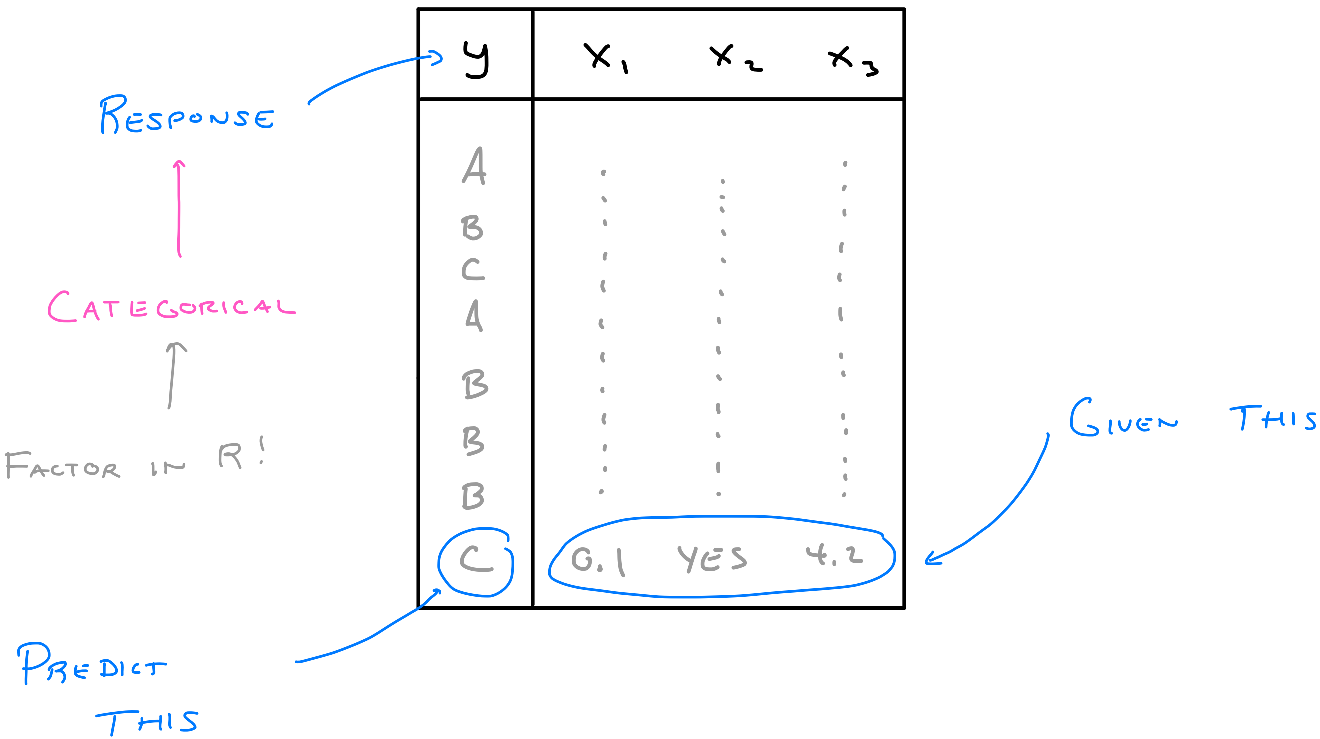
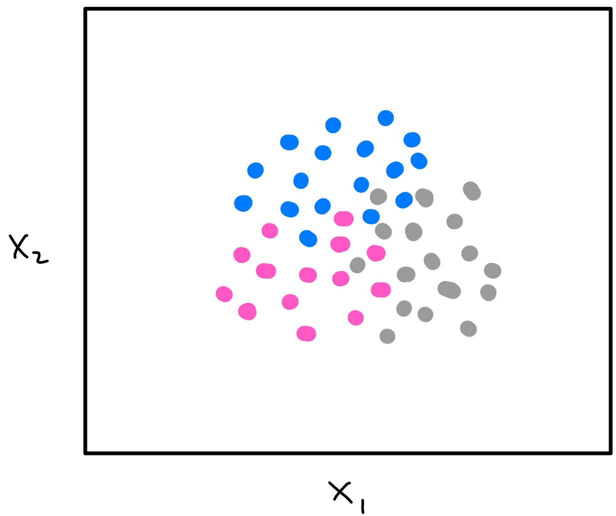
CLASSIFICATION

↳ AN INTRODUCTION

Data View


Response
y
x_1
x_2
x_3
Categorical
Factor in R!
A
B
C
A
B
B
B
C
0.1
YES
4.2
Given this
Predict this

VISUAL DATA VIEW



PROBABILITY VIEW

P FEATURES   K CATEGORIES

$$(X, Y) \in \mathbb{R}^P \times \{1, 2, \ldots K\}$$

FEATURES   RESPONSE

FIND A CLASSIFIER $C(x)$ THAT MINIMIZES

$$P[C(x) \neq Y]$$

PROBABILITY OF MISCLASSIFICATION

WHERE $C : \mathbb{R}^P \longrightarrow \{1, 2, 3, \ldots K\}$

INPUT FEATURES   OUTPUT CATEGORY

BAYES CLASSIFIER

← MINIMIZES PROBABILITY OF MISCLASSIFICATION

$$C^{B}(x) \triangleq \underset{k \in \{1,\ldots,K\}}{\mathrm{ARGMAX}}\; P[Y = k \mid X = x]$$

GIVEN FEATURE VECTOR x, CLASSIFY OBSERVATION AS THE CATEGORY WITH THE HIGHEST PROBABILITY

DUH?

Example

		$X=0$	$X=1$	
Y	A	0.1	0.1	0.2
	B	0.2	0.1	0.3
	C	0.1	0.4	0.5
		0.4	0.6	

Joint distribution of (X,Y)

Marginal distribution of X

$$C^B(x=0) = ?$$

$$\frac{P[X=0 \cap Y=A]}{P[X=0]}$$

$$P[Y \mid X=0] = \begin{cases} 0.25 & y=A \\ 0.50 & y=B \\ 0.25 & y=C \end{cases}$$

Conditional probability of $Y \mid X=0$

$$C^B(x=0) = B$$

$$C^B(x=1) = C$$

BAYES ERROR

← AVERAGE MISSCLASSIFICATION USING BAYES CLASSIFIER

$$1 - E_x\left[\max_k P[Y=k \mid X=x]\right]$$

"IRREDUCIBLE ERROR"

		X: 0	X: 1	
Y	A	0.1	0.1	0.2
	B	0.2	0.1	0.3
	C	0.1	0.4	0.5
		0.4	0.6	

$$= 1 - \left[P[Y=B \mid X=0]\,P[X=0] + P[Y=C \mid X=1]\,P[X=1]\right]$$

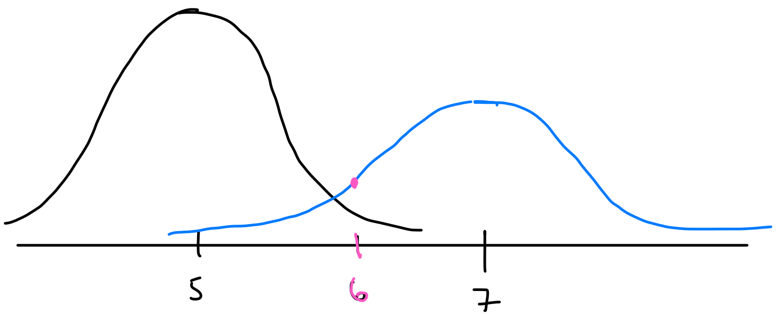
$$= 1 - \left[\left(\frac{0.2}{0.4}\right)(0.4) + \left(\frac{0.4}{0.6}\right)(0.6)\right]$$

$$= 1 - [0.2 + 0.4] = 0.4$$

Example

$$X \mid Y=0 \sim N(\mu=5, \sigma=1) \qquad f_0(x)$$

$$X \mid Y=1 \sim N(\mu=7, \sigma=2) \qquad f_1(x)$$



$$\pi_0 = P[Y=0] = 0.6$$

$$\pi_1 = P[Y=1] = 0.4$$

$$C^B(x=6) = ?$$

Calculate $\quad P[Y=0 \mid X=6] = \dfrac{\pi_0 f_0(6)}{\pi_0 f_0(6) + \pi_1 f_1(6)} = \ldots$

In practice

Don't know $P[Y=k \mid X=x]$!!!

Estimate it !!!

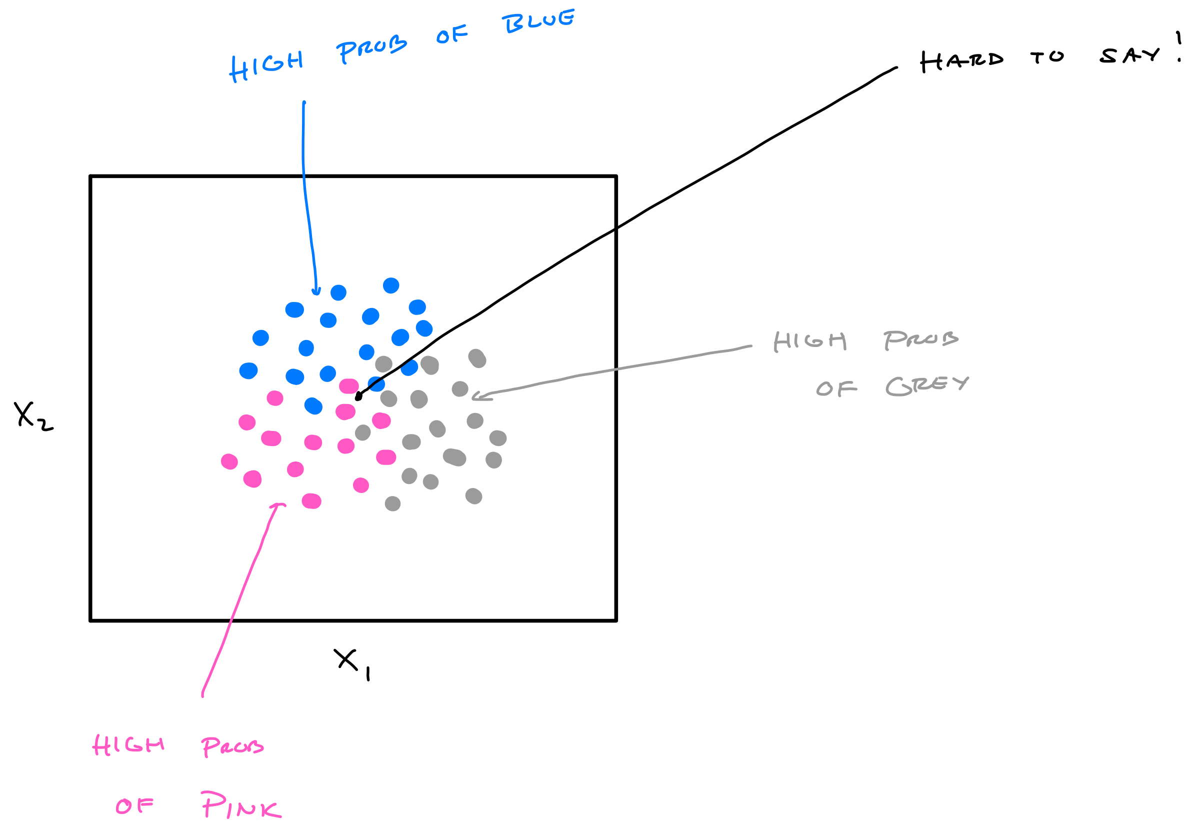
Learned classifier

$$\hat{C}(x) = \underset{k}{\operatorname{argmax}} \underbrace{\hat{P}[Y=k \mid X=x]}_{\text{Estimate of conditional probability}}$$

A "guess" for

$C^B(x)$

How?


HIGH PROB OF BLUE
HARD TO SAY!
HIGH PROB OF GREY
X_2
X_1
HIGH PROB OF PINK

Estimating Conditional Probabilities

KNN	w/	caret::knn3	predict(mod, data, type)
Trees	w/	rpart::rpart	
Linear models	w/	glm nnet::nnet	

Ensure response is a factor

METRICS

WOULD LIKE $P\left[C(x) \neq Y\right]$

SETTLE FOR $\frac{1}{n}\sum_{i=1}^{n} I\left(C(x_i) \neq y_i\right)$

MISCLASSIFICATION

$$I\left(C(x_i) \neq y_i\right) = \begin{cases} 1 & C(x_i) \neq y_i \\ 0 & C(x_i) = y_i \end{cases}$$

$\frac{1}{n}\sum_{i=1}^{n} I\left(C(x_i) = y_i\right)$ ← ACCURACY

$C(x)$ PLACEHOLDER

$\hat{C}(x)$ LEARNED

$C^B(x)$ BAYES

BINARY CLASSIFICATION

$Y = 0$ or $Y = 1$

"NEGATIVE" "POSITIVE"

$$p(x) \triangleq P[Y = 1 \mid X = x]$$

$$1 - p(x) = P[Y = 0 \mid X = x]$$

$$C^B(x) = \begin{cases} 1 & p(x) \geq 0.5 \\ 0 & \text{ELSE} \end{cases}$$

METRICS

FP/TP

FN/TN

ETC