Erasmus School of Economics

Machine Learning

FEM31002

Introduction

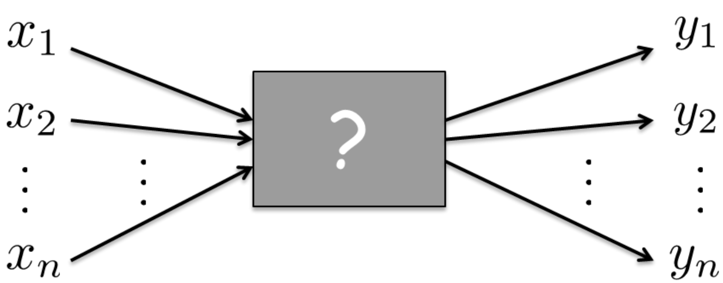
Part 2

Ilker Birbil

birbil@ese.eur.nl

ERASMUS UNIVERSITEIT ROTTERDAM

$$Y = f(X) + \epsilon \xrightarrow{\text{approximation?}} \hat{Y} = \hat{f}(X)$$



training data

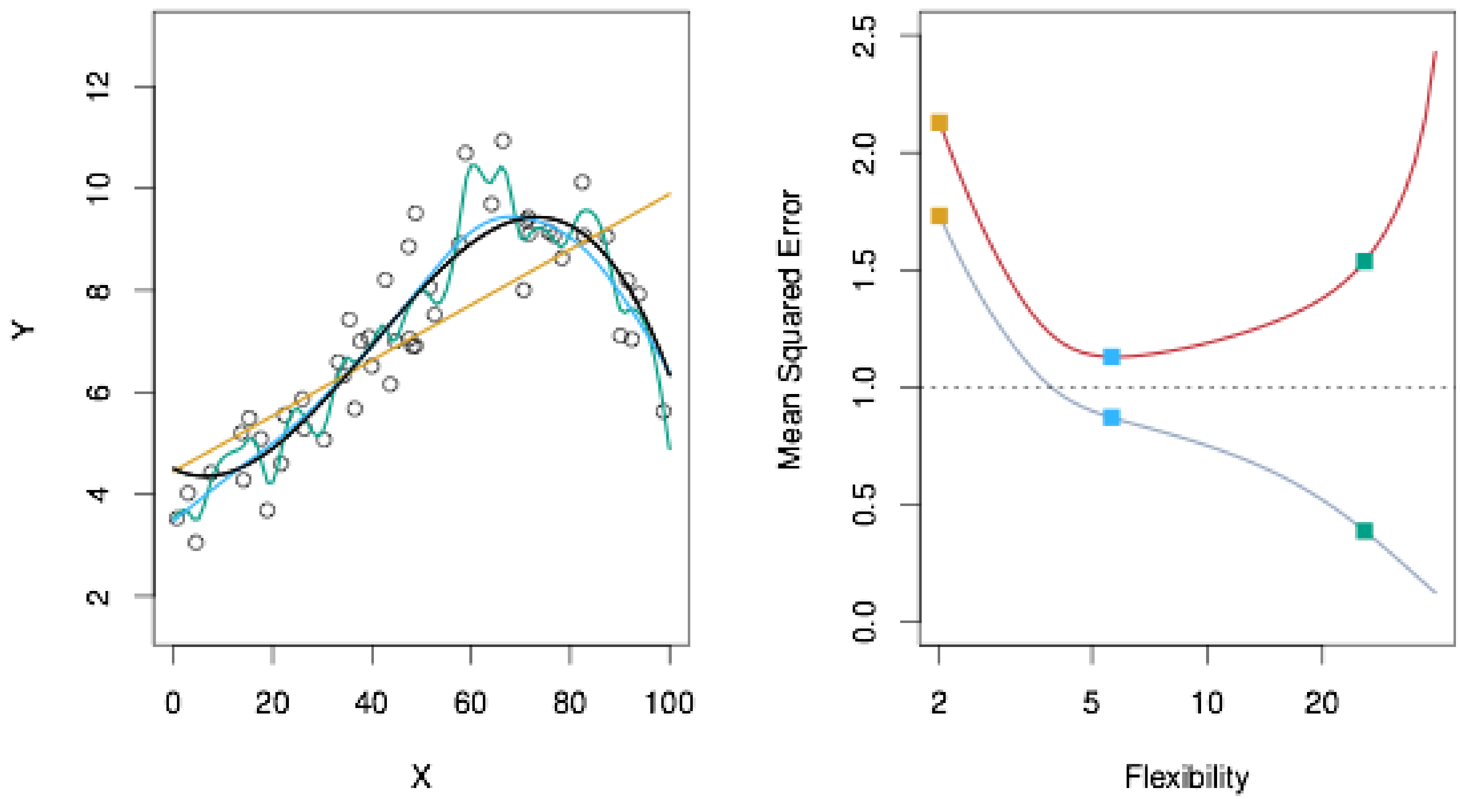
$$\{(x_i, y_i) : 1, \ldots, n\}$$

Training Mean Square Error (MSE)

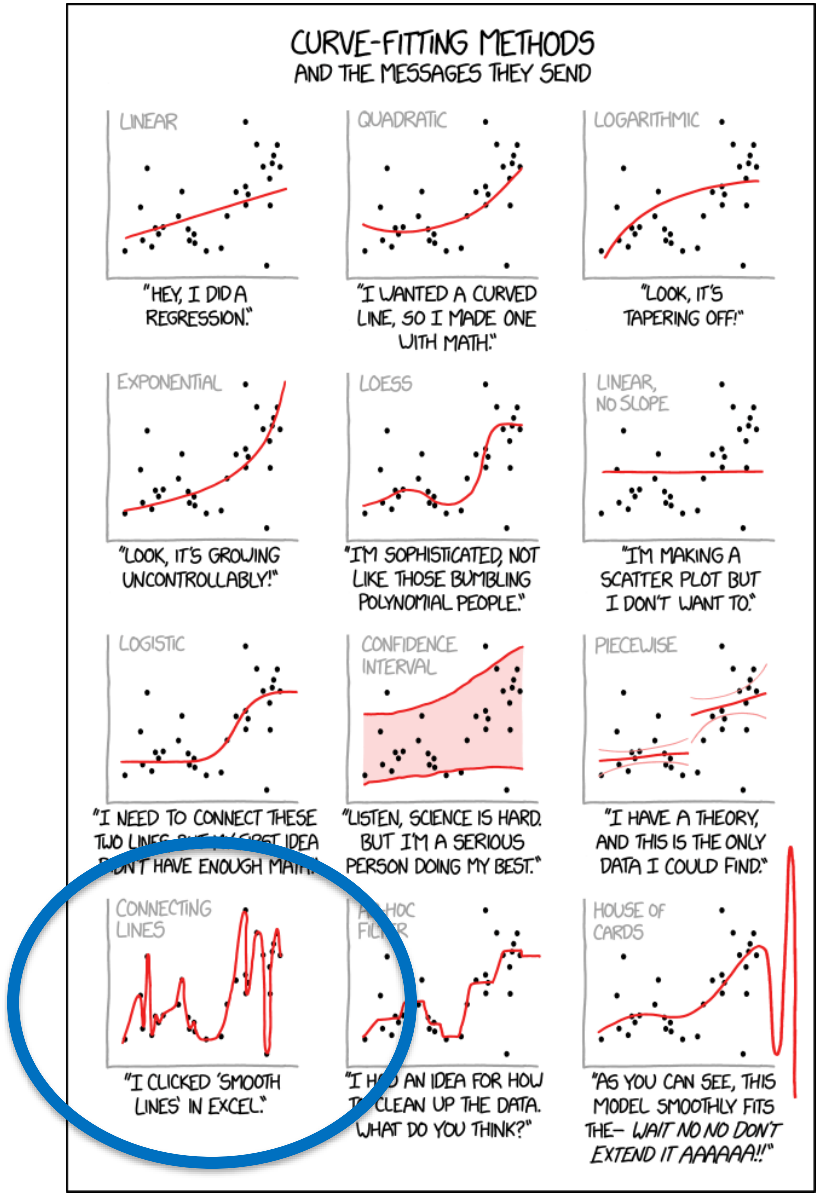
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

(x_0, y_0) : test data ???

$$\mathbb{E}(Y-\hat{Y})^2 = \left(f(X)-\hat{f}(X)\right)^2 + \mathrm{Var}(\epsilon)$$



overfitting


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

(source)

Expected **test** MSE:

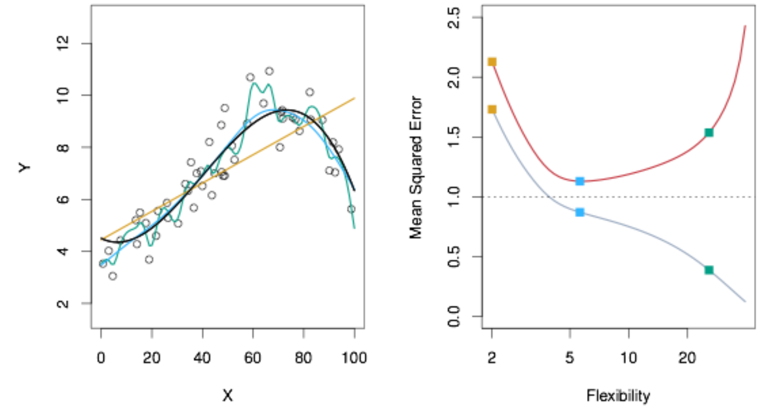
$$\mathbb{E}(y_0 - \hat{f}(x_0))^2 = \mathbb{E}\left(\left(f(x_0) + \epsilon - \hat{f}(x_0)\right)^2\right)$$

$$= \mathbb{E}\left((f(x_0))^2 + 2\epsilon f(x_0) + \epsilon^2 - 2(f(x_0) + \epsilon)\hat{f}(x_0) + (\hat{f}(x_0))^2\right)$$

$$= \mathbb{E}\big((f(x_0))^2\big) + 2\mathbb{E}\big(\epsilon f(x_0)\big) - 2\mathbb{E}\big(f(x_0)\hat{f}(x_0)\big) + \mathbb{E}\big((\hat{f}(x_0))^2\big) + \mathrm{Var}(\epsilon)$$

$$= (f(x_0))^2 - 2f(x_0)\mathbb{E}\big(\hat{f}(x_0)\big) + \mathbb{E}\big((\hat{f}(x_0))^2\big) + \mathrm{Var}(\epsilon) + \mathbb{E}\big(\hat{f}(x_0)\big)^2 - \mathbb{E}\big(\hat{f}(x_0)\big)^2$$

$$= \underbrace{\mathbb{E}\big((\hat{f}(x_0))^2\big) - \mathbb{E}\big(\hat{f}(x_0)\big)^2}_{\text{variance}} + \underbrace{\big(\mathbb{E}(\hat{f}(x_0)) - f(x_0)\big)^2}_{\text{bias}} + \mathrm{Var}(\epsilon)$$



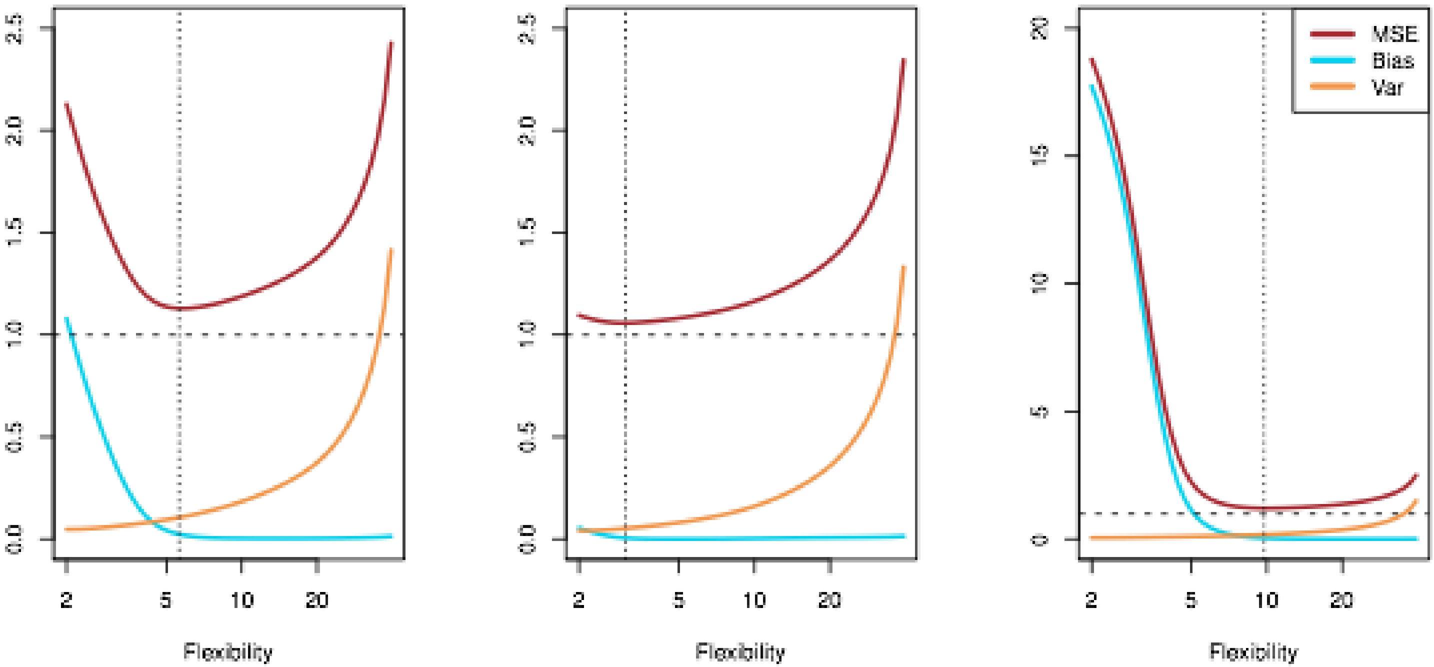
Expected **test** MSE:

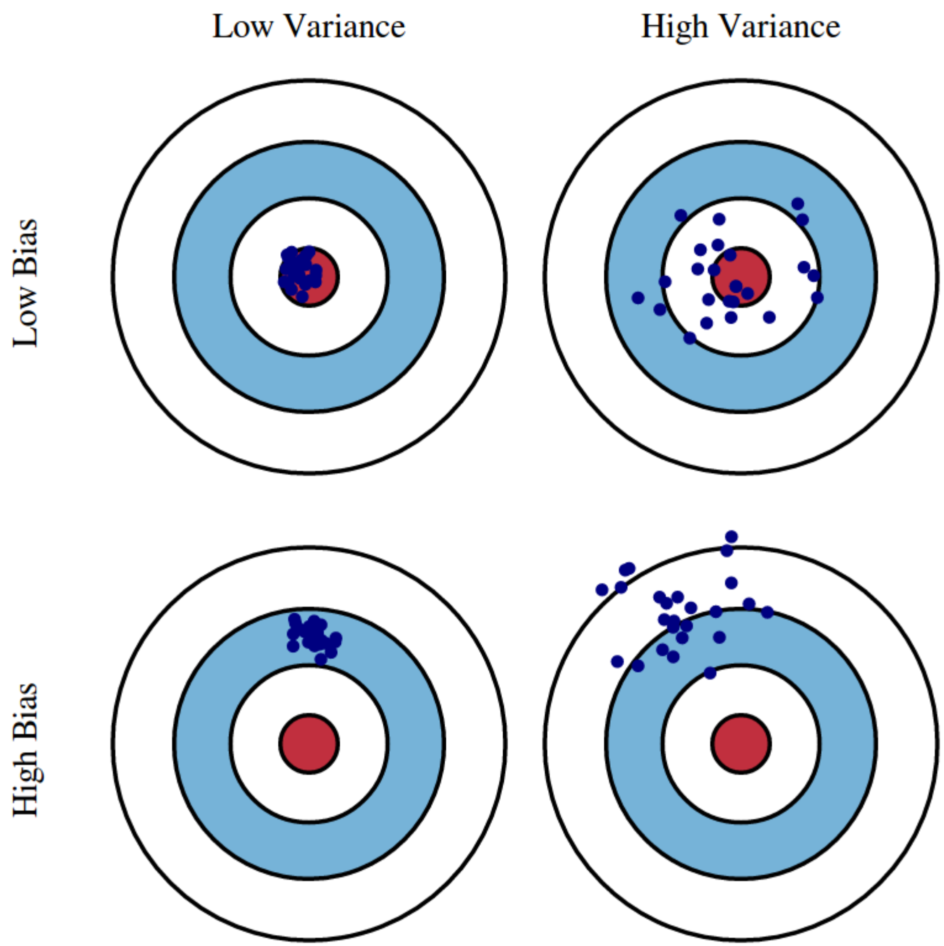
$$
\begin{aligned}
\mathbb{E}(y_0 - \hat{f}(x_0))^2 \quad &= \dots \\
&= \mathbb{E}\big((\hat{f}(x_0))^2\big) - \mathbb{E}\big(\hat{f}(x_0)\big)^2 + \big(\mathbb{E}(\hat{f}(x_0)) - f(x_0)\big)^2 + \mathrm{Var}(\epsilon) \\
&= \mathrm{Var}(\hat{f}(x_0)) + \big(\mathrm{Bias}(\hat{f}(x_0))\big)^2 + \mathrm{Var}(\epsilon)
\end{aligned}
$$

Roughly:

Model Complexity ↑	Variance ↑	Bias ↓

$$\mathbb{E}(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\hat{f}(x_0)) + \left(\mathrm{Bias}(\hat{f}(x_0))\right)^2 + \mathrm{Var}(\epsilon)$$





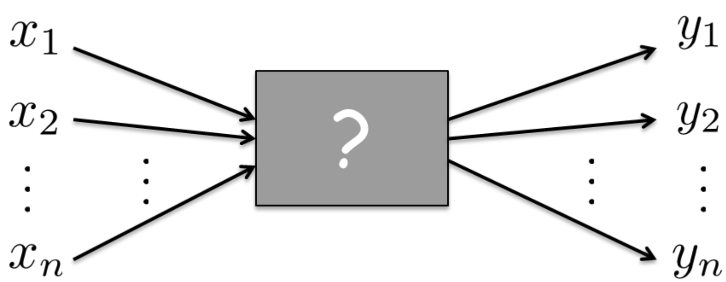
(source)

CLASSIFICATION

input variables

output variables (categorical)

$X = (X_1, X_2, \ldots, X_p)$



training data

$$\{(x_i, y_i) : 1, \ldots, n\}$$

Training Error Rate

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

(x_0, y_0) : test data ???

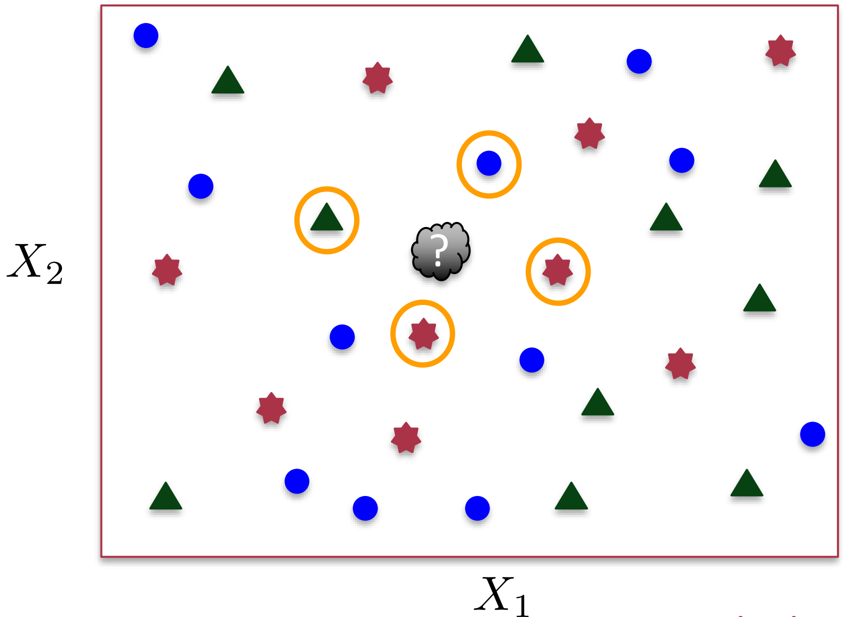
Confusion Matrix

		Predicted Class		
		-	**+**	**Total**
True Class	**-**	True Negative (TN)	False Positive (FP)	N
	+	False Negative (FN)	True Positive (TP)	P
	Total	N*	P*	

		Predicted Class		
		-	+	**Total**
True Class	-	TN	FP	N
	+	FN	TP	P
	Total	N*	P*	

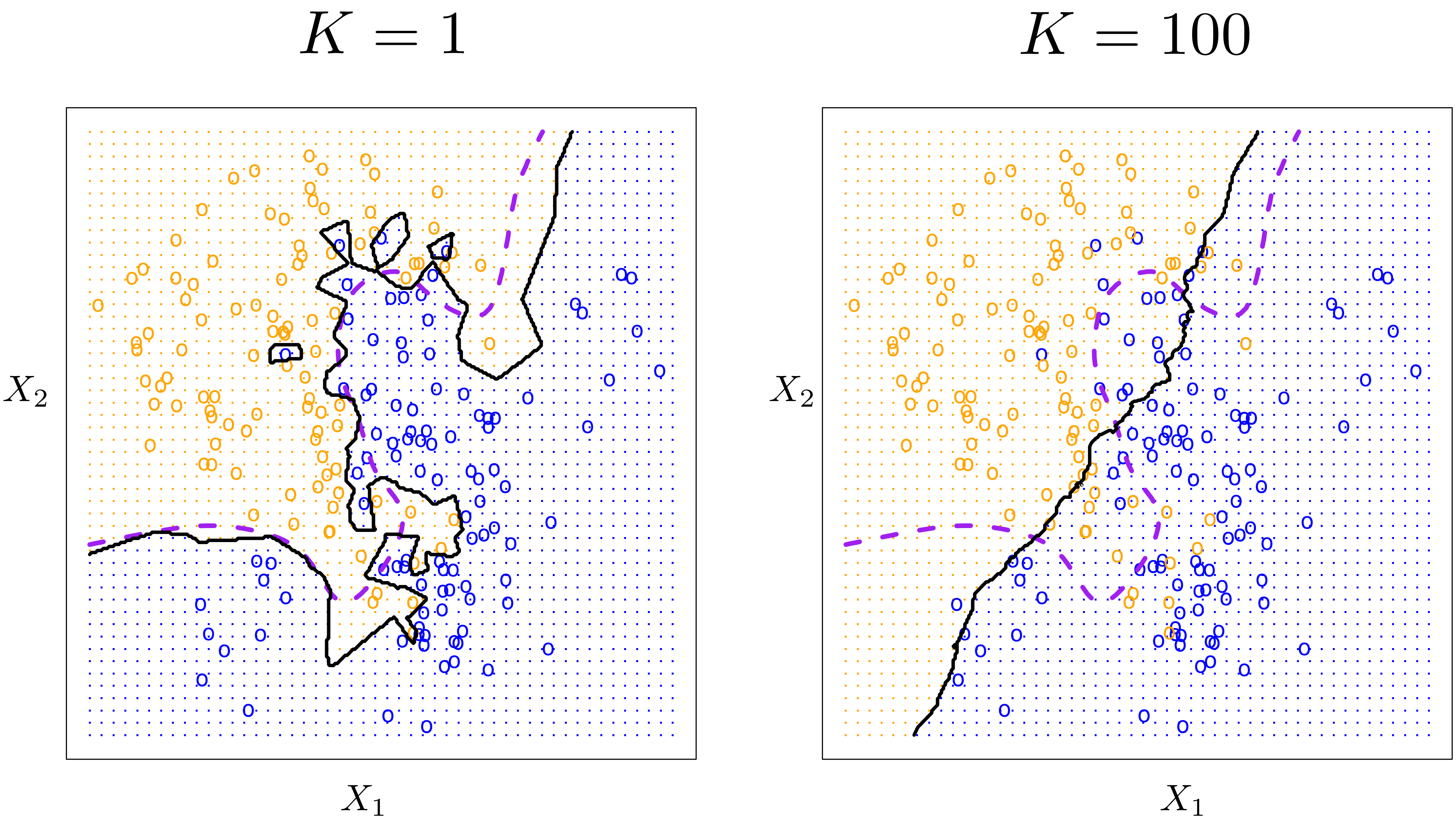
Name	**Defn.**	**Synonyms**
False Positive Rate	FP/N	type I error, (1-specificity)
True Positive Rate	TP/P	(1-type II error), power, sensitivity, recall
Positive Pred. Value	TP/P*	precision, (1-false discovery proportion)
Negative Pred. Value	TN/N*	

K – Nearest Neighbors



$$\mathbb{P}(Y=j\,|\,X=x_0)=\frac{1}{K}\overbrace{\sum_{i\in\mathcal{N}_0} I(y_i=j)}^{\text{majority voting}}$$

set of K
closest points to x_0



Voronoi Tessellation

Regression

X, Y random: $f(X) = ?$

$$\mathbb{E}((Y - f(X))^2) = \mathbb{E}(\mathbb{E}((Y - f(X))^2 \mid X))$$

$$f(x) = \arg\min_u \underbrace{\mathbb{E}((Y - u)^2 \mid X = x)}_{h(u)}$$

$$\frac{\partial h(u)}{\partial u} = 2u - 2\mathbb{E}(Y|X = x) = 0 \implies f(x) = \mathbb{E}(Y|X = x)$$

$$\boxed{f(x_0) = \mathbb{E}(Y|X = x_0)}$$

$$\hat{f}(x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} y_i$$

K-NN approximation to regression

Classification

X, Y random: $\hat{Y}(X) = ?$

$$\mathbb{E}(I(Y \neq \hat{Y}(X))) = \mathbb{E}(\mathbb{E}(I(Y \neq \hat{Y}(X)) \mid X))$$

$$\begin{aligned}
\hat{Y}(x) &= \arg\min_j \; \mathbb{E}(I(Y \neq j) \mid X = x) \\
&= \arg\min_j \; \sum_k I(k \neq j)\mathbb{P}(Y = k | X = x) \\
&= \arg\min_j \; (1 - \mathbb{P}(Y = j | X = x)) \\
&= \arg\max_j \; \mathbb{P}(Y = j | X = x)
\end{aligned}$$

Bayes Classifier


(source)

$$\hat{y}_0 = \arg\max_j \ \mathbb{P}(Y = j | X = x_0)$$

Bayes Error Rate

$$\mathbb{E}\Big(1 - \max_j \ \mathbb{P}(Y = j|X)\Big) = 1 - \mathbb{E}\Big(\max_j \ \mathbb{P}(Y = j|X)\Big)$$