

CS 307

SPRING 2024

DALPIAZ

CLASSIFICATION

WITH KNN

CLASSIFICATION

↳ AN INTRODUCTION

DATA VIEW

TARGET
↑
CATEGORICAL

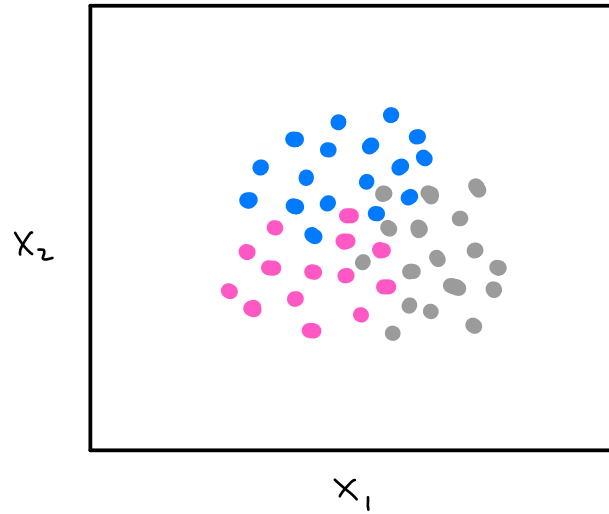
FEATURES

| y | x ₁ | x ₂ | x ₃ |
|---|----------------|----------------|----------------|
| A | ⋮ | ⋮ | ⋮ |
| B | ⋮ | ⋮ | ⋮ |
| C | ⋮ | ⋮ | ⋮ |
| A | ⋮ | ⋮ | ⋮ |
| B | ⋮ | ⋮ | ⋮ |
| B | ⋮ | ⋮ | ⋮ |
| B | ⋮ | ⋮ | ⋮ |
| C | 0.1 | YES | 4.2 |

GIVEN THIS

PREDICT
THIS

VISUAL DATA VIEW



PROBABILITY VIEW

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

Diagram illustrating the probability view of a classification problem. The input (X, Y) is shown as a pair of variables. X is labeled "FEATURES" and is associated with \mathbb{R}^P , which is labeled "P FEATURES". Y is labeled "RESPONSE" and is associated with the set $\{1, 2, \dots, K\}$, which is labeled "K CATEGORIES".

FIND A CLASSIFIER $C(x)$ THAT MINIMIZES

$$P[C(x) \neq Y] \leftarrow \text{PROBABILITY OF MISCLASSIFICATION}$$

WHERE

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

Diagram illustrating the classifier function C . The input is \mathbb{R}^P , labeled "INPUT FEATURES". The output is $\{1, 2, 3, \dots, K\}$, labeled "OUTPUT CATEGORY".

BAYES CLASSIFIER

← MINIMIZES PROBABILITY
OF MISCLASSIFICATION

$$C^B(x) \triangleq \underset{k \in \{1, \dots, K\}}{\operatorname{ARGMAX}} P[Y = k | X = x]$$

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

DUH?

EXAMPLE

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

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

Joint Distribution of (X, Y)

| | X | | | |
|---|---|-----|-----|-----|
| | 0 | 1 | | |
| Y | A | 0.1 | 0.1 | 0.2 |
| | B | 0.2 | 0.1 | 0.3 |
| | C | 0.1 | 0.4 | 0.5 |

Marginal Distribution of X

0.4 0.6

Conditional Probability of Y | X=0

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

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

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

BAYES ERROR

← AVERAGE MISCLASSIFICATION
USING BAYES CLASSIFIER

$$1 - E_x \left[\max_k P[Y=k | X=x] \right] \quad \text{"IRREDUCIBLE ERROR"}$$

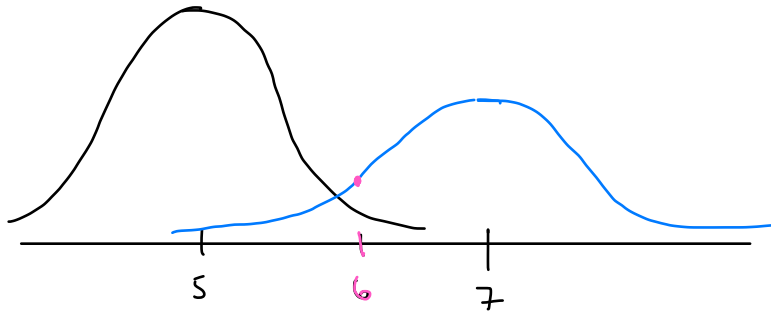
| | X | | |
|---|-------|-------|-----|
| | 0 | 1 | |
| A | 0.1 ✓ | 0.1 ✓ | 0.2 |
| B | 0.2 ✗ | 0.1 ✓ | 0.3 |
| C | 0.1 ✓ | 0.4 ✗ | 0.5 |
| | 0.4 | 0.6 | |

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

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

$$= 1 - [0.2 + 0.4] = \underline{0.4}$$

EXAMPLE



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

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

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

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

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

CALCULATE

$$P[Y=0 | X=6] = \frac{\pi_0 f_0(6)}{\pi_0 f_0(6) + \pi_1 f_1(6)} = \dots \text{ TO SCIPY!}$$

$$P[Y=0 | X=b] = \frac{\pi_0 f_0(b)}{\pi_0 f_0(b) + \pi_1 f_1(b)}$$

$$P[Y=1 | X=b] = \frac{\pi_1 f_1(b)}{\pi_0 f_0(b) + \pi_1 f_1(b)}$$

SAME DENOMINATOR

ONLY NEED NUMERATOR
FOR CLASSIFICATION

IN PRACTICE

DON'T KNOW $P[Y = k | X = x]$!!!
...

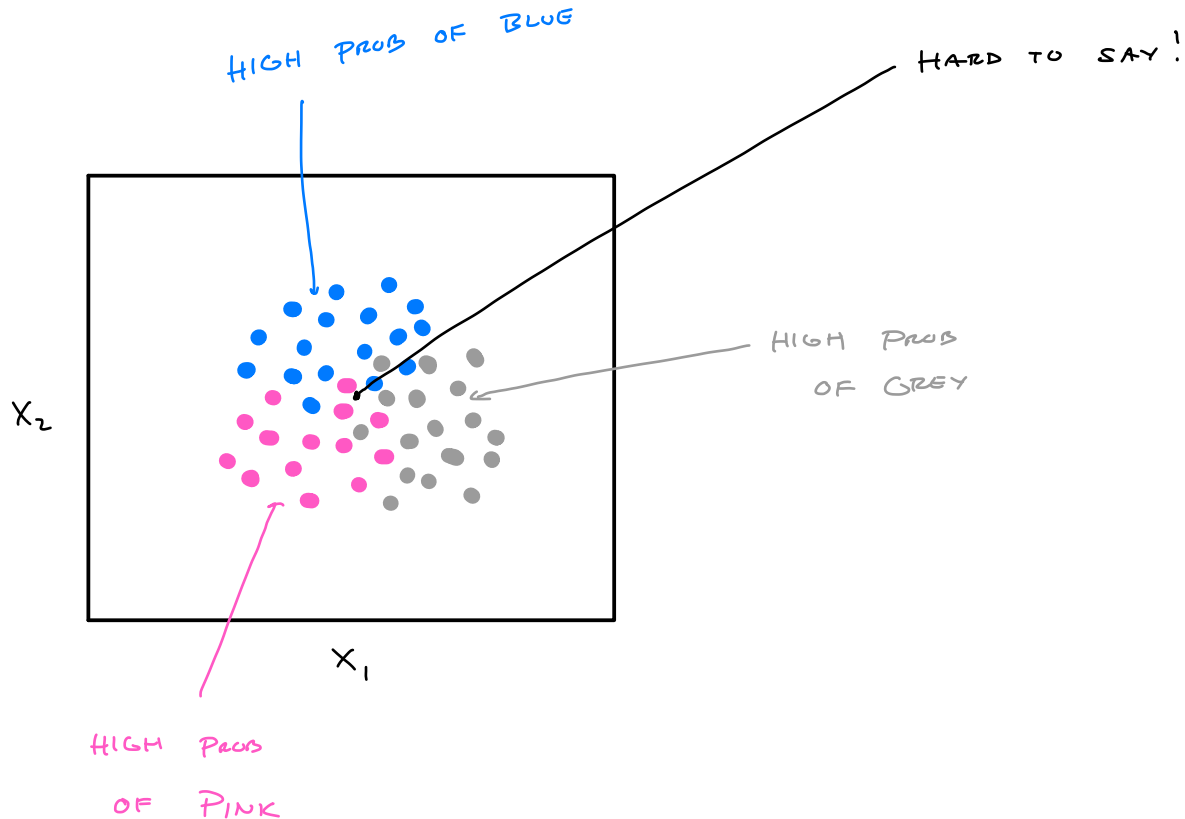
ESTIMATE IT !!!

LEARNED
CLASSIFIER

$$\hat{C}(x) = \underset{k}{\operatorname{ARGMAX}} \underbrace{\hat{P}[Y = k | X = x]}_{\text{ESTIMATE OF CONDITIONAL PROBABILITY}}$$

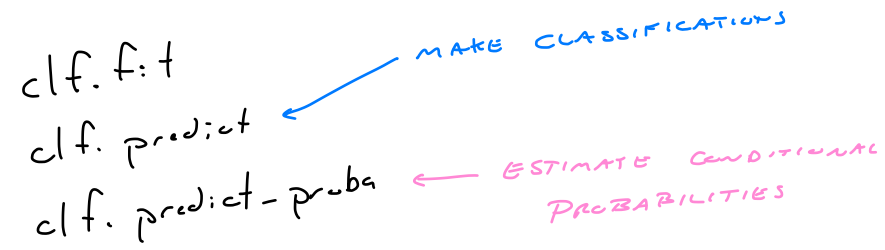
A "GUESS"
FOR $C^B(x)$

How?



ESTIMATING CONDITIONAL PROBABILITIES

| | | | |
|---------------|----|---|-----------------------|
| KNN | w/ | sklearn.neighbors.KNeighbors <u>Classifier</u> | } FAMILIAR INTERFACES |
| TREES | w/ | sklearn.tree.DecisionTree <u>Classifier</u> | |
| LINEAR MODELS | w/ | sklearn.linear_model.Logistic Regression <i>(est. prob.)</i> | |



METRICS

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

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

MISCLASSIFICATION

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

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

ACCURACY

$C(x)$ PLACEHOLDER

$\hat{C}(x)$ LEARNED

$C^b(x)$ BAYES

KNN CLASSIFICATION

- MOSTLY THE SAME AS REGRESSION
- FOCUS ON ESTIMATING

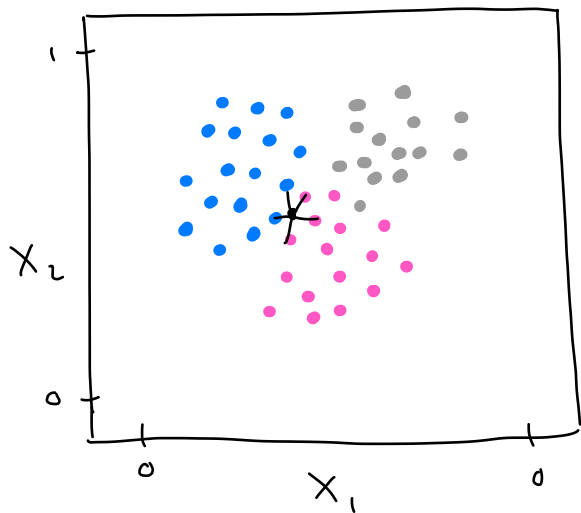
$$P[Y=g | X=x]$$

K_{NN} CLASSIFICATION

How TO ESTIMATE?

$$\hat{P}[Y=g | X=x] = \frac{1}{K} \sum_{\{i: x_i \in \mathcal{N}_K(x, D)\}} I(y_i = g)$$

$$K=5$$

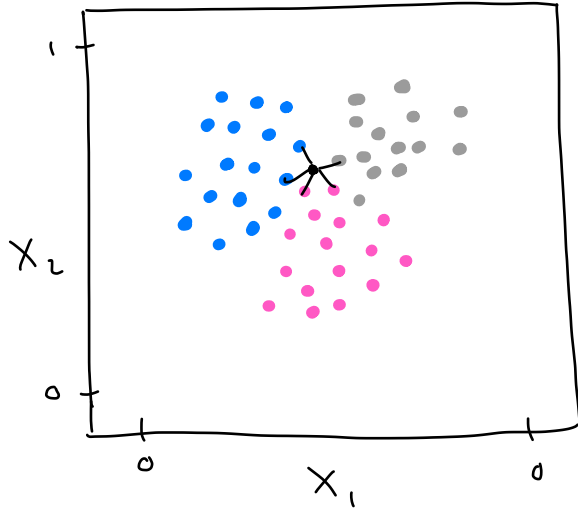


$$\hat{P}[Y = \bullet \mid X = \bullet] = 0/5$$

$$\hat{P}[Y = \bullet \mid X = \bullet] = 2/5$$

$$\hat{P}[Y = \bullet \mid X = \bullet] = 3/5$$

$k=5$



$$\hat{P}[Y = \bullet \mid X = \bullet] = 1/5$$

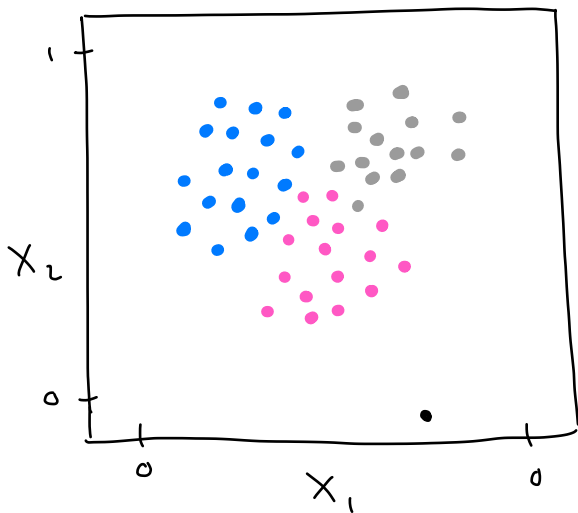
$$\hat{P}[Y = \bullet \mid X = \bullet] = 2/5$$

$$\hat{P}[Y = \bullet \mid X = \bullet] = 2/5$$

How to CLASSIFY?

↳ DEPENDS WHO WROTE CODE!

$$K=5$$



$$\hat{P}[Y = \bullet \mid X = \bullet] = 0/5$$

$$\hat{P}[Y = \bullet \mid X = \bullet] = 0/5$$

$$\hat{P}[Y = \bullet \mid X = \bullet] = 5/5$$

BUT SHOULD YOU CLASSIFY?