EPFL

Lecture 05
14.10.2024

Naive Bayes Classifier

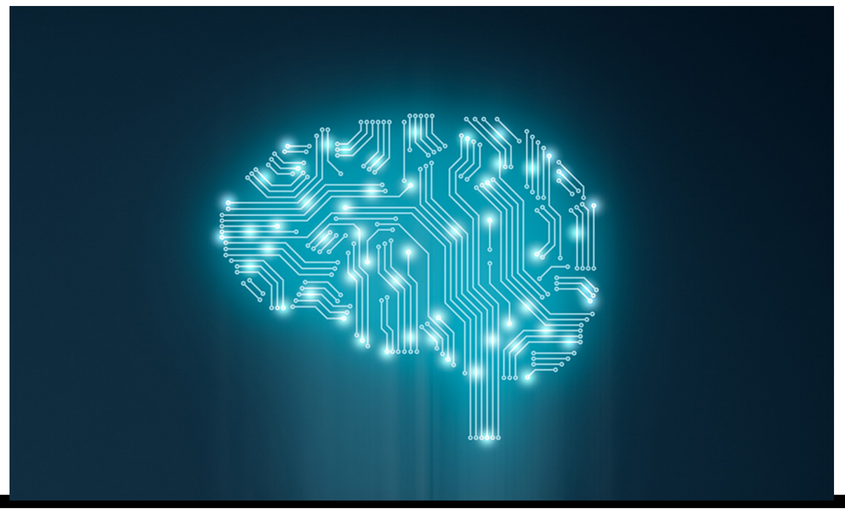
Outline

- Review of probability
  - Conditional distribution
  - Independence
  - Bayes rule

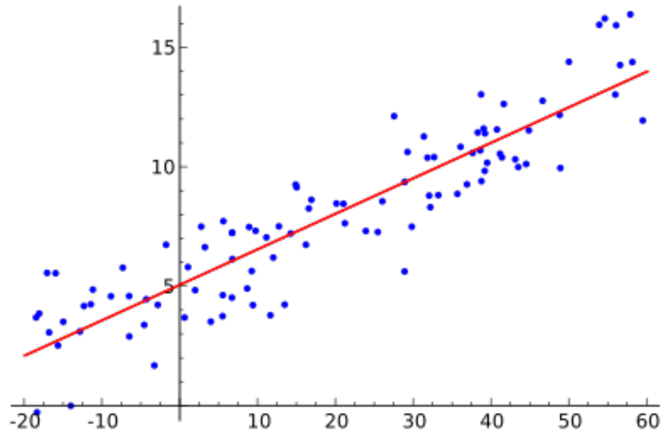
- Naive Bayes classifier
  - Finite-valued features
  - Continuous-valued features (Gaussian Naive Bayes classifier)

- Announcements:
  - Exercise hours: quiz 1

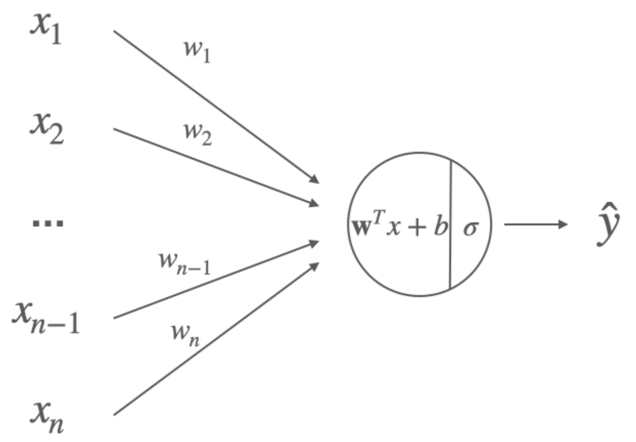
Introduction



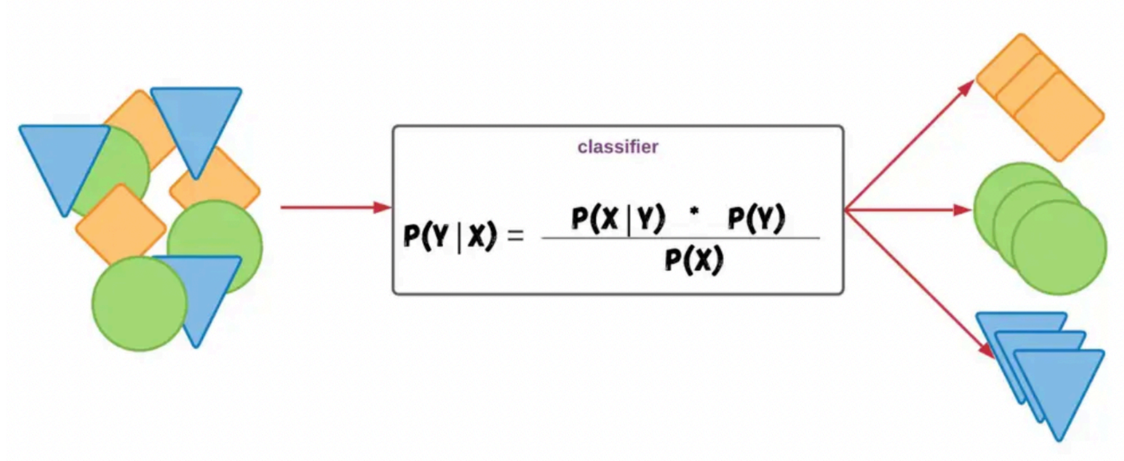
Linear regression



Logistic regression



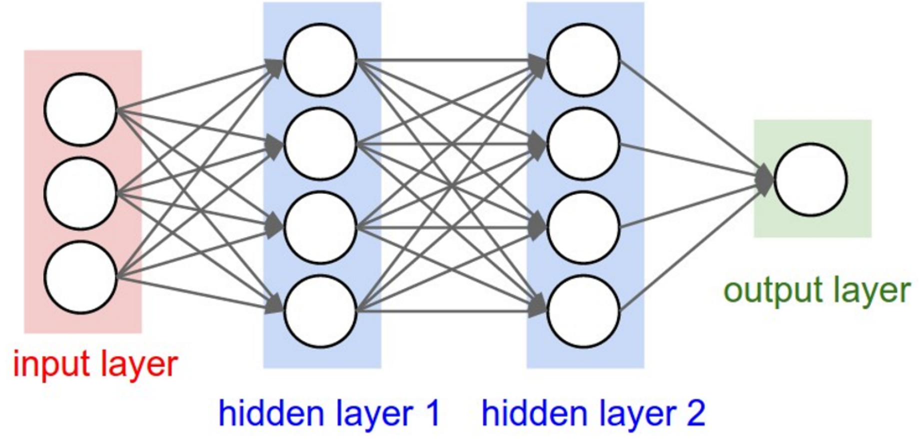
Naive Bayes



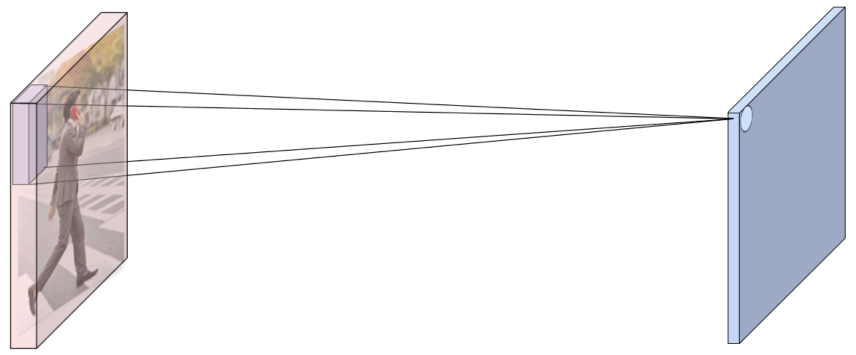
KNN

AI ethics

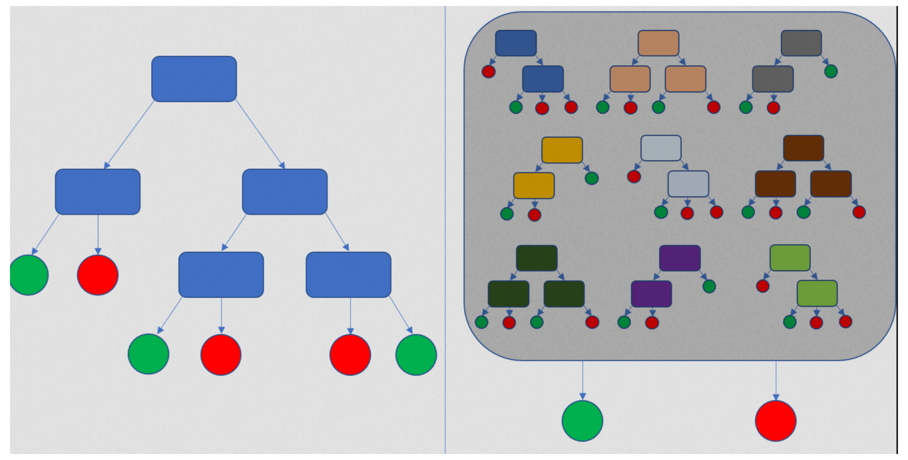
Neural networks



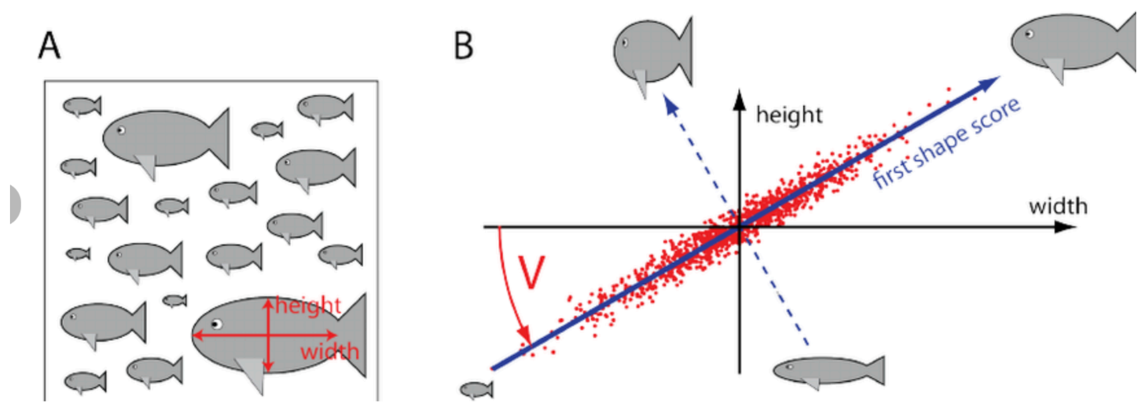
Convolutional neural networks



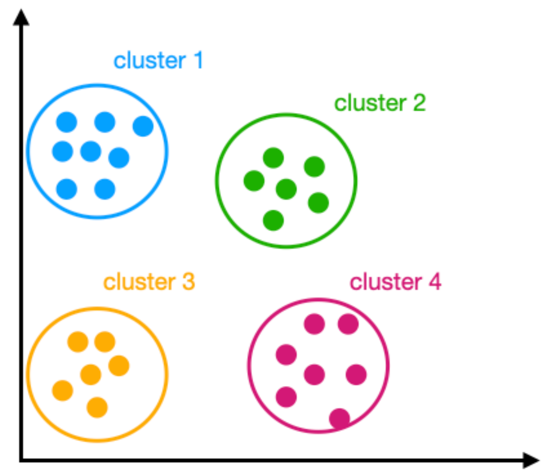
Decision-trees



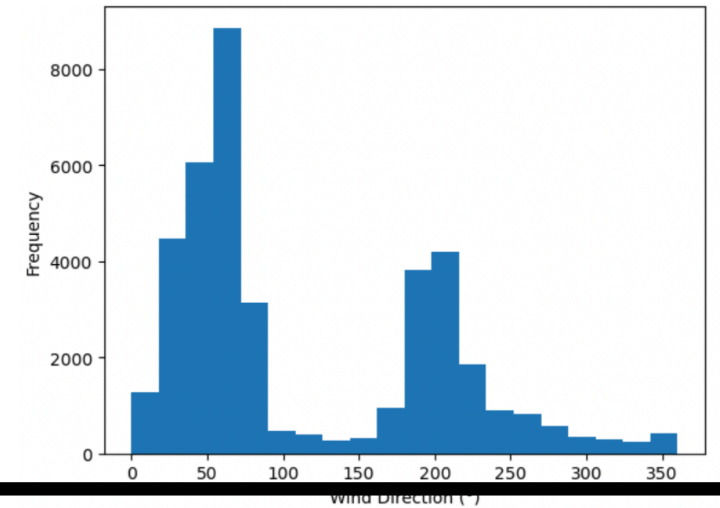
Dimensionality reduction



Clustering



Reinforcement learning



Naive Bayes Classifier

Supervised learning - probabilistic classifier

Recall probabilistic interpretation of logistic regression

$\{x^i, y^i\}_{i=1}^N \quad x^i \in \mathbb{R}^d, \quad y^i \in \{1, 2, \dots, K\}$

a probabilistic classifier gives probability of $x^{test} \in \mathbb{R}^d$ having label $c \in \{1, 2, \dots, K\}$.

example: multinomial logistic regression:

$$p(y=c \mid x) = \frac{e^{z_c}}{\sum_{j=1}^{K} e^{z_j}}, \quad z_c = w_c^T x + b_c$$

$$z_j = w_j^T x + b_j$$

We will now discuss another approach to probabilistic classification

EPFL

Probability review
Probability, random variable

- probability of an event A, $Pr(A)$: how likely it is for event A to happen, ex: event A: die rolls a "6"

- Random variable X: a variable that takes different values with certain probabilities. X_{all}: set of all values X can take,

$$\sum_{x \in X_{all}} \underbrace{Pr(X = x)}_{\text{Prob. of } X \text{ taking value } x} = 1.$$

dice ... $X_{all} = \{1, 2, \dots, 6\}$,

X: the number the die rolls at $P(X=x) = \frac{1}{6}$

dice example

"1"	"2"	"3"
"4"	"5"	"6"

"1"	"2"	"3"
"4"	"5"	"6"

event A

"1"	"2"	"3"
"4"	"5"	"6"

$1 - Pr(A)$

"1"	"2"	"3"
"4"	"5"	"6"

$\sum_{x=1}^{6} Pr(X=x) = 1$

EPFL

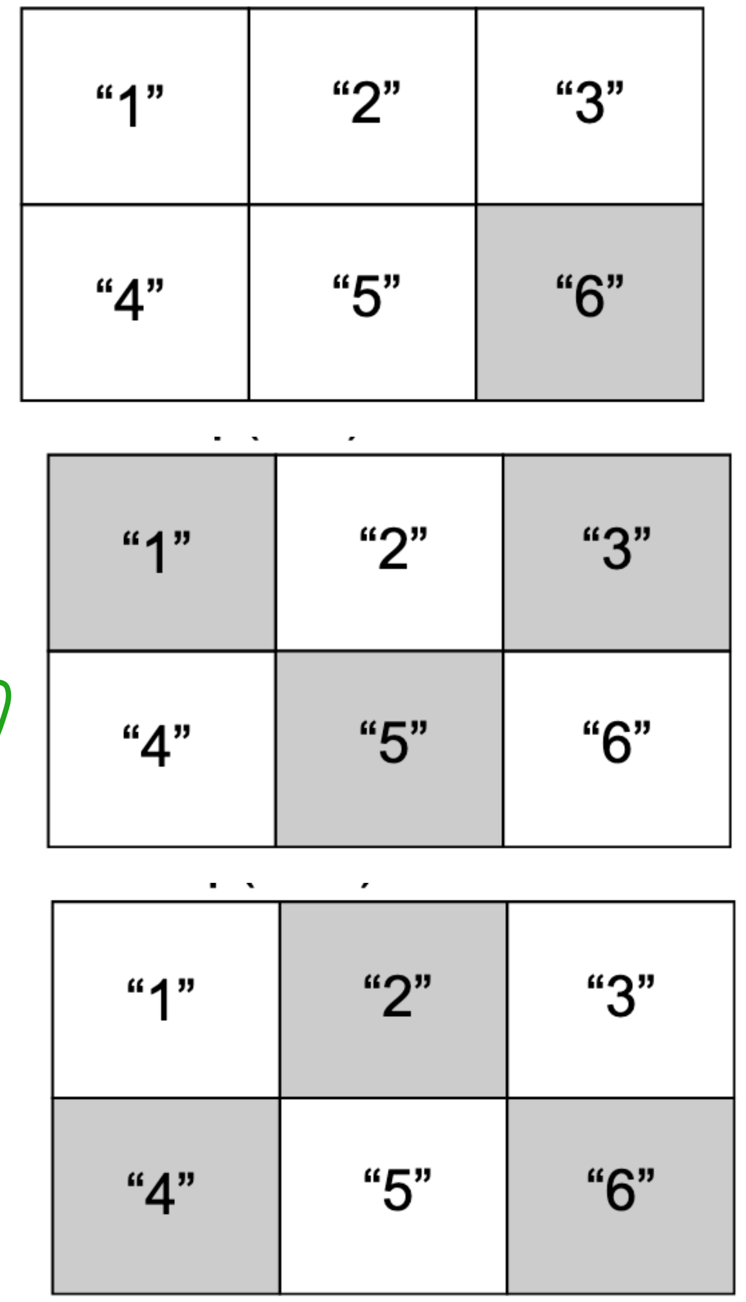
Probability review
Joint probability, independence

- Joint probability of two events A, B,

$P(A, B)$: prob of events A and B both happening

$P(\text{odd}, 6) = 0$, $P(\text{even}, 6) = \frac{1}{6}$

$P(X=6)$



$P(\text{odd}) = 1/2$

$P(\text{even}) = 1/2$

- Marginalisation rule

$$\sum_{x \in X_{all}} P(A, X=x) = P(A)$$

$$\sum_{x \in X_{all}} P(\text{odd}, X=x) = \underbrace{P(\text{odd}, 1)}_{1/6} + \underbrace{P(\text{odd}, 2)}_{0} + \underbrace{P(\text{odd}, 3)}_{1/6} + \dots + \underbrace{P(\text{odd}, 6)}_{0} = 1/2$$

- Independence of two events

A, B are indep. if $P(A, B) = P(A)P(B)$

we call two random variables X, Y indep if $\forall$ values they can take

$$P(X=x, Y=y) = P(X=x)P(Y=y)$$

Probability review
conditional distribution

B : even

“1”	“2”	“3”
“4”	“5”	“6”

“1”	“2”	“3”
“4”	“5”	“6”

- Conditional distribution $P(A \mid B) = \dfrac{P(A, B)}{P(B)}$ ★

prob. of event A given that event B has occured

$P(6 \mid \text{even}) = 1/3$

• Marginalization $\sum_{x \in X_{all}} P(x \mid B) = \sum_{x \in X_{all}} \dfrac{P(x, B)}{P(B)} = \dfrac{1}{P(B)} \sum_{x \in X_{all}} P(x, B) = \dfrac{P(B)}{P(B)} = 1$

- Conditional independence

events A_1, A_2 are conditionally independent given event B:

$$P(A_1, A_2 \mid B) = P(A_1 \mid B)\, P(A_2 \mid B)$$

- Product rule from ★ : $P(A, B) = P(A \mid B)\, P(B) =$

$P(A, B) = P(B \mid A)\, P(A)$

Probability review
Bayes rule

- Recall product rule $P(A,B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A)$

- Bayes rule $P(B \mid A) = \dfrac{P(A \mid B)\,P(B)}{P(A)}$

consider now $P(B \mid A_1, A_2) = \dfrac{P(A_1, A_2 \mid B)\,P(B)}{P(A_1, A_2)}$

... can be extended to $\{A_i\}_{i=1}^{d}$ events.

Bayes rule for classification

$\{x^i, y^i\}_{i=1}^N$, $y^i \in \{1, 2, \ldots, K\}$

x^i taking finitely many values.

Probability to observe this x knowing the label c

Probability to label with c knowing x

Probability to label with class c

$$P(y = c \mid x) = \frac{P(x \mid y = c)P(y = c)}{P(x)}$$

Probability of the x

$P(y)$ is called the **prior probability**

Describes the probability to encounter the data labeled y

Bayes rule for classification

$$P(y=c\,|\,x) = \frac{P(x\,|\,y=c)P(y=c)}{P(x)}, \quad \hat{y} = \underset{c\in\{1,2,\dots,K\}}{\operatorname{argmax}} P(y=c\,|\,x)$$

example : $y \in \{0,1\}$

$P(y=0\,|\,x) > P(y=1\,|\,x)$? $\begin{cases} \text{yes} & \text{label as class } 0 \\ \text{no} & \text{label as class } 1 \end{cases}$

$$\frac{P(x\,|\,y=0)P(y=0)}{P(x)} > \frac{P(x\,|\,y=1)P(y=1)}{P(x)} \iff P(x\,|\,y=0)P(y=0) > P(x\,|\,y=1)P(y=1)$$

Spam detection example

Bag of words to encode text to features

List all words encountered in the set of texts you have

Use a Boolean feature for the presence (1) or absence (0) of each word → $x_i \in \{0,1\}$ absence or presence of word i

In spam email detection, data: emails, labelled as: spam (class 1), or no-spam (class 0)

Suppose your emails contain "Dear", "Friend", "Lunch", "Money".

Data	"Dear" x_1	"Friend" x_2	"Lunch" x_3	"Money" x_4	class
email 1	1	1	1	0	non-spam
email 2	0	0	0	1	spam
⋮					
email N					

EPFL

Spam detection example

Use of Base’s rule for spam detection

$\{x^i, y^i\}_{i=1}^{N} = D$

Need to calculate $P(x \mid y = \text{"spam"})\, P(y = \text{"spam"})$ and

$P(x \mid y = \text{"no.spam"})\, P(y = \text{"no.spam"})$.

$$P(y = \text{"spam"}) = \frac{\#\ \text{spam emails in } D}{N},$$

$$P(y = \text{"no.spam"}) = \frac{\#\ \text{non.spam emails in } D}{N}$$

$P(x \mid y = \text{"spam"})$?   example: $x = (1, 0, 0, 1)$.

$P((1,0,0,1) \mid y = 1)$, $P((1,0,0,1) \mid y = 0)$?

EPFL

Naive Bayes assumption

Conditional independence of different features given the label $x \in \mathbb{R}^d$

$$P(x \mid y) = \prod_{j=1}^{d} P(x_j \mid y)$$

the d features of our data are conditionally independent given the label y

$$P((1,0,0,1) \mid y = \underbrace{\text{"spam"}}_{1}) = P(x_1 = 1 \mid 1)\, P(x_2 = 0 \mid 1) P(x_3 = 0 \mid 1) P(x_4 = 1 \mid 1)$$

$$P(x_j = 1 \mid 1) = \frac{\#\text{ of spam emails containing word } j}{\#\text{ of spam emails.}} \qquad j = 1, 2, \dots, d$$

(in our example $d = 4$)

$$P(x_j = 0 \mid 1) = \frac{\#\text{ of spam emails not containing word } j}{\#\text{ of spam emails}}$$

Naive Bayes classifier

$x \in \{0,1\}^d = \{0,1\} \times \{0,1\} \times \dots \times \{0,1\}$

continuing example ...

$$P(y=1 \mid x) = \frac{\prod_{j=1}^{d} P(x_j \mid y=1)\, P(y=1)}{P(x)}$$

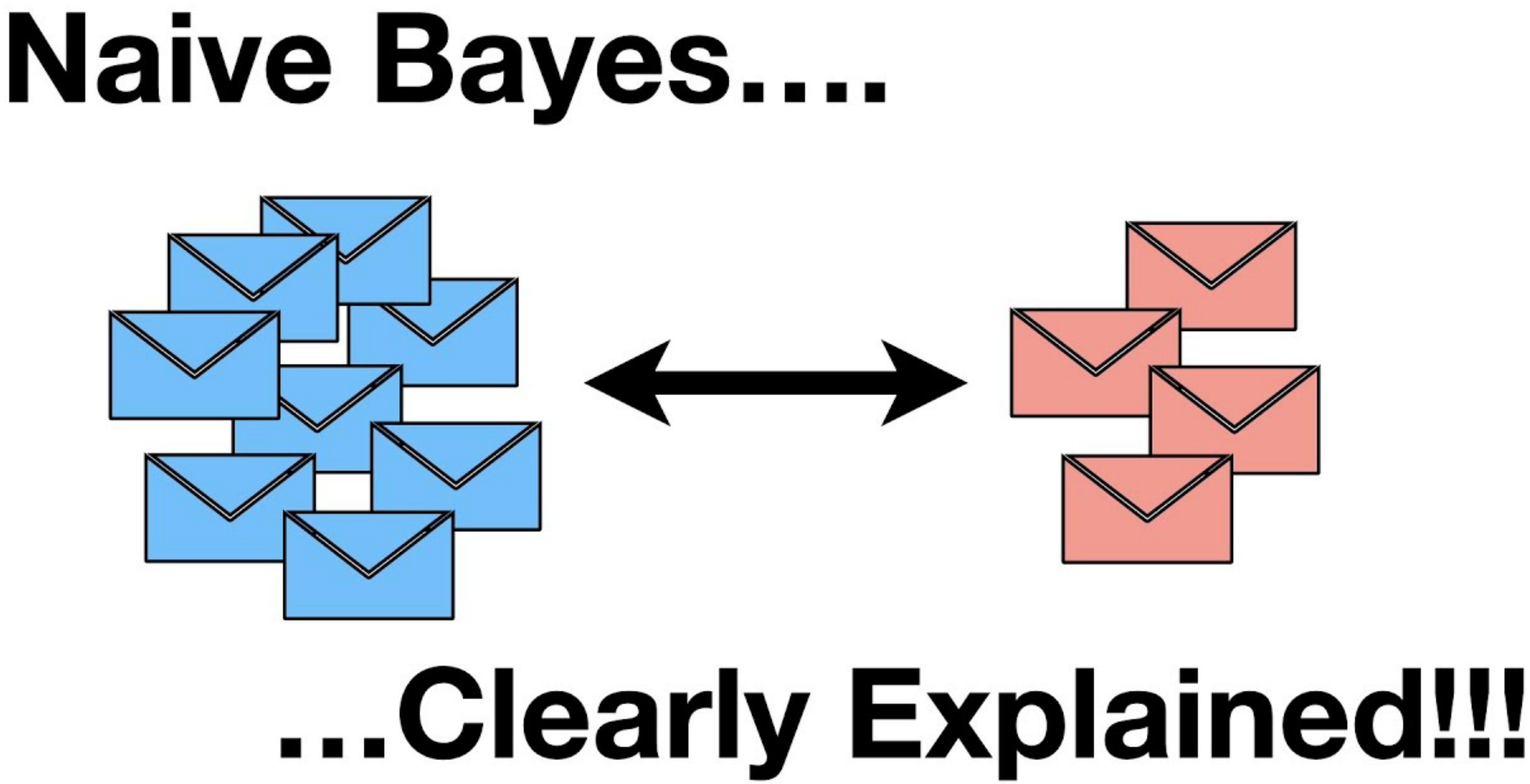
d : # of features (here words)

to determine $P(y=1 \mid x) > P(y=0 \mid x)$ we don't need to compute denominator $P(x)$ (same in both sides of the inequality)

Need to compare ...

$$\prod_{j=1}^{d} P(x_j \mid y=1)\, P(y=1) \overset{?}{>} \prod_{j=1}^{d} P(x_j \mid y=0)\, P(y=0)$$

Naive Bayes classifier for spam detection
Example with worked out numbers



https://www.youtube.com/watch?v=O2L2Uv9pdDA

Activity

- Watch: https://www.youtube.com/watch?v=O2L2Uv9pdDA
  - This takes 15 minutes
- After watching the video, answer the following questions
  1. How are the features defined differently than the binary scores we defined earlier?
  2. How would you evaluate the accuracy of the classifier?
  3. Are there any hyper parameters to tune for this classifier?
  4. Is there any other approach you could imagine using for this classification?

- Next class: discuss your answers with your two nearest neighbours
- Pick one person to represent your answers to class

Naive Bayes Classifier
Continuous-valued features

Bayes rule
Continuous-valued features

$y \in \{1, 2, \dots, K\}$

$x \in \mathbb{R}^d$ continuous-valued

$$P(y = c \,|\, x) = \frac{f_{x|y=c}(x)P(y = c)}{f_x(x)}$$

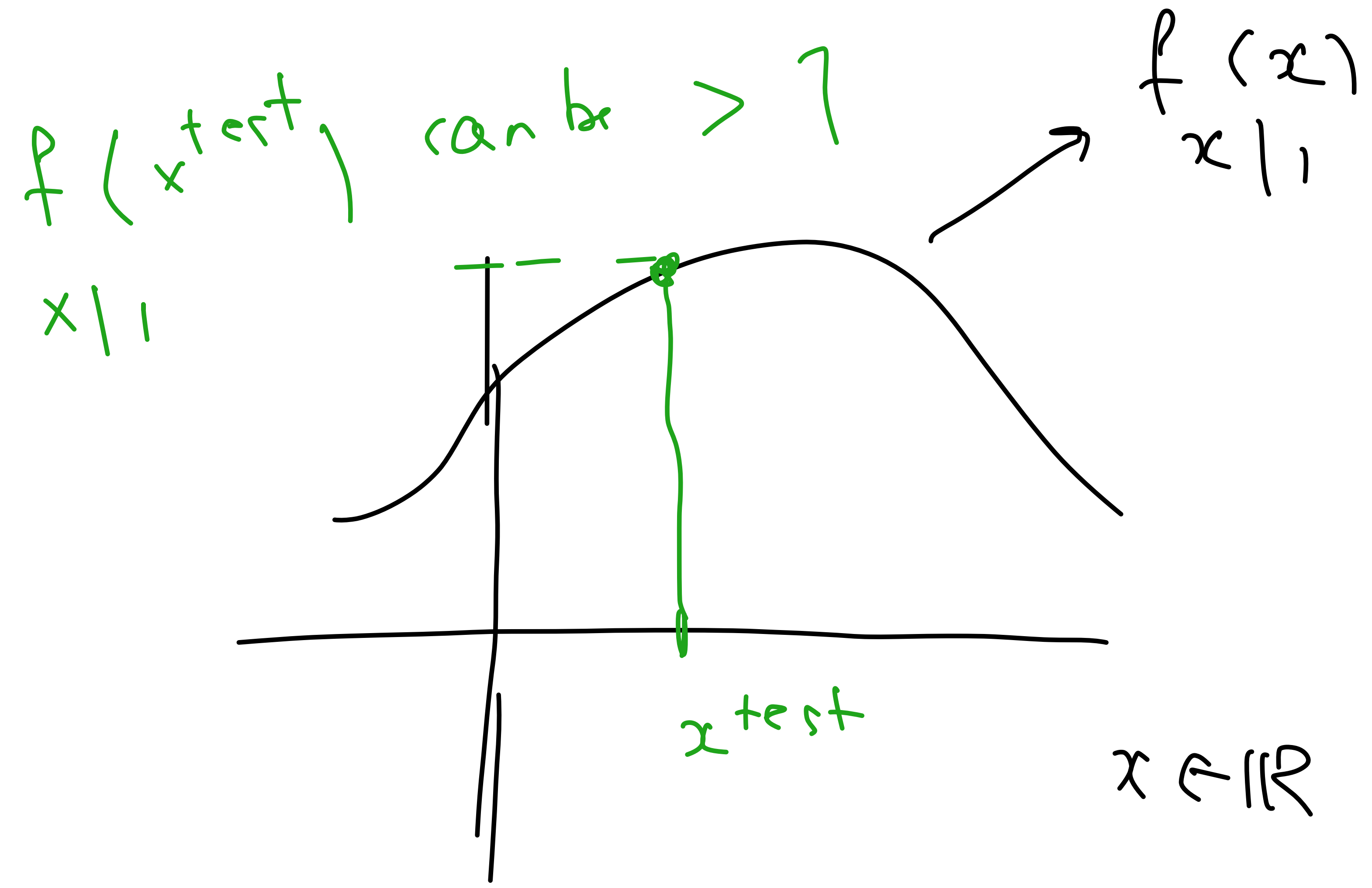
$f_{x|y=c}(x)$ Probability density function: probability distribution for a continuous random variable

Probability of $x = v$ is 0.

Probability of $x \in [a_1, a_2] = \int_{a_1}^{a_2} f_{x|1}(x)\,dx$

$x \in \mathbb{R}^d$, $S \subseteq \mathbb{R}^d$

$P(x \in S) = \int_S f_{x|1}(x)\,dx$.



Naive Bayes assumption
Continuous features

$$P(y = c \,|\, x) = \frac{f_{x|y=c}(x) P(y = c)}{f_x(x)} = \frac{\prod_{j=1}^{d} f_{x_j|y=c}(x_j) \, P(y=c)}{f_{x}(x)}$$

Naive Bayes assumption: says given class c, the features are conditionally independent.

$x \in \mathbb{R}^d$ $\quad f_{x|y=c}(x)$: conditional density of x given class c

$$f_{x|y=c}(x) = \prod_{j=1}^{d} f_{x_j|y=c}(x_j)$$

Gaussian Naive Bayes

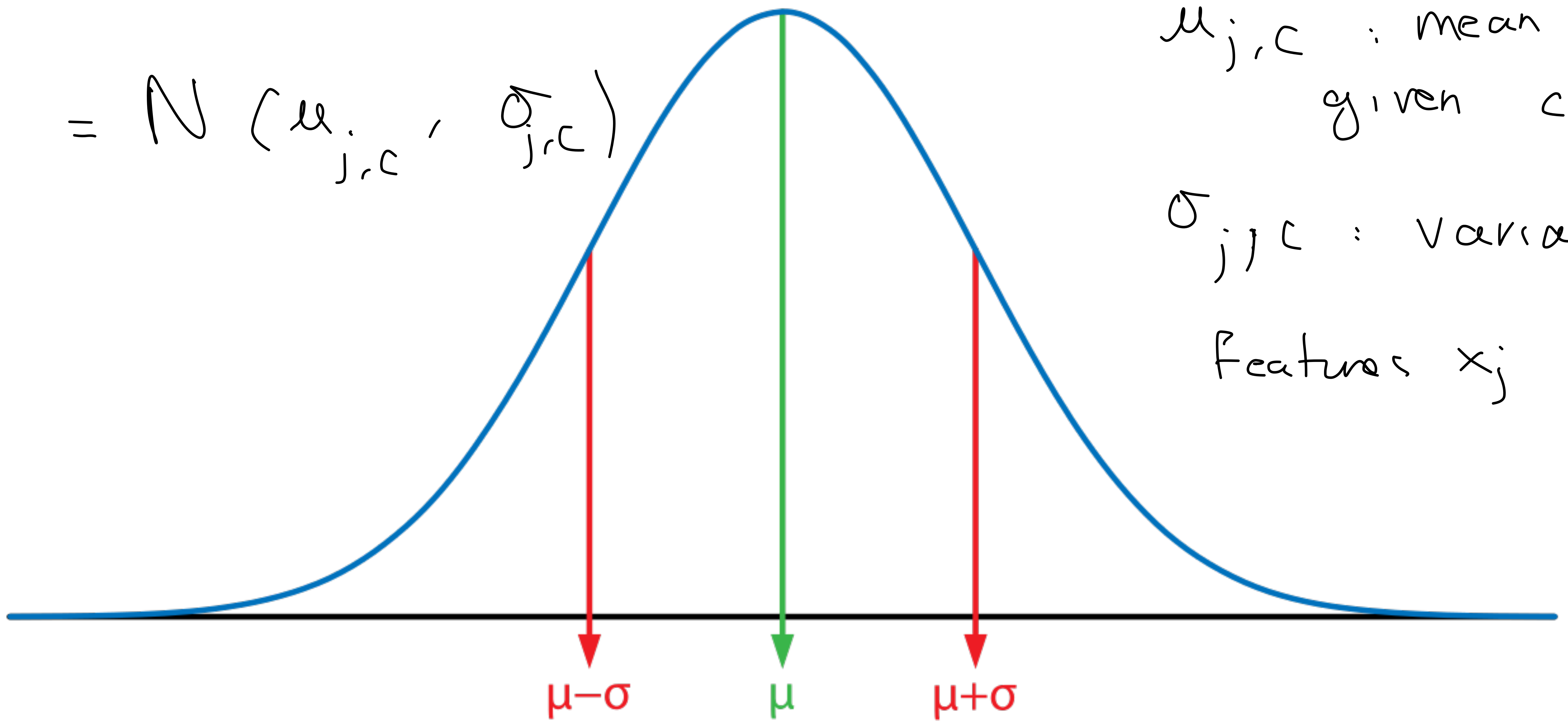
Assumes that the probability density function for each feature follows a gaussian distribution

$f_{x_j|c}(x_j)$: conditional density of feature j given class c.

$f_{x_j|c}(x_j) = N(\mu_{j,c}, \sigma_{j,c})$

$\mu_{j,c}$: mean of features x_j given class c

$\sigma_{j,c}$: variance of features x_j given class c.



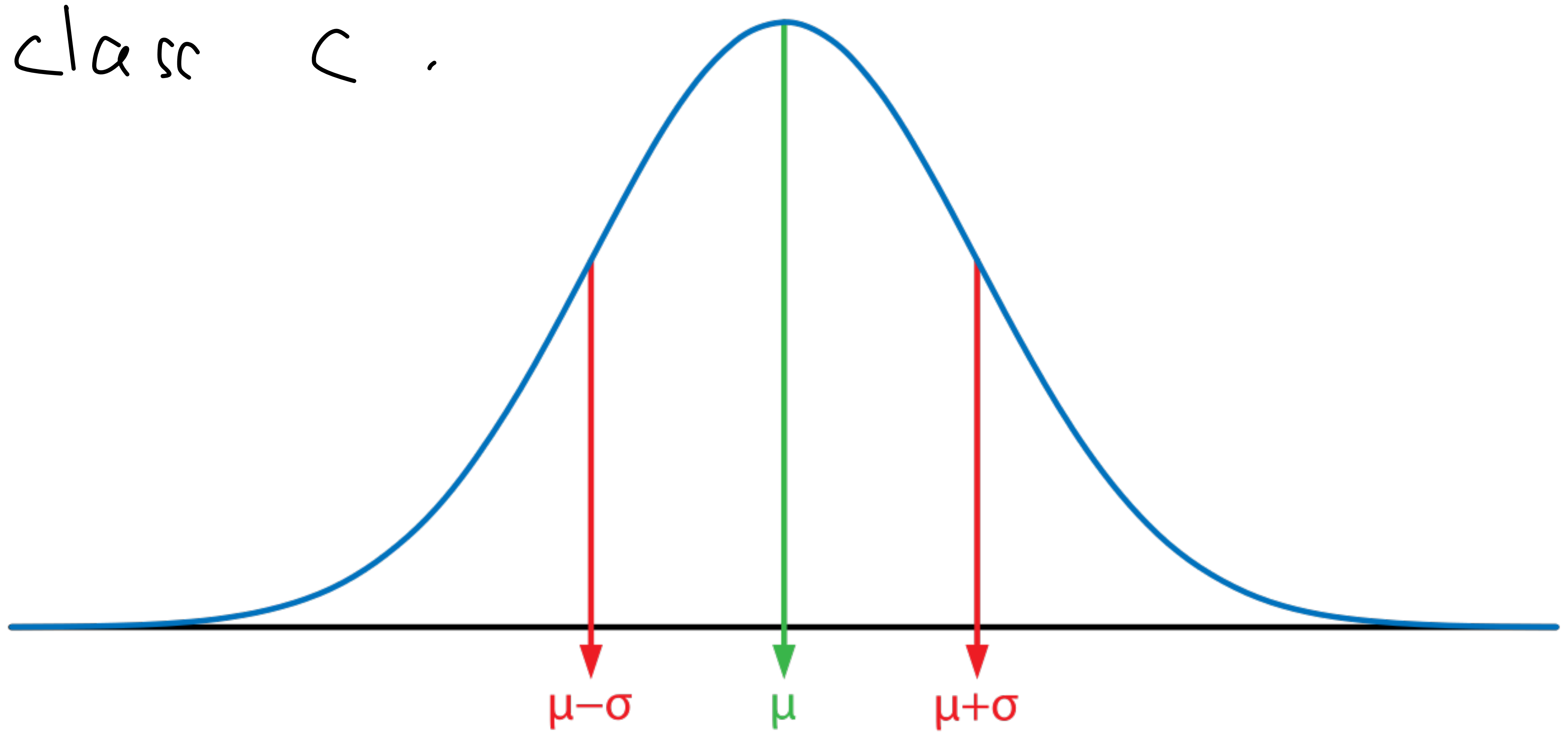
Gaussian Naive Bayes

$\{ x^i, y^i \}_{i=1}^{N}, \ x^i \in \mathbb{R}^d$

Estimating the conditional Gaussian distribution for each feature using sample data

Recall $\mu_{j,c} = \frac{1}{|I_c|} \sum_{i \in I_c} x_j^i$ empirical mean

I_c : indices of data corresponding to class c.



$$\sigma_{j,c}^2 = \frac{\sum_{i \in I_c} (x_j^i - \mu_{j,c})^2}{|I_c| - 1},$$

empirical variance

$$N(\mu_{j,c}, \sigma_{j,c}) = f_{x_j | y=c}(x_j)$$

Gaussian Naive Bayes

Classifier prediction

$$P(y = c \mid x) = \frac{f_{x|y=c}(x)P(y = c)}{f_x(x)} = \frac{\prod_{j=1}^{d} f_{x_j|y=c}(x_j)\, P(y=c)}{f_x(x)}$$

We used data to approximate the density function . .

$$f_{x_j|y=c}(x_j) = N(\mu_{j,c}, \sigma_{j,c})$$

& to also determine $P(y=c)$ $\forall\, c \in \{1, 2, \dots, K\}$

$x^{test} \in \mathbb{R}^d \implies$ evaluate $\underbrace{\prod_{j=1}^{d} f_{x_j|y=c}(x_j^{test})\, P(y=c)}$ for $c \in \{1, 2, \dots, K\}$

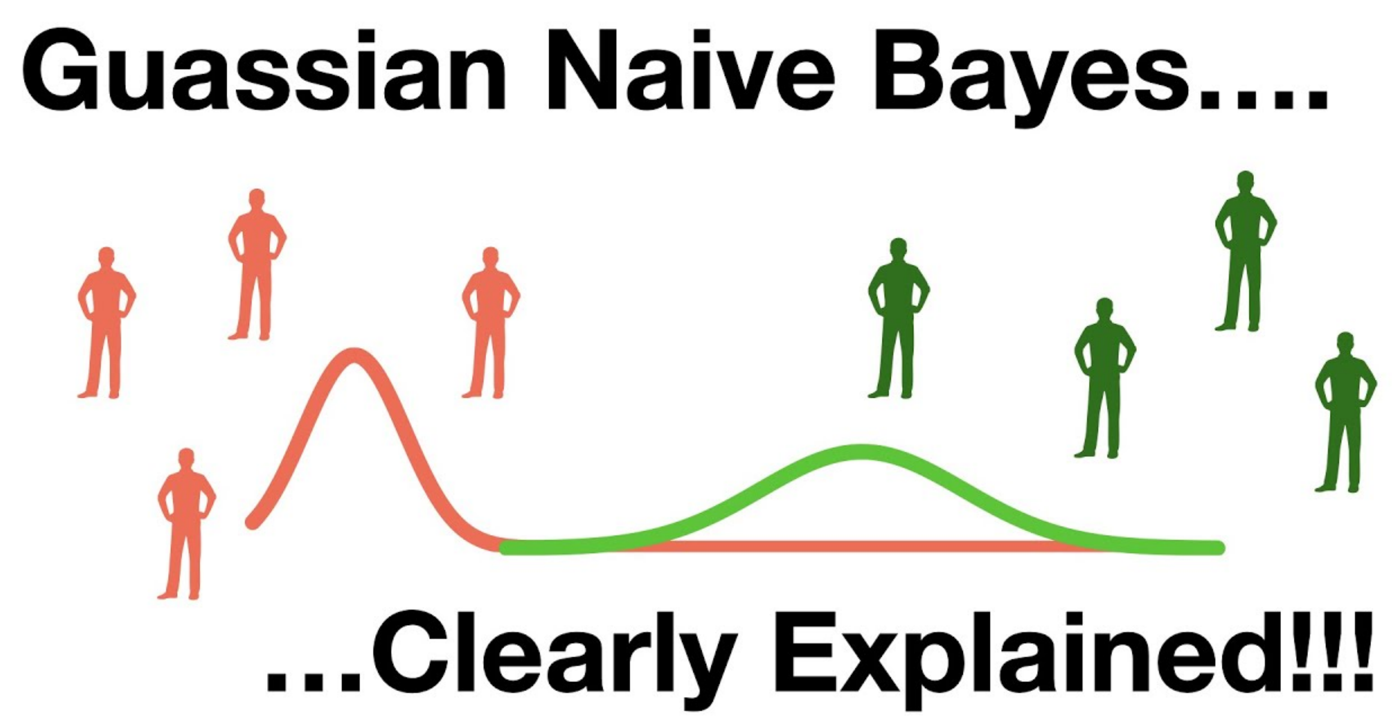
& gives label $y^{test} = \underset{c}{\operatorname{argmax}}$ of

Gaussian Naive Bayes

Naive Bayes assumption and final classifier

$$P(y = c \mid x) = \frac{f_{x_1|y=c}(x_1) f_{x_2|y=c}(x_2) \ldots f_{x_d|y=c}(x_d) P(y = c)}{f_x(x)}$$

you can see a worked out example in the video below.



Naive Bayes
Summary

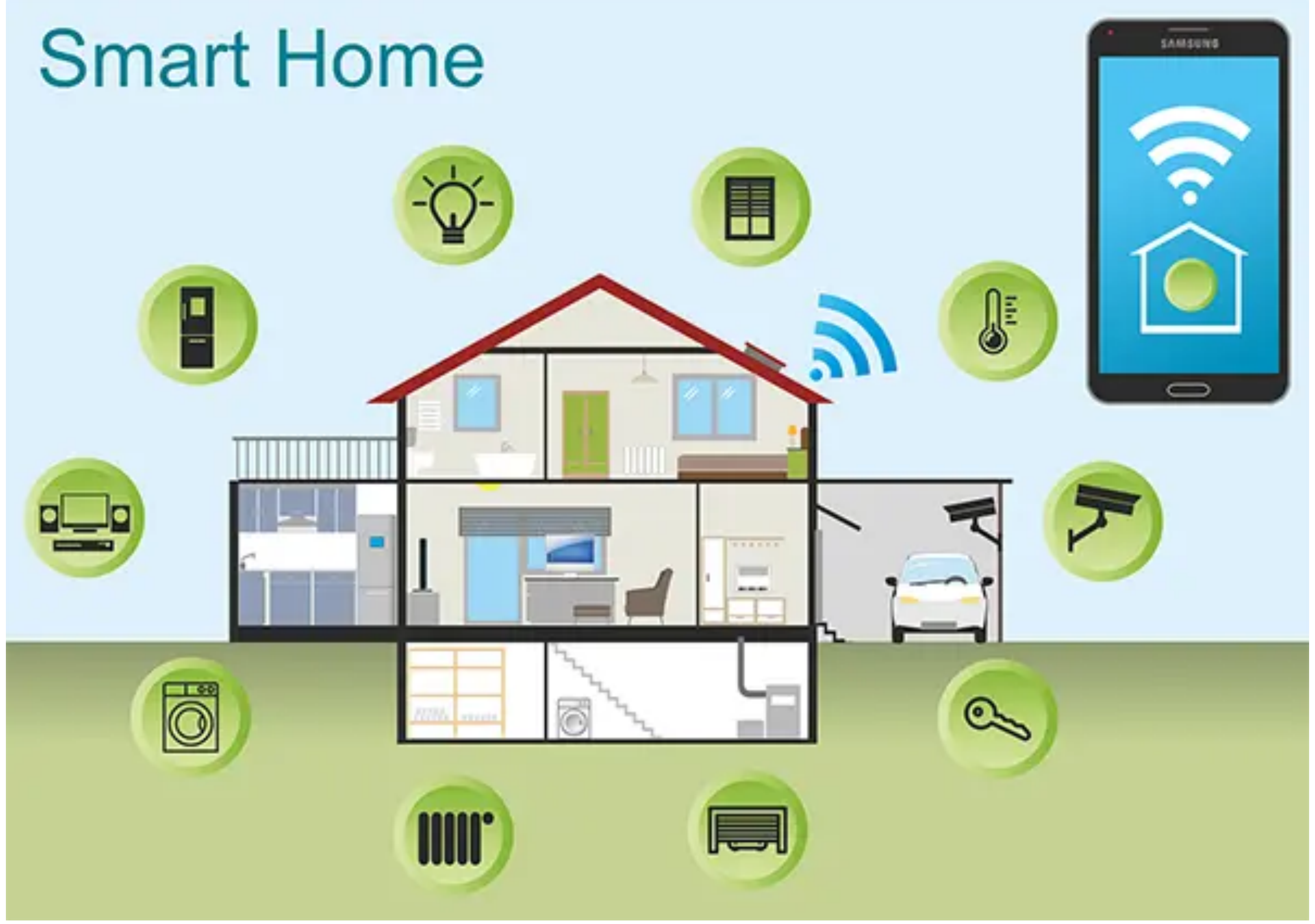
- Probabilistic classifier
  - Features can be finite- or continuous-valued

- Uses Bayes rule and empirical probabilities computed based on data

- Assumes different features are independent given the class label $\Rightarrow$ Naive assumption
  - In practice hard to verify the assumption but works well on certain problems

Naive Bayes classifier
Application in python exercises

- **Dataset:** Smart Home Device Efficiency
- **Goals:** Evaluate the efficiency of a smart home device



Quiz 7

ML an approach to AI & uses $\boxed{\text{data}}$ to

learn a predictor (supervised learng) $\{x^i, y^i\}_{i=1}^d$

a pattern (unsupervised learng)

supervised learng

$y \in \mathbb{R}^m$ regression

→ linear regression

⋮

classification $y \in \{1, 2, \dots, K\}$

{ logistic regression

⋮

$D = \{x^i, y^i\}_{i=1}^{N}$ $\qquad y^i \in \{1, 2, \dots, K\}$

model $\quad f_{w,b}(x)$

$$D \longrightarrow \boxed{\begin{array}{c}\text{loss function}\\ w, b\end{array}} \longrightarrow \boxed{\text{training}} \longrightarrow w^*, b^*$$

$$x^{test} \longrightarrow \boxed{f_{w^*,b^*}(x^{test})} \longrightarrow y^{test} \in \{1, 2, \dots, K\}$$

- training set, validation set, test.
- underfitting & overfitting. • feature expansion.