UNC CHARLOTTE

The WILLIAM STATES LEE COLLEGE of ENGINEERING

Introduction to ML
Lecture 6: Naive Bayes Classifier

Hamed Tabkhi
Department of Electrical and Computer Engineering,
University of North Carolina Charlotte (UNCC)
htabkhiv@uncc.edu

UNC CHARLOTTE

Naive Bayes Classifier

Thomas Bayes

1702 - 1761

English statistician

Naive Bayes Classifier

Some basics

The probability of two events A and B happening, P(A∩B), is the probability of A, P(A), times the probability of B given that A has occurred, P(B|A).

$$P(A \cap B) = P(A)P(B|A) \tag{1}$$

On the other hand, the probability of A and B is also equal to the probability of B times the probability of A given B.

$$P(A \cap B) = P(B)P(A|B) \tag{2}$$

Equating the two yields: **(1) = (2)**

$$P(B)P(A|B) = P(A)P(B|A) \tag{3}$$

Bayes Theorem or Bayes Rule

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

UNC CHARLOTTE

Naive Bayes Classifier

- Bayesian classifiers use **Bayes theorem**, which says

$$\boxed{p(c_j \mid d)} = \frac{\boxed{p(d \mid c_j)}\,\boxed{p(c_j)}}{p(d)}$$

posterior probability likelihood prior

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

feature d

- $p(c_j \mid d)$ = probability of instance d being in class c_j,
  This is what we are trying to compute
- $p(d \mid c_j)$ = probability of generating instance d given class c_j,
  We can imagine that being in class c_j, causes you to have feature d with some probability
- $p(c_j)$ = probability of occurrence of class c_j,
  This is just how frequent the class c_j, is in our database
- $p(d)$ = probability of instance d occurring

  This can actually be ignored, since it is the same for all classes
  Can be viewed as normalization factor

The following slides are adapted from Eamonn Keogh and Eric Eaton.

UNC CHARLOTTE

Naive Bayes Classifier

Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

UNC CHARLOTTE

Assume that we have two classes

c_1 = male, and c_2 = female.

We have a person whose sex we do not know, say "*drew*" or *d*.

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, I.e which is greater $p(\text{male} \mid drew)$ or $p(\text{female} \mid drew)$

(Note: "Drew can be a male or female name")

Drew Barrymore

Drew Carey

What is the probability of being called "*drew*" given that you are a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

What is the probability of being a male?

What is the probability of being named "*drew*"?

(actually irrelevant, since it is that same for all classes)

Officer Drew

This is Officer Drew

Is Officer Drew a Male or Female?

Luckily, we have a small database with names and sex.

We can use it to apply Bayes rule…

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

$$p(\text{male} \mid \textit{drew}) = \frac{p(\textit{drew} \mid \text{male})\, p(\text{male})}{p(\textit{drew})}$$

The probability of being a male

feature	label
Name	**Sex**
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Officer Drew

1. The probability of being a male

$$p(male \mid drew) = \frac{p(drew \mid male)\, p(male)}{p(drew)}$$

$p(drew \mid male)$?

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

UNC CHARLOTTE

Officer Drew

1. The probability of being a male

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

$p(drew \mid \text{male})$ =1/3

$\frac{1}{3}$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Officer Drew

1. The probability of being a male

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$p(\text{male}) = 3/8$

Officer Drew

1. The probability of being a male

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(drew) = 3/8$$

UNC CHARLOTTE

Officer Drew

1. The probability of being a male

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

$$p(\text{male} \mid drew) = \frac{\mathbf{1/3 * 3/8}}{\mathbf{3/8}} = \frac{\mathbf{0.125}}{3/8}$$

UNC CHARLOTTE

Officer Drew

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

We can follow the same procedure to compute the probability of "Drew" being a female.

p(drew|female)

p(female)

$$p(\text{female} \mid drew) = \frac{2/5 \; * \; 5/8}{3/8} = \frac{0.250}{3/8}$$

UNC CHARLOTTE

Officer Drew

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid drew) = \frac{1/3 \; * \; 3/8}{3/8} = \frac{0.125}{3/8}$$

$$\wedge$$

$$p(\text{female} \mid drew) = \frac{2/5 \; * \; 5/8}{3/8} = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female.

Officer Drew IS a female!

Officer Drew

$$p(\text{male} \mid drew) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

Naive Bayes Classifier

So far we have only considered Bayes Classification when we have one attribute/feature (the "*name*"). But we may have many features.

How do we use all the features?

Naive Bayes Classifier

- "naive" conditional independence assumptions

Independence

- Two events A, B are independent, if (the following are equivalent)
  - $P(A, B) = P(A) * P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$

Naive Bayes Classifier

- "naive" conditional independence assumptions

Assume that each feature x_i is conditionally independent of every other feature x_j for $j \neq i$, given the category C_k

That means:

$$p(x_i \mid \cancel{x_{i+1}, \ldots, x_n}, C_k) = p(x_i \mid C_k)$$

Naive Bayes Classifier

- Thus, the joint model can be expressed as

$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) \boxed{\propto} \, & p(C_k, x_1, \dots, x_n) \\ & = p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\ & = p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

Denote " proportional "

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{\boxed{p(\mathbf{x})}}$$

Naive Bayes Classifier

Using the Naïve Bayes Classifier

To classify a new point $\mathbf{x}$

Find the maximum

$$\hat{y} = \underset{k \in \{1,...,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

Any potential problem ?

There could be cases where the classification could be multivariate. Therefore, we need to find the class **y** with maximum probability.

UNC CHARLOTTE

Naive Bayes Classifier

Using the Naïve Bayes Classifier

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

We are multiplying lots of small numbers

Danger of underflow!

- $0.5^{57} = 7$ E -18

a number of smaller absolute value than the computer can actually represent in memory on its CPU.

Solution?

Naive Bayes Classifier

We are multiplying lots of small numbers
Danger of underflow!

- $0.5^{57} = 7$ E -18

Solution? $\boxed{\text{Use logs and add!}}$

– In practice, we use log-probabilities to prevent underflow

$$\arg\max_{y_k} \; \log P(Y = y_k) + \sum_{j=1}^{n} \log P(X_j = x_j \mid Y = y_k)$$

class label {1, … , K}

UNC CHARLOTTE

Problems with Naive Bayes Classifier

- Naive Bayes assumption
  - Usually, features are not conditionally independent:

$$\boxed{P(X_1 \ldots X_n | Y) \neq \prod_i P(X_i | Y)}$$

  - The naïve Bayes assumption is often violated, yet it performs surprisingly well in many cases.

More deeper investigation: Zhang, Harry. "The optimality of naive Bayes." *AAAI* (2004).

Problems with Naive Bayes Classifier

- Notice that some probabilities estimated by counting might be zero

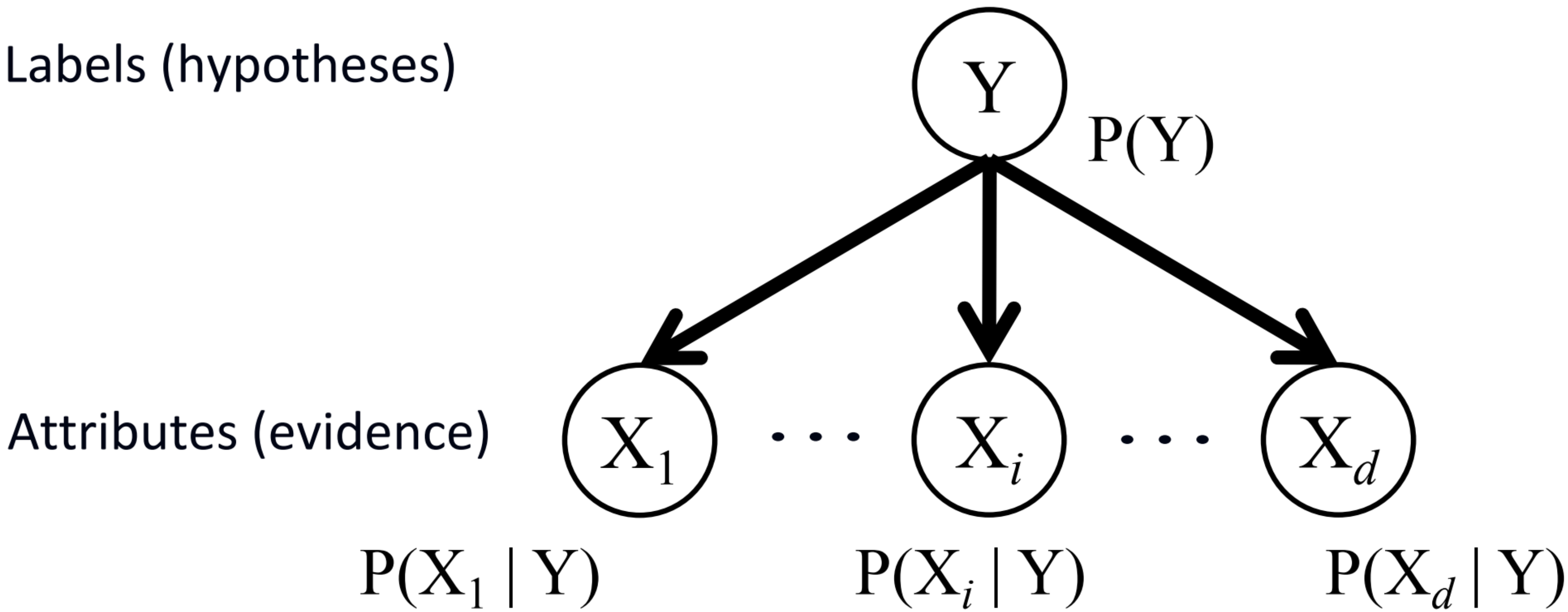
- Fix by using Laplace smoothing:
  - Likelihood:

$$P(X_i = x|Y = y) = \frac{Count(X_i = x, Y = y) \boxed{+1}}{\sum_{x'} Count(X_i = x', Y = y) \boxed{+ |\text{values}(X_i)|}}$$

$|\text{values}(X_i)|$ is the number of values X_i can take on

UNC CHARLOTTE

The Naïve Bayes Graphical Model

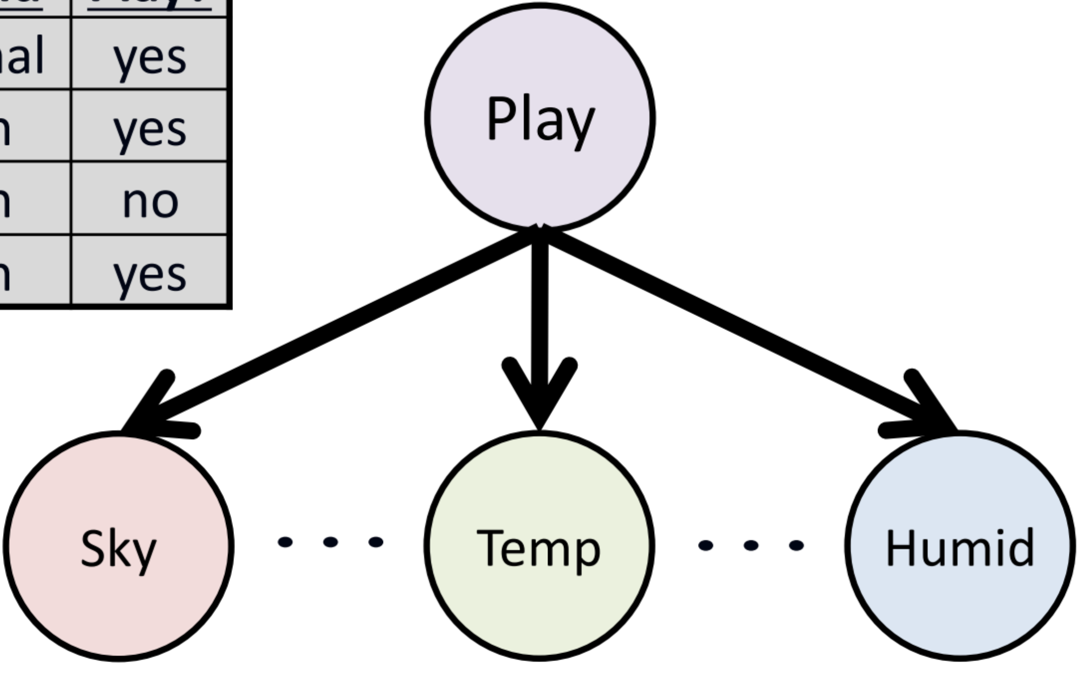


- Nodes denote random variables
- Edges denote dependency
- Each node has an associated conditional probability table (CPT), conditioned upon its parents

Example NB Graphical Model

Data:

Sky	Temp	Humid	Play?
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes

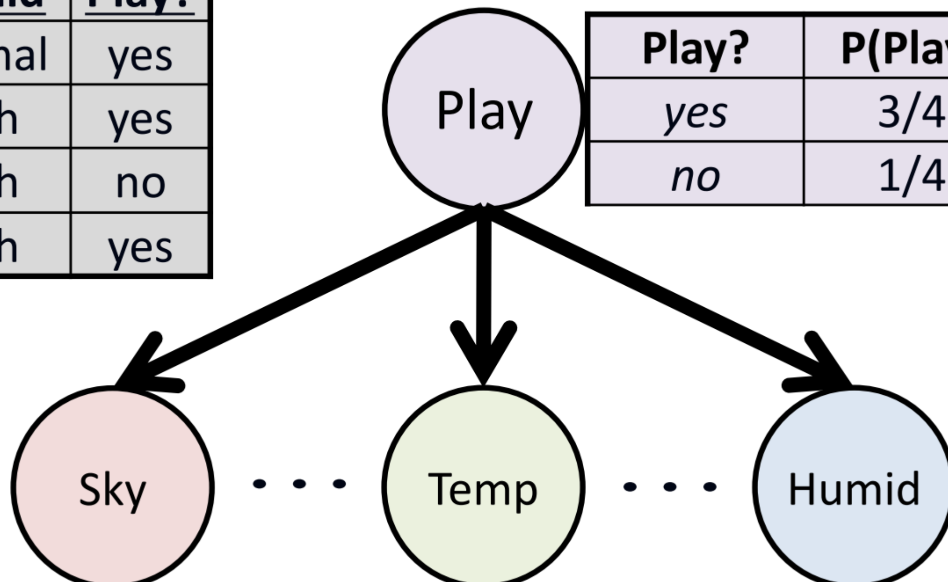


UNC CHARLOTTE

Example NB Graphical Model

Data:

Sky	Temp	Humid	Play?
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Play?	P(Play)
yes	3/4
no	1/4

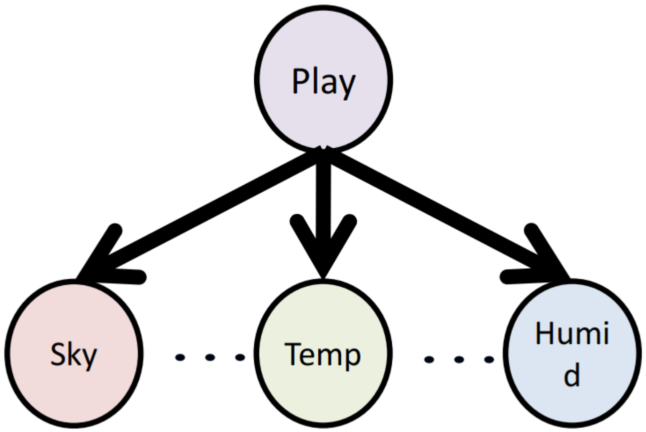
Laplace smoothing is applied

Sky	Play?	P(Sky | Play)
sunny	*yes*	4/5
rainy	*yes*	1/5
sunny	*no*	1/3
rainy	*no*	2/3

Temp	Play?	P(Temp | Play)
warm	*yes*	4/5
cold	*yes*	1/5
warm	*no*	1/3
cold	*no*	2/3

Humid	Play?	P(Humid | Play)
high	*yes*	3/5
norm	*yes*	2/5
high	*no*	2/3
norm	*no*	1/3

Example Using NB for Classification



Play?	P(Play)
yes	3/4
no	1/4

Temp	Play?	P(Temp | Play)
warm	*yes*	4/5
cold	*yes*	1/5
warm	*no*	1/3
cold	*no*	2/3

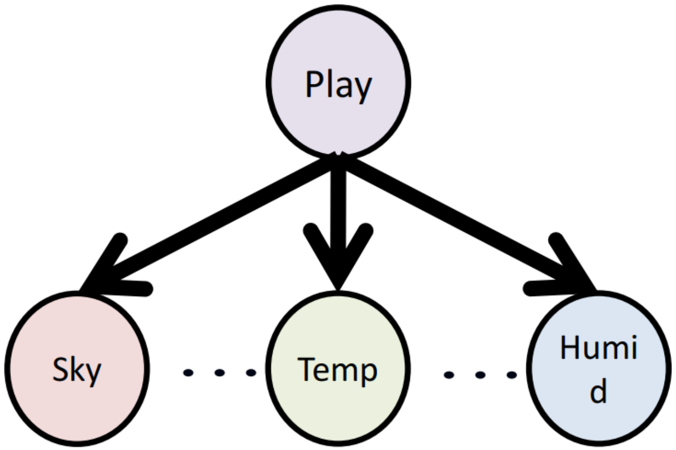
Sky	Play?	P(Sky | Play)
sunny	*yes*	4/5
rainy	*yes*	1/5
sunny	*no*	1/3
rainy	*no*	2/3

Humid	Play?	P(Humid | Play)
high	*yes*	3/5
norm	*yes*	2/5
high	*no*	2/3
norm	*no*	1/3

$$h(\mathbf{x}) = \arg\max_{y_k} \ \log P(Y = y_k) + \sum_{j=1}^{d} \log P(X_j = x_j \mid Y = y_k)$$

Goal: Predict label for $\mathbf{x}$ = (rainy, warm, normal)

Example Using NB for Classification



Play?	P(Play)
yes	3/4
no	1/4

Temp	Play?	P(Temp | Play)
warm	*yes*	4/5
cold	*yes*	1/5
warm	*no*	1/3
cold	*no*	2/3

Sky	Play?	P(Sky | Play)
sunny	*yes*	4/5
rainy	*yes*	1/5
sunny	*no*	1/3
rainy	*no*	2/3

Humid	Play?	P(Humid | Play)
high	*yes*	3/5
norm	*yes*	2/5
high	*no*	2/3
norm	*no*	1/3

Predict label for:

$\mathbf{x}$ = (rainy, warm, normal)

$$P(\text{play} \mid \mathbf{x}) \propto \log P(\text{play}) + \log P(\text{rainy} \mid \text{play}) + \log P(\text{warm} \mid \text{play}) + \log P(\text{normal} \mid \text{play})$$

$$\propto \log 3/4 + \log 1/5 + \log 4/5 + \log 2/5 = \boxed{-1.319 \quad \text{predict PLAY}}$$

$$P(\neg\text{play} \mid \mathbf{x}) \propto \log P(\neg\text{play}) + \log P(\text{rainy} \mid \neg\text{play}) + \log P(\text{warm} \mid \neg\text{play}) + \log P(\text{normal} \mid \neg\text{play})$$

$$\propto \log 1/4 + \log 2/3 + \log 1/3 + \log 1/3 = -1.732$$

Continuous features (Gaussian Naive Bayes)

- Suppose the training data contains a continuous attribute x

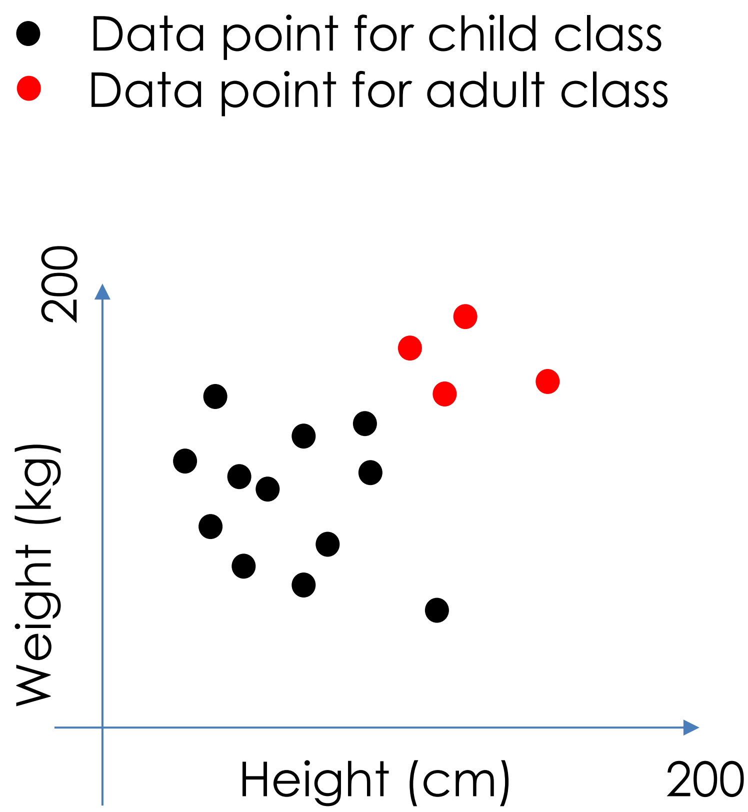
- We first segment the data by the class, and then compute the mean and variance of x in each class

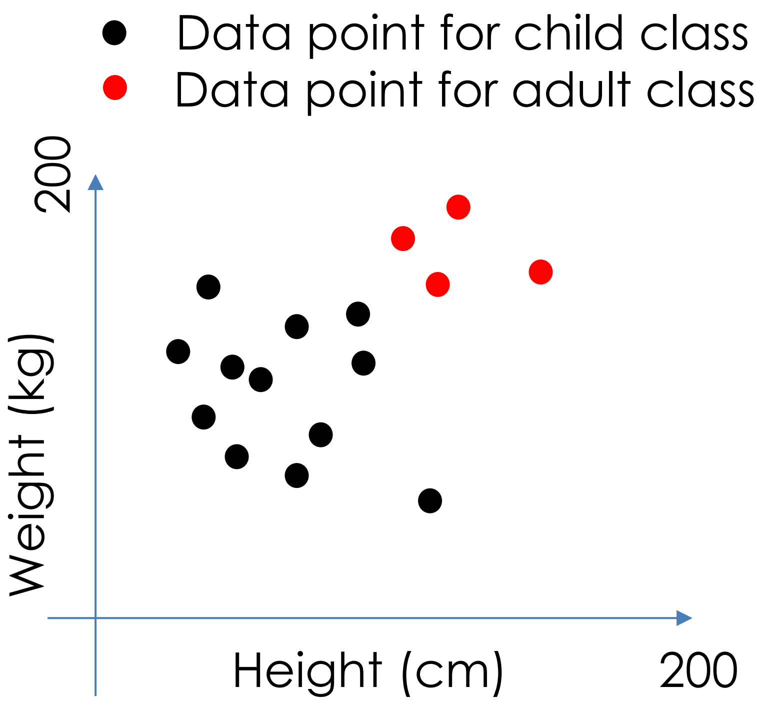
- For each class, compute

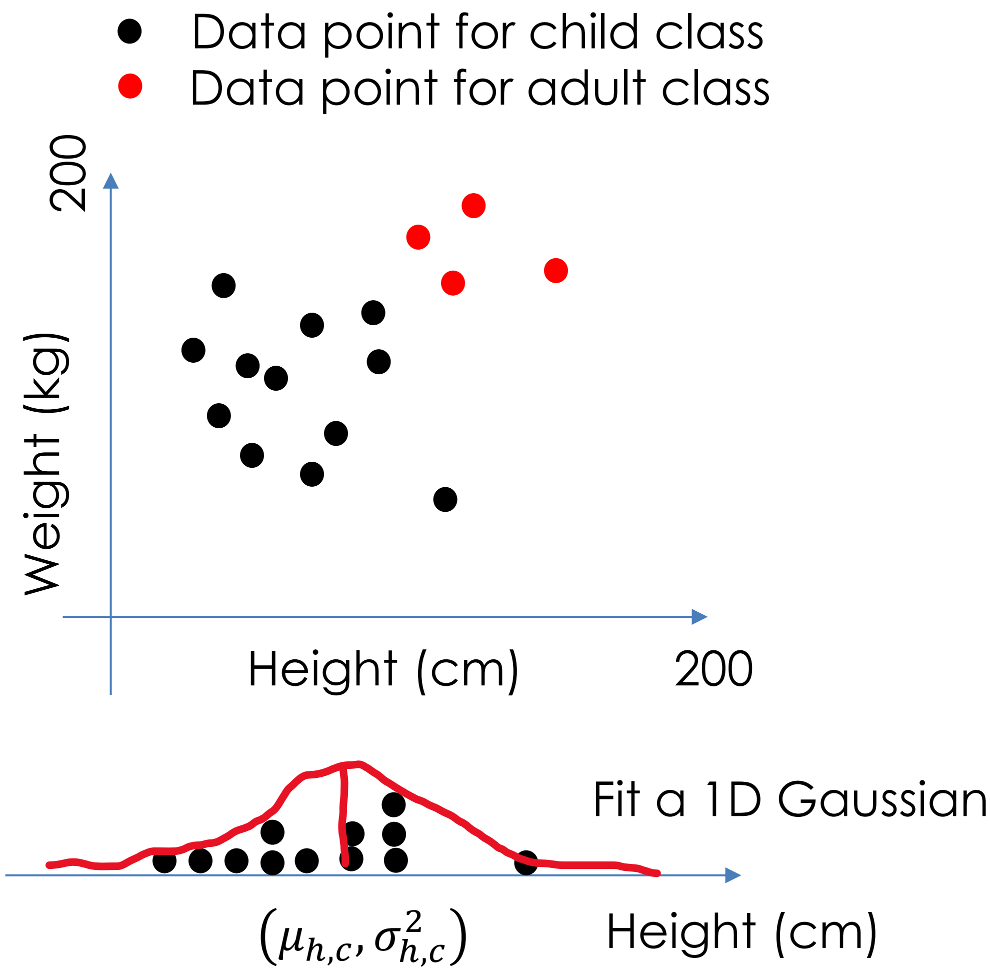
μ_k : the mean of x in class C_k

σ_k^2 : the variance of x in class C_k

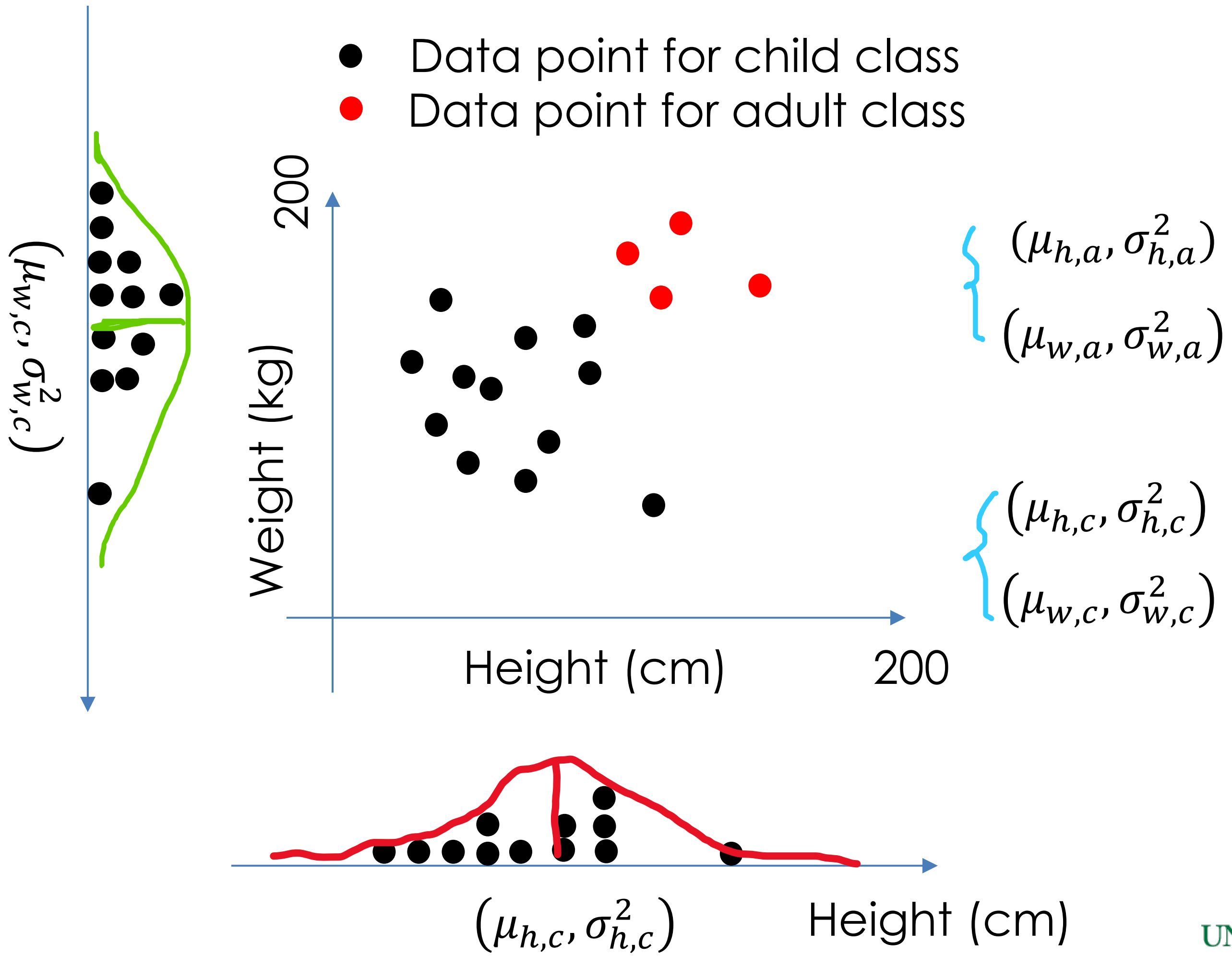
UNC CHARLOTTE

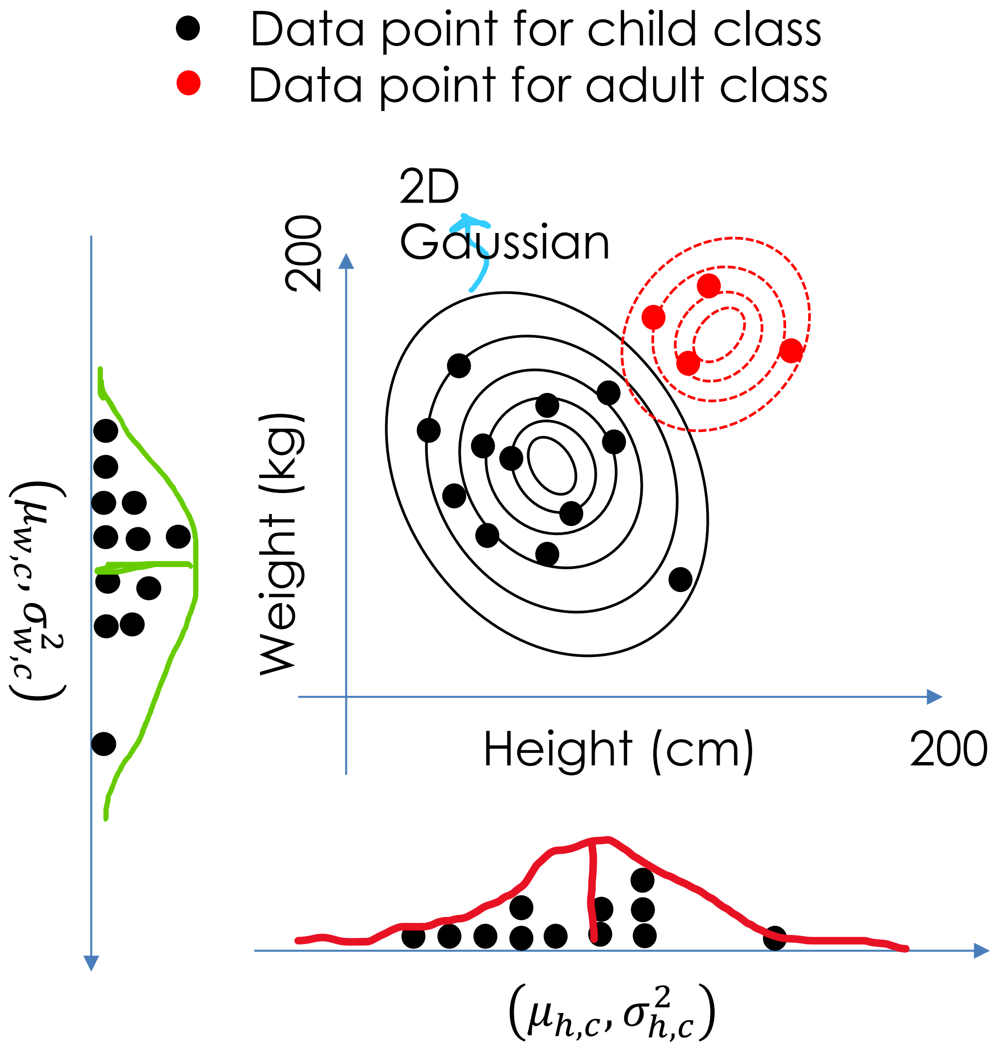

Data point for child class
Data point for adult class
200
Weight (kg)
Height (cm)
200


Data point for child class
Data point for adult class
200
Weight (kg)
Height (cm)
200



Data point for child class
Data point for adult class
200
Weight (kg)
Height (cm)
200
Fit a 1D Gaussian
$(\mu_{h,c}, \sigma^2_{h,c})$
Height (cm)


Data point for child class
Data point for adult class
200
Weight (kg)
Height (cm)
200
$(\mu_{h,a}, \sigma^2_{h,a})$
$(\mu_{w,a}, \sigma^2_{w,a})$
$(\mu_{h,c}, \sigma^2_{h,c})$
$(\mu_{w,c}, \sigma^2_{w,c})$
$(\mu_{w,c}, \sigma^2_{w,c})$
$(\mu_{h,c}, \sigma^2_{h,c})$
Height (cm)



Data point for child class
Data point for adult class
2D Gaussian
200
Weight (kg)
Height (cm)
200
$(\mu_{w,c}, \sigma^2_{w,c})$
$(\mu_{h,c}, \sigma^2_{h,c})$

Example

- Example

Problem: classify whether a given person is **a male or a female** based on the measured features. The features include height, weight, and foot size.

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9