  Method of *least-squares*
- N = #observations
- Minimize the *residual sum of squares*

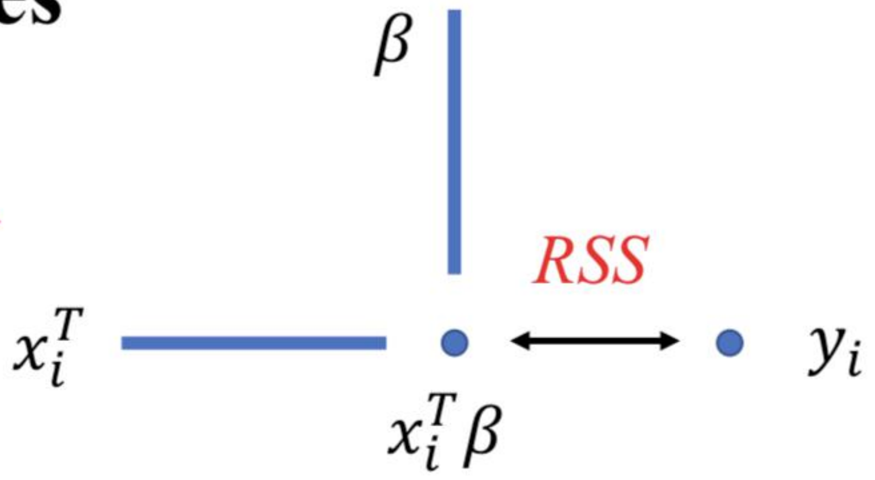
$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \boxed{(y_i - x_i^T\beta)^2}$$

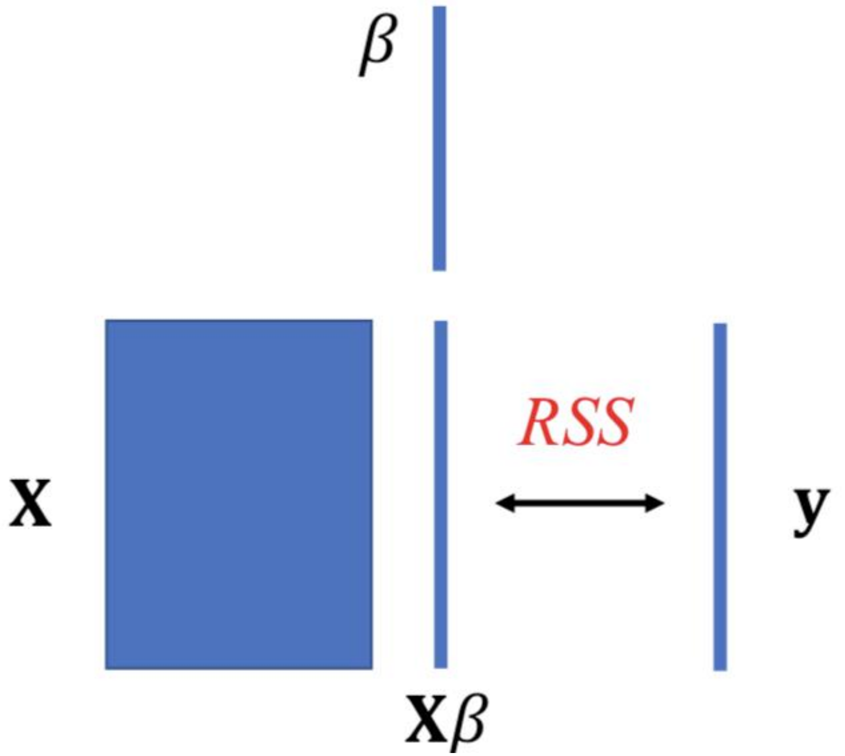
Or equivalently,

$$\mathrm{RSS}(\beta) = \boxed{(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)}$$
$$= \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

- This quadratic function always has a global minimum, but it may not be unique.

Q: What is the difference among x_i, x_i^T, $\mathbf{x}$, X and $\mathbf{X}$?





- *(scalar to scalar)* $\mathrm{d}f = f'(x)dx$
- *(scalar to vector)* $df = \sum_i \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$
- $\frac{\partial Ax}{\partial x} = A^T, \frac{\partial x^T A}{\partial x} = A$

$$
\begin{aligned}
RSS(\beta) &= (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) \\
&= y^T y - 2\beta^T \mathbf{X}^T y + \beta^T \mathbf{X}^T \mathbf{X}\beta \\
\frac{\partial RSS(\beta)}{\partial \beta} &= -2\mathbf{X}^T y + 2\mathbf{X}^T \mathbf{X}\beta = 0 \\
&\Rightarrow \quad \mathbf{X}^T \mathbf{X}\beta = \mathbf{X}^T y \\
&\Rightarrow \quad \hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^T y
\end{aligned}
$$

$$\mathrm{RSS}(\beta) = (\mathbf{X}\beta - \mathbf{y})^T \mathbf{W}(\mathbf{X}\beta - \mathbf{y})$$

$$\begin{aligned}
\nabla_\beta \mathrm{RSS}(\beta) &= \frac{\partial\, \mathrm{RSS}(\beta)}{\partial \beta} \\
&= \frac{\partial}{\partial \beta}(\mathbf{X}\beta - \mathbf{y})^T \mathbf{W}(\mathbf{X}\beta - \mathbf{y}) \\
&= 2\mathbf{X}^T \mathbf{W}(\mathbf{X}\beta - \mathbf{y}) \\
&= 0 \\
\Rightarrow \hat{\beta} &= (\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T \mathbf{W}\mathbf{y}.
\end{aligned}$$

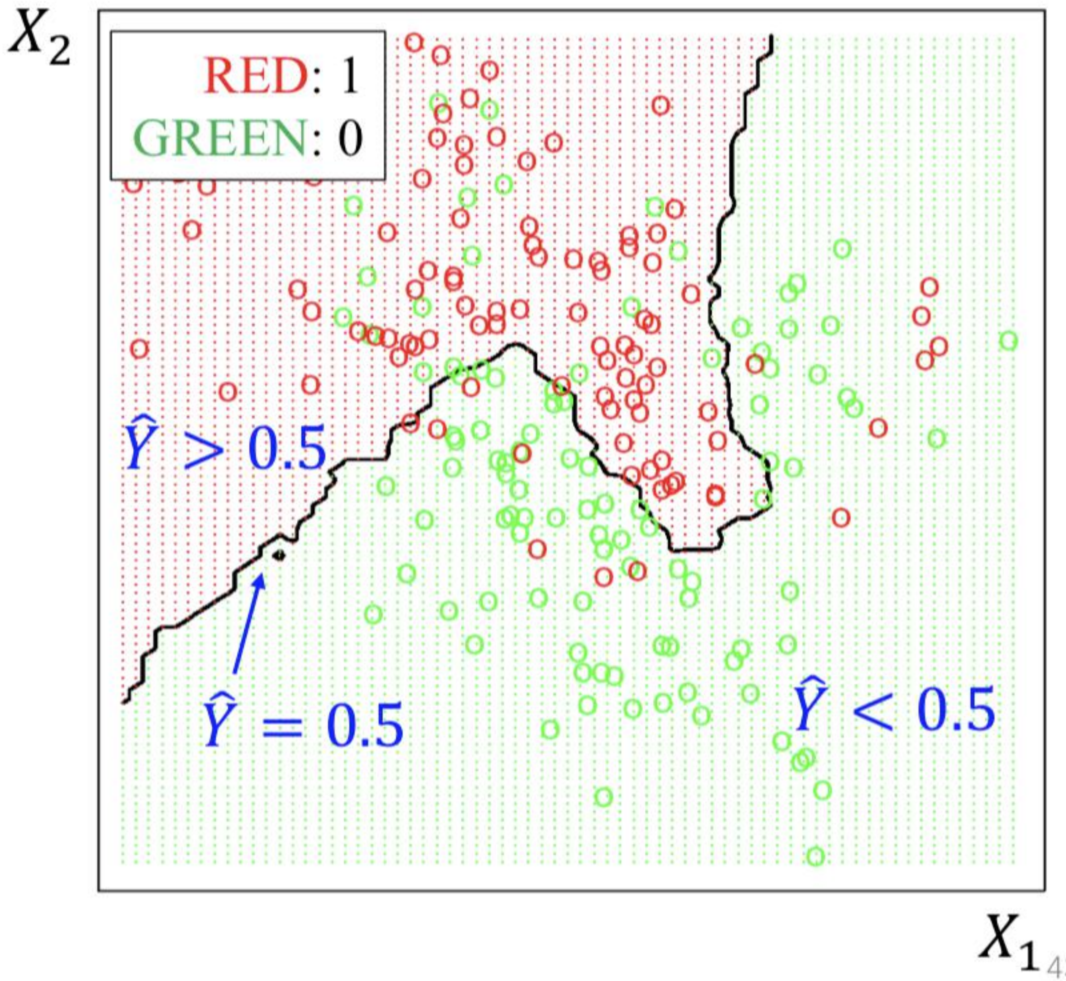
Simple Approach 2: Nearest Neighbors

- Use observations in the training set closest to the given input.

$$\hat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i.$$

- $N_k(x)$ is the set of the k closest points to x is the training sample
- Average the outcome of the k closest training sample points
- Fewer misclassifications

15-nearest neighbors averaging



RSS： training error

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2$$
$$= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$
$$= \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

EPE： generalization error， out-of sample error

$$\mathrm{EPE}(f) = \mathrm{E}(Y - f(X))^2$$
$$= \int (y - f(x))^2 \Pr(dx, dy).$$

SE: $L(Y, f(X)) = (Y - f(X))^2.$

MSE: mean squared error, in-sample error

$$\begin{aligned}\mathrm{MSE}(x_0) &= \mathrm{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\ &= \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 \\ &\quad + [\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\ &= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0)\end{aligned}$$

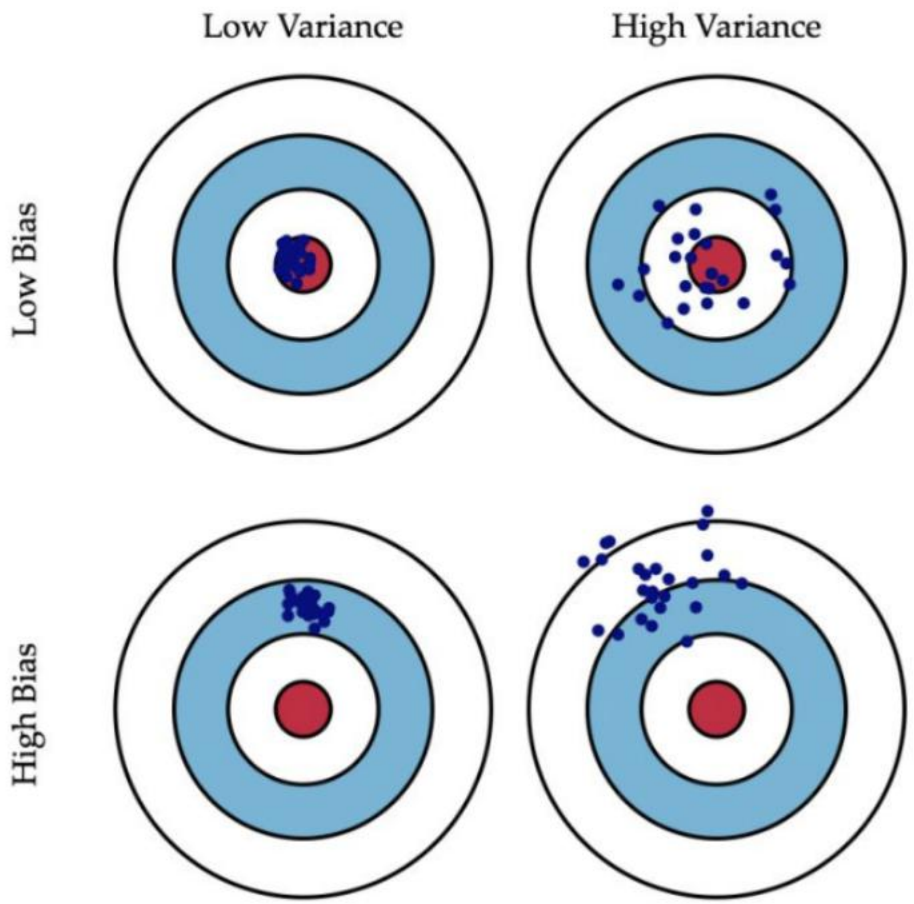


Figure 1: The difference between Bias and Var.

Regression function: $f(x) = E(Y|X = x)$.

$$
\begin{aligned}
\frac{\partial}{\partial f}\mathbb{E}_{Y|X}[(Y-f)^2|X=x] &= \frac{\partial}{\partial f}\int [y-f]^2 \Pr(y|x)dy \\
&= \int \frac{\partial}{\partial f}[y-f]^2 \Pr(y|x)dy \\
&\Rightarrow 2\int y \Pr(y|x)dy = 2f\int \Pr(y|x)dy \\
&\Rightarrow 2\mathbb{E}[Y|X=x] = 2f \\
&\Rightarrow \hat{f}(x) = \mathbb{E}[Y|X=x].
\end{aligned}
$$

$$\hat{G}(x) = \underset{k \in \mathcal{G}}{\operatorname{argmax}} \Pr(G = k | X = x)$$

$$\mathrm{EPE} = \mathrm{E}_X \sum_{k=1}^{K} L[\mathcal{G}_k, \hat{G}(X)] \Pr(\mathcal{G}_k | X)$$

- Bayesian methods
- Formula for joint probabilities

$$\Pr(X,Y) = \Pr(Y|X)\Pr(X)$$
$$= \Pr(X|Y)\Pr(Y)$$

- Bayes's theorem

Likelihood

Prior probability for Y

$$\Pr(Y|X) = \frac{\Pr(X|Y)\Pr(Y)}{\Pr(X)}$$

Posterior probability for Y

Evidence