# Naive Bayes

MR. HUYNH NAM

# Guess the gender

## The Dataset

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

## Example

Is this officer a Male or a Female?

**Officer Drew**

# The Dataset

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

# Guess the gender

Assume that we have two classes

$C_1$= male, $C_2$= female

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

Drew can be a male or a female name

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(male | drew) or P(female | drew)

# Bayes approach

$$\text{P(class|data)} = \frac{\text{P(class)} * \text{P(data|class)}}{\text{P(data)}}$$

$$posterior = \frac{prior \times likelihood}{evidence}$$

# Result

Officer Drew is female


**Officer Drew**

| | Outlook | Play Golf |
|---|---|---|
| 0 | Rainy | No |
| 1 | Rainy | No |
| 2 | Overcast | Yes |
| 3 | Sunny | Yes |
| 4 | Sunny | Yes |
| 5 | Sunny | No |
| 6 | Overcast | Yes |
| 7 | Rainy | No |
| 8 | Rainy | Yes |
| 9 | Sunny | Yes |
| 10 | Rainy | Yes |
| 11 | Overcast | Yes |
| 12 | Overcast | Yes |
| 13 | Sunny | No |

1. Cho bảng dữ liệu thu thập về tình hình thời tiết và quyết định đi chơi.

2. Hãy xây dựng mô hình Bayes Classifier

3. Hãy dùng mô hình Bayes để dự báo nếu hôm nay thời tiết u ám thì có đi chơi không.

## The Dataset

| Over 170cm | Eye | Hair length | Sex |
|---|---|---|---|
| No | Blue | Short | Male |
| Yes | Brown | Long | Female |
| No | Blue | Long | Female |
| No | Blue | Long | Female |
| Yes | Brown | Short | Male |
| No | Blue | Long | Female |
| Yes | Brown | Short | Female |
| Yes | Blue | Long | Male |

# How to deal with multiple attributes?

So far we have only considered Bayes Classification when we have one attribute (the "antennae length", or the "name"). But we may have many features.

Ex: Height, Eye Color, Hair Length, and so on.

How do we use all the features?

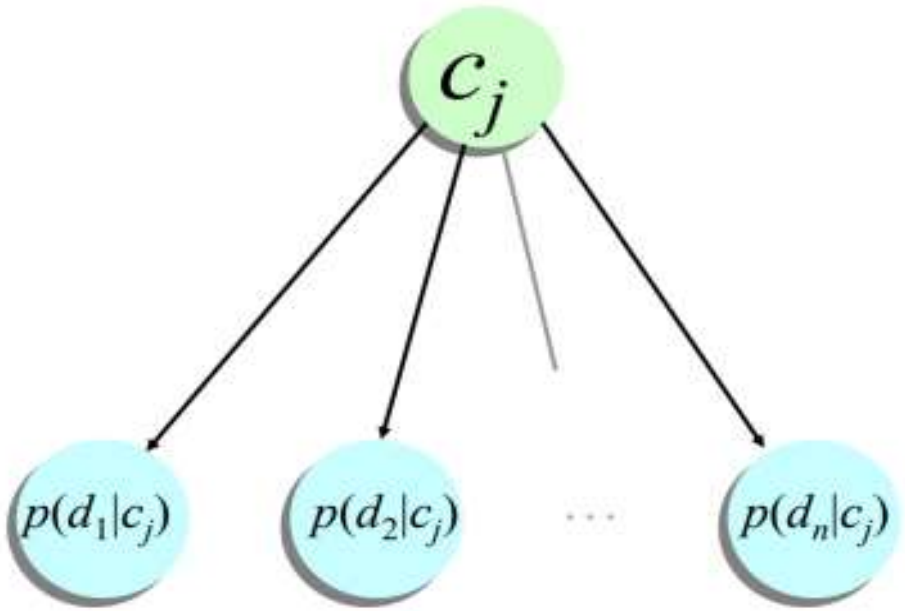
$$P(\text{male} \mid \text{Height, Eye, Hair Length}) \propto P(\text{Height, Eye, Hair Length} \mid \text{male})\ P(\text{male})$$

Computing Probability P(Height, Eye, Hair Length | male) is **infeasible**!

# Naïve Bayes Classification

Assume all input features are class conditionally independent!

P(male| Height, Eye, Hair Length) ∝ P(Height| male) P(Eye|male) P(Hair Length | male)  P(male)




Note: Direction of arrow from class to feature

How to choose class

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

## The Dataset

| Name | Over 170cm | Eye | Hair length | Sex |
|---|---|---|---|---|
| Drew | No | Blue | Short | Male |
| Claudia | Yes | Brown | Long | Female |
| Drew | No | Blue | Long | Female |
| Drew | No | Blue | Long | Female |
| Alberto | Yes | Brown | Short | Male |
| Karin | No | Blue | Long | Female |
| Nina | Yes | Brown | Short | Female |
| Sergio | Yes | Blue | Long | Male |

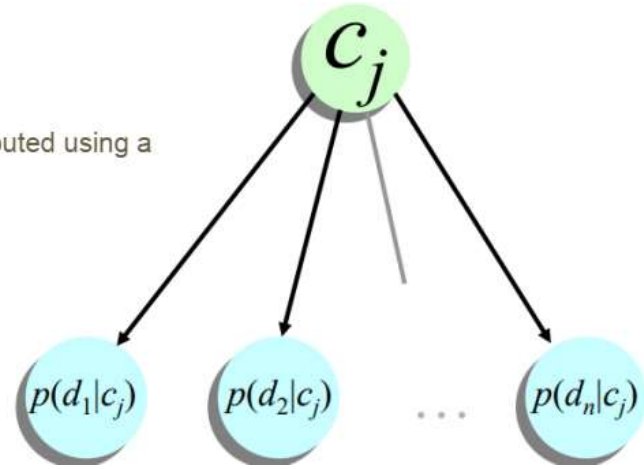
1. Cho bảng dữ liệu thu thập về name, height hơn 170 cm, eye color, hair length

2. Hãy xây dựng mô hình Bayes Classifier

3. Hãy dùng mô hình Bayes để dự báo nếu tôi biết thông tin : chiều cao trên 170cm, mắt nâu và tóc ngắn thì người đó sẽ có giới tính nào.

Naive Bayes is fast and space efficient

The conditional probability tables can be computed using a single pass over the data



| | Outlook | Temperature | Humidity | Windy | Play Golf |
|---|---|---|---|---|---|
| 0 | Rainy | Hot | High | False | No |
| 1 | Rainy | Hot | High | True | No |
| 2 | Overcast | Hot | High | False | Yes |
| 3 | Sunny | Mild | High | False | Yes |
| 4 | Sunny | Cool | Normal | False | Yes |
| 5 | Sunny | Cool | Normal | True | No |
| 6 | Overcast | Cool | Normal | True | Yes |
| 7 | Rainy | Mild | High | False | No |
| 8 | Rainy | Cool | Normal | False | Yes |
| 9 | Sunny | Mild | Normal | False | Yes |
| 10 | Rainy | Mild | Normal | True | Yes |
| 11 | Overcast | Mild | High | True | Yes |
| 12 | Overcast | Hot | Normal | False | Yes |
| 13 | Sunny | Mild | High | True | No |

1. Cho bảng dữ liệu thu thập về tình hình thời tiết và quyết định đi chơi.

2. Hãy xây dựng mô hình Bayes Classifier

3. Hãy dùng mô hình Bayes để dự báo nếu hôm nay: thời tiết u ám, trời nóng, độ ẩm cao và không có gió thì có đi chơi không.

# Evaluating Naive Bayes Classifier

- **Advantages**

  Fast to train (just one scan of database) and classify

  Not sensitive to irrelevant features

  Handles discrete data well

  Handles streaming data well

- **Disadvantage**

  Assumes independence of features - Losing the accuracy.

  Not really good for continuous attribute values.

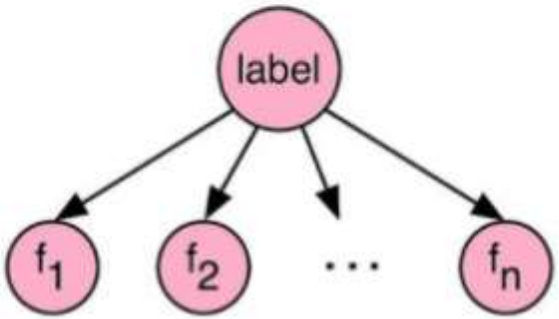
# Issues with Naive Bayes Classifier - 1

**ZERO CONDITIONAL PROBABILITY PROBLEM**

- Incomplete training data
- A given class and feature value never occur together in the training set.
- This is problematic since it will wipe out all information in the other probabilities when they are multiplied.
- Conditional probability will be zero and the whole construction collapses!

# Issues with Naive Bayes Classifier - 1

## ZERO CONDITIONAL PROBABILITY PROBLEM




**Example:** A feature **F2** and the class **label** do not occur in training set.

Since, P(F2 | label) = 0

P(label | F1, F2, F3) = P(label) P(F1|label) P(F2|label) P(F3|label)

**P(label | F1, F2, F3) = 0**

# Correction to Zero Probability Problem

## Laplace Smoothing

- To eliminate zeros joint probability, we use **add-one or Laplace smoothing**
- Adds arbitrary low probabilities in such cases so that the probability computation does not become zero.
- **Basic Idea: Pretend that you saw every feature-class outcome pair k extra times.**

# Laplace Smoothing

$X_i$ = The i-th attribute in dataset D.

$x_i$ = A particular value of the $X_i$ attribute in dataset D.

N = Total number of tuples in dataset D.

k = Laplace Smoothing Factor.

Count ($X_i = x_i$) = Number of tuples where the attribute $X_i$ takes the value $x_i$

$|X_i|$ = Number of different values attribute $X_i$ can take.

$$P_{Lap,k}(X_i = x_i) = \frac{count(X_i = x_i) + k}{N + k|X_i|}$$

# Laplace Smoothing

Count ($X_i = x_i$ , Y =y) = Joint probability of $X_i = x_i$ and Y =y appearing together in the dataset.

$|X_i|$ = Number of different values attribute $X_i$ can take.

$$P_{Lap,k}(X_i = x_i|Y = y) = \frac{count(X_i = x_i, Y = y) + k}{count(Y = y) + k|X_i|}$$

## The Dataset

| Name | Over 170cm | Eye | Hair length | Sex |
|---|---|---|---|---|
| Drew | No | Blue | Short | Male |
| Claudia | Yes | Brown | Long | Female |
| Drew | No | Blue | Long | Female |
| Drew | No | Blue | Long | Female |
| Alberto | Yes | Brown | Short | Male |
| Karin | No | Blue | Long | Female |
| Nina | Yes | Brown | Short | Female |
| Sergio | Yes | Blue | Long | Male |

Sử dụng Laplace Smoothing với k = 1 để thực hiện

1. Cho bảng dữ liệu thu thập về name, height hơn 170 cm, eye color, hair length

2. Hãy xây dựng mô hình Bayes Classifier

3. Hãy dùng mô hình Bayes để dự báo nếu tôi biết thông tin : chiều cao trên 170cm, mắt nâu và tóc ngắn thì người đó sẽ có giới tính nào.

# Issues with Naive Bayes Classifier - 2

**CONTINUOUS VARIABLES**

- When an attribute is continuous, computing the probabilities by the traditional method of frequency counts is not possible.

***Solution ---- May lead to loss in classification accuracy***

**Discretization:** Convert the attribute values to discrete values - Binning

**Probability Density Functions:** To compute probability densities instead of actual probabilities.

# Concept of Probability Density Function (PDF)

- Finding P(X = x) for a continuous random variable X is not going to work.
- **Solution - Find the probability that X falls in some interval [a, b], i.e. find P(a <= X <= b) ------> PDF comes to the rescue.**

**$f_X(x)$ = Probability Density Function of attribute X.**

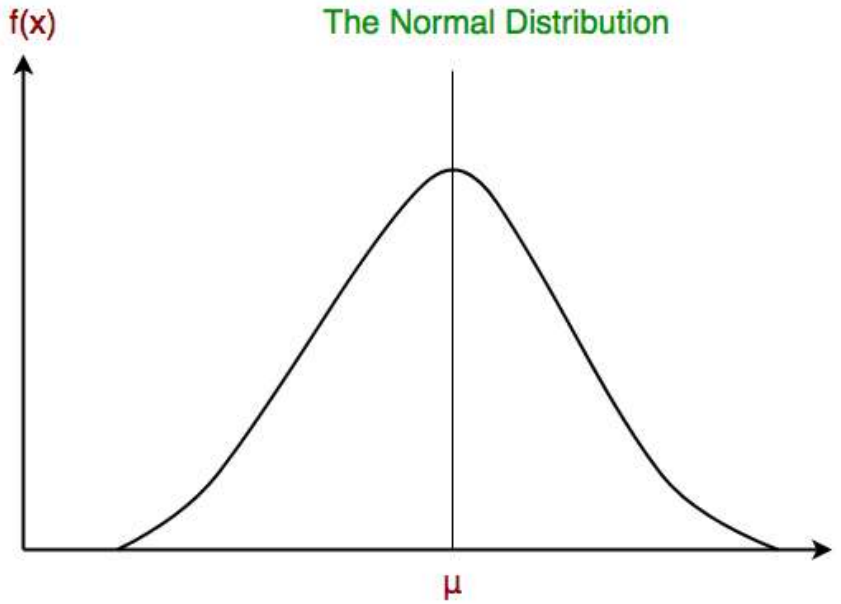
$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)\, dx.$$

**Gaussian Naive Bayes classifier**

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution.

**PDF equation:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$




$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

| ti | ci | drug |
|---|---|---|
| 5.1 | 3.5 | X |
| 4.9 | 3 | X |
| 4.7 | 3.2 | X |
| 4.6 | 3.1 | X |
| 7 | 3.2 | Y |
| 6.4 | 3.2 | Y |
| 6.9 | 3.1 | Y |
| 5.5 | 2.3 | Y |
| 6.3 | 2.5 | Z |
| 6.5 | 3 | Z |
| 6.2 | 3.4 | Z |
| 5.9 | 3 | Z |

Xây dựng mô hình Bayes Gaussian đề xuất cấp thuốc dựa trên chỉ số xét nghiệm máu: ti và ci

**Ví dụ: [ti, ci] = [`6.9, 3.1`] thì sẽ cấp thuốc loại gì**

THANK YOU