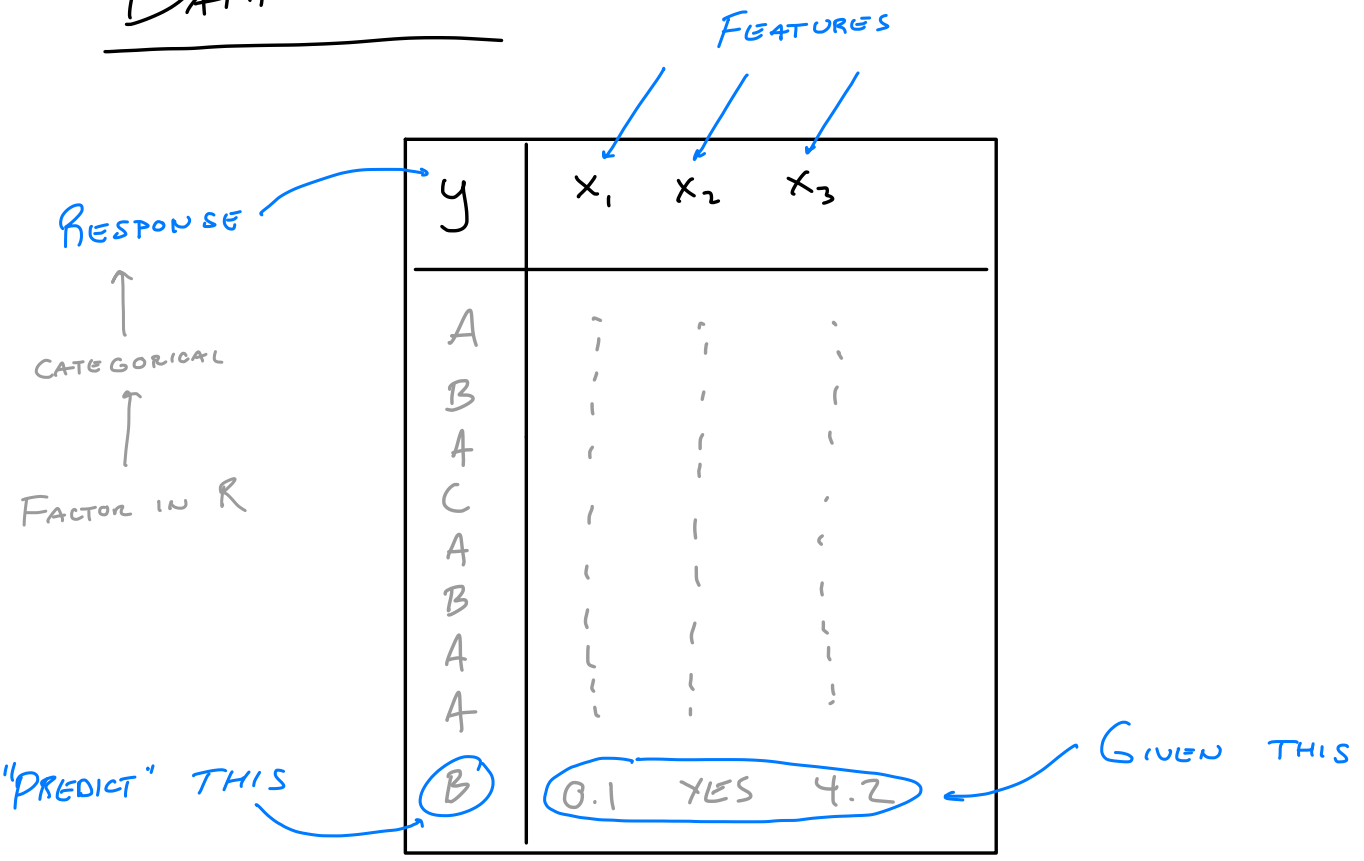
CLASSIFICATION

↳ AN INTRODUCTION

DALPIAZ

02/17/2020

Data View


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

Probability View

P features

K categories

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

X: features

Y: response

Find a classifier $C(x)$ that minimizes

x: features

$C(x)$ output: one of the K categories

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

→ misclassification rate

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

Bayes Classifier

← minimizes misclassification

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

Given feature vector x, classify observation as the category with the highest probability.

Duh?

Example

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

Y \ X	0	1	
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

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

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

Conditional Distribution of $Y \mid X=0$

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

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

Bayes Error

← Average misclassification using Bayes classifier

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

("irreducible error")

Y \ X	0	1	
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	

$$= 1 - \left[\max_k P[Y = k \mid X = 0] P[X = 0] + \max_k P[Y = k \mid X = 1] P[X = 1] \right]$$

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

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

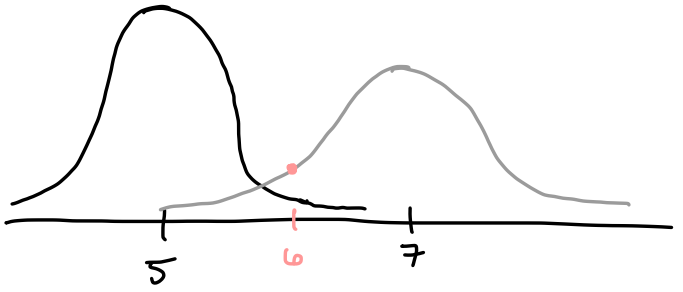
Example

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

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

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

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



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

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

Binary Classification

$Y = 0$ or $Y = 1$

"negative" "positive"

Later

FP / TP

FN / TN

ETC

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

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

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

In Practice

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

Estimate it!

Classifier

$$C(x) = \underset{k}{\text{argmax}}\ \hat{P}[Y = k \mid X = x]$$

$\hat{P}[Y = k \mid X = x]$: Estimate of conditional distribution

A "guess" for

$C^B(x)$

How?

ESTIMATING CONDITIONAL DISTRIBUTIONS

KNN w/ caret::knn3()

TREES w/ rpart::rpart()

LINEAR MODELS w/ glm() ? nnet::nnet() ?

} predict(mod, data, type)

MAKE SURE RESPONSE VARIABLE IS A FACTOR!