PS5841

Data Science in Finance & Insurance

Variance Bias

Yubo Wang

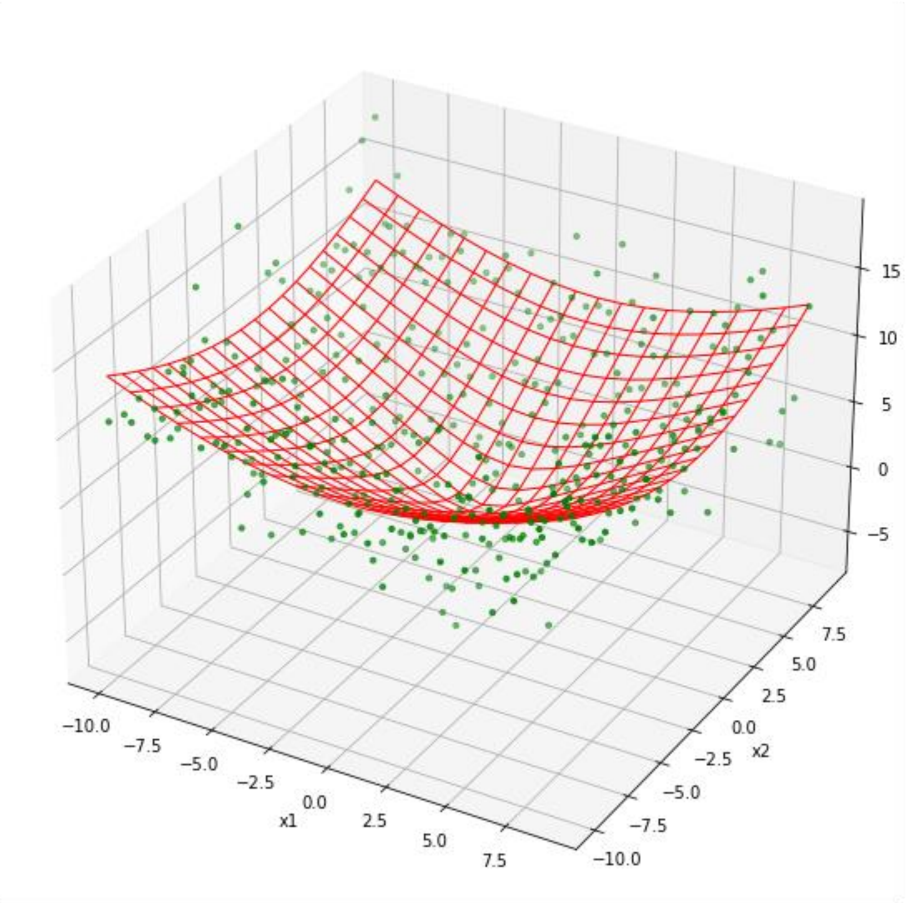
Spring 2022

COLUMBIA
UNIVERSITY

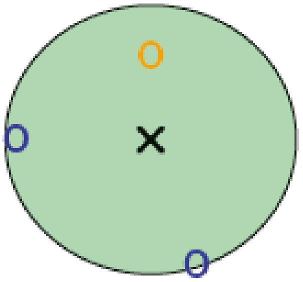
Quantitative Response

- $\hat{y} = E(Y|X)$ has the lowest possible test MSE:
$$E\left((y-\hat{y})^2\right)$$
- $EPE(\hat{y})$
$$= \sigma_\epsilon^2 + [E_\tau(\hat{y}_0) - f(x_0)]^2 + E_\tau([\hat{y}_0 - E_\tau(\hat{y}_0)]^2)$$
$$= \sigma_\epsilon^2 + bias^2(\hat{y}_0) + Var(\hat{y}_0)$$
- A more flexible model tends to have a higher variance than a less flexible one

Example : Hyperboloid of Two Sheets



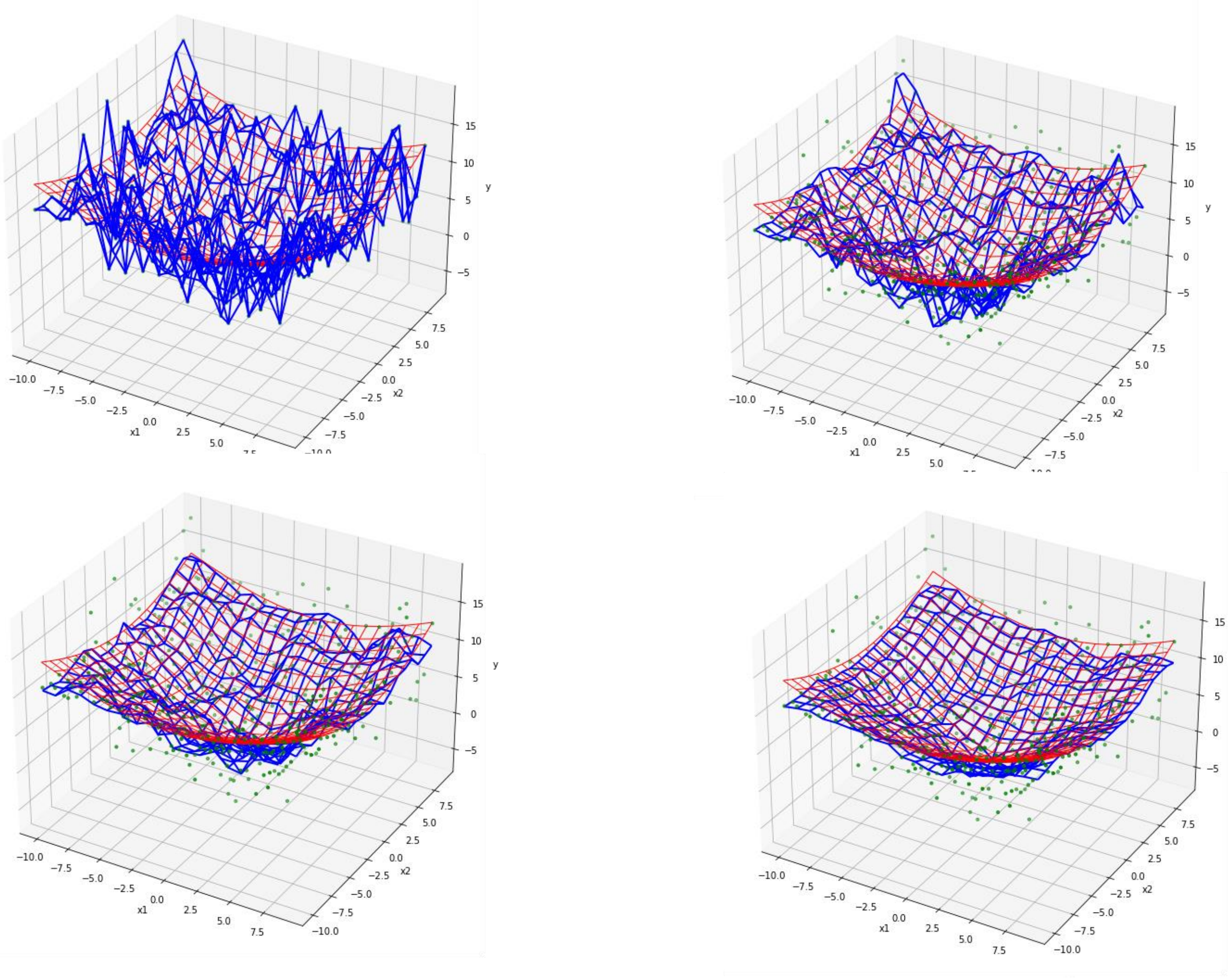
K-Nearest Neighbor Regression



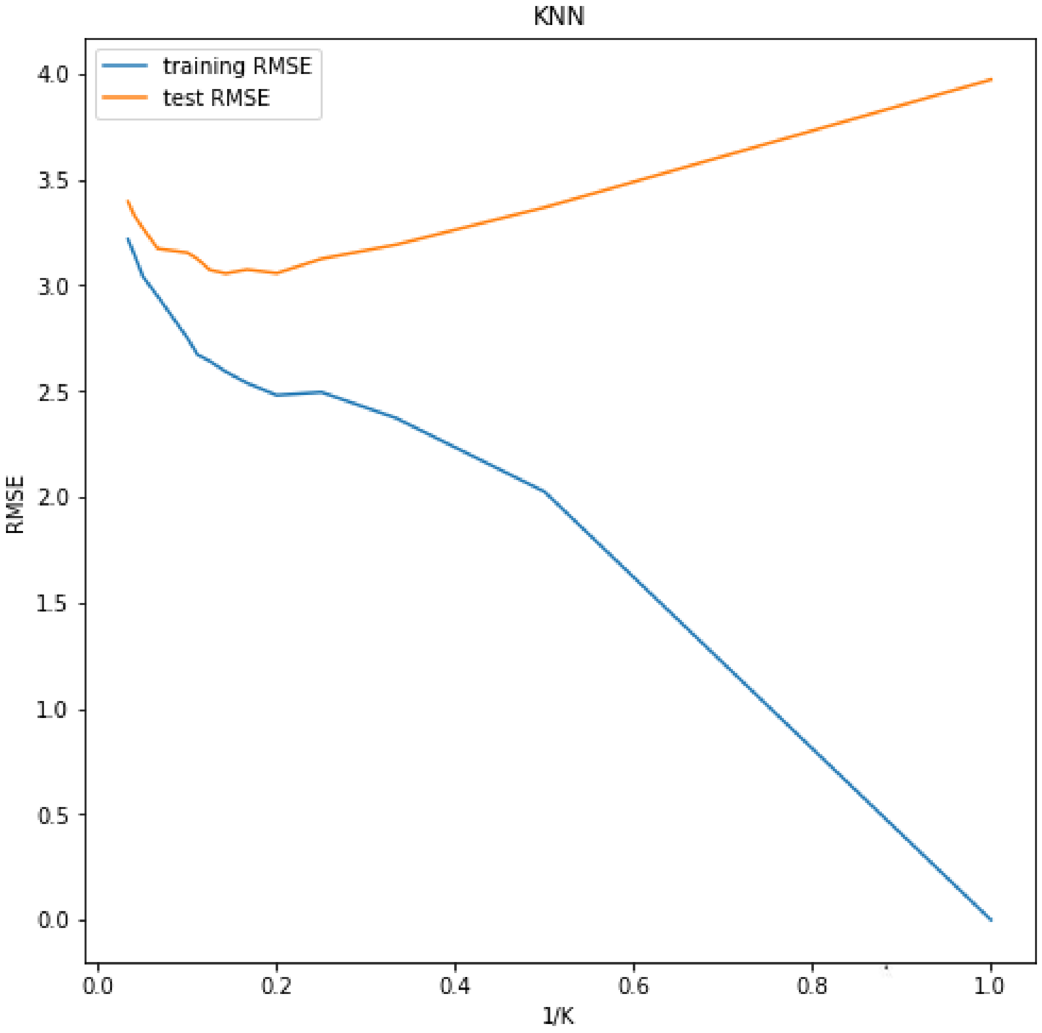
$$\hat{y} = E(Y|X = \boldsymbol{x}_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} y_i$$

- Predicted response is the mean response in the neighborhood

Example : Hyperboloid of Two Sheets KNN (K=1,3,7,14)



Example : Hyperboloid of Two Sheets
KNN: RMSE vs. 1/K



Qualitative Response

- Bayes Classifier $\hat{y} = \underset{k}{\operatorname{argmax}} \Pr(Y = k|X)$
- Bayes Classifier has the lowest possible test error rate: $E(I_{y \neq \hat{y}})$
- Overall Bayes error rate

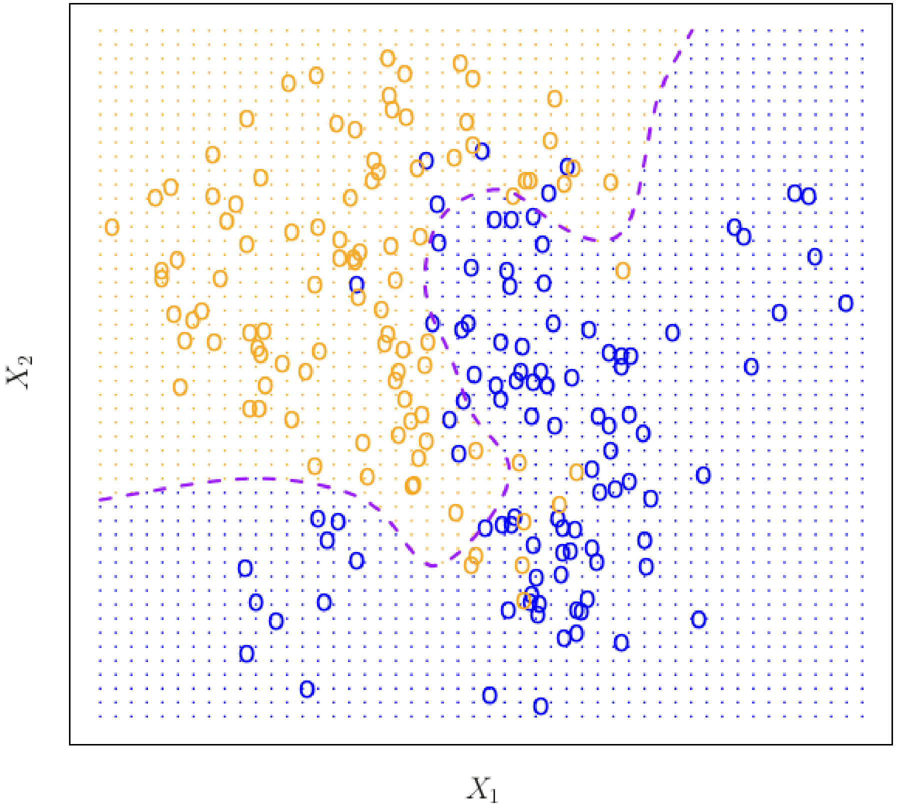
$$1 - E\left(\max_{k} \Pr(Y = k|X)\right)$$

  – The expectation averages the probability over all possible values of X

- A more flexible model tends to have a error rate than a less flexible one

Decision Boundary

Bayes Decision Boundary



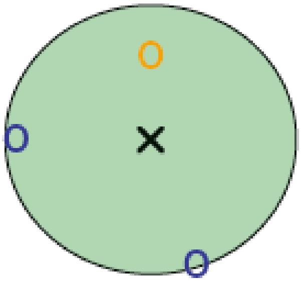
Sample from a Discrete Distribution

$$X = \begin{cases} x_1, & \Pr(X = x_1) = p_1 \\ & \vdots \\ x_k, & \Pr(X = x_k) = p_k \end{cases}$$

$$U \sim \text{Uniform}(0,1)$$

$$x = \begin{cases} x_1, & 0 \leq u < p_1 \\ x_2, & p_1 \leq u < p_1 + p_2 \\ & \vdots \\ x_k, & p_1 + \cdots + p_{k-1} \leq u \end{cases}$$

K-Nearest Neighbor Classifier

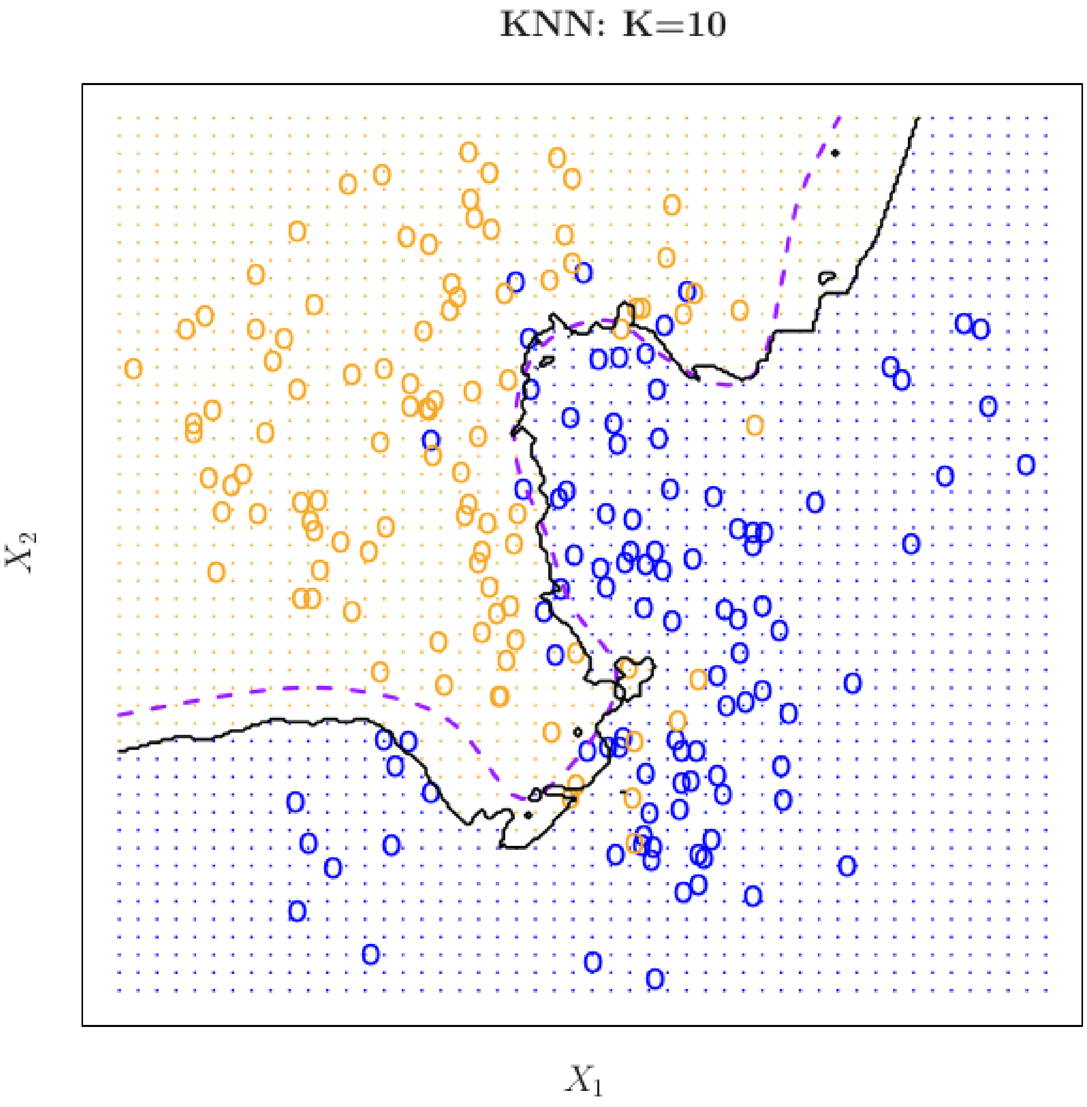


$$\Pr(Y = j | X = \boldsymbol{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

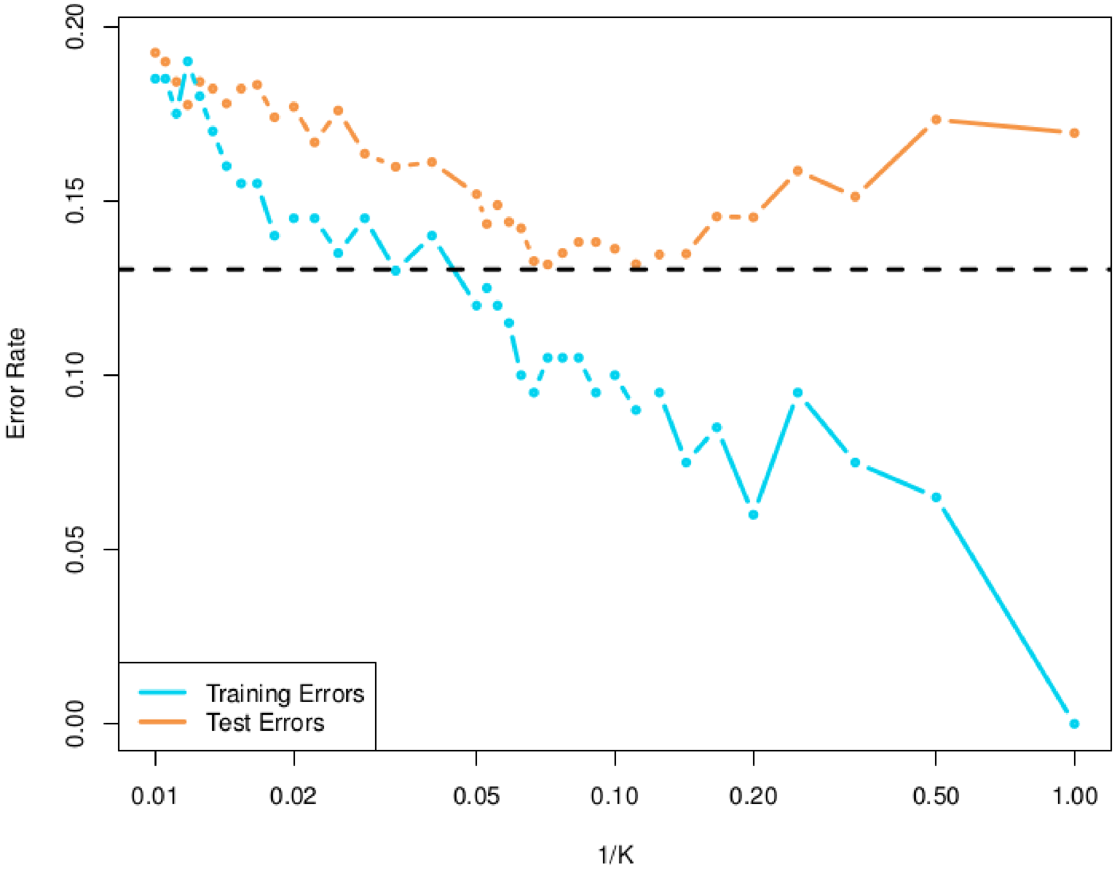
- Classifies $\boldsymbol{x}_0$ to the class with the highest probability

Bayes vs KNN

KNN: K=10



KNN: Error Rate vs 1/K



That was