# Minería de Datos
# IIC2433

## Naïve Bayes

Vicente Domínguez

Basado en diapositivas del prof. Denis Parra

# ¿Qué vimos las clases pasadas?

- Random Forest

# ¿Qué veremos esta clase?

- Otra forma de clasificar: Naïve Bayes

# Probabilidades condicionales y conjuntas

*Ejemplo*


Al tirar una moneda

- ¿Cuál es la probabilidad de que salga sello

- ¿Cuál el es la probabilidad de que salga dos veces sello al tirarla dos veces?

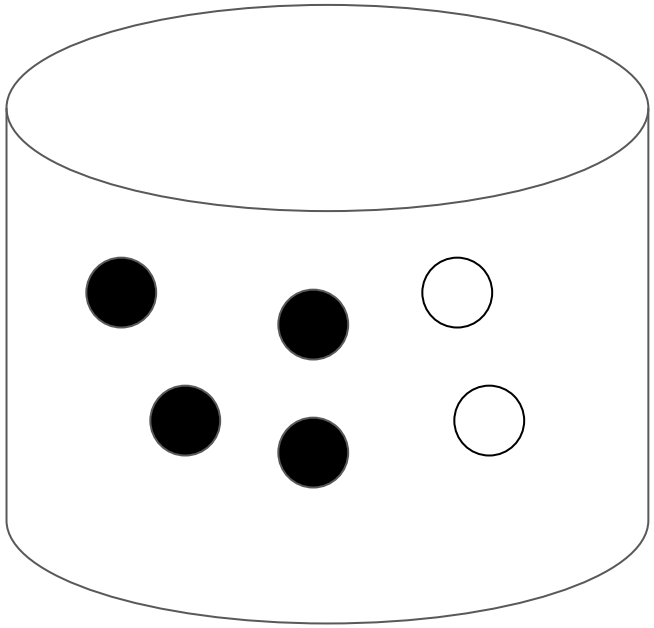
# Probabilidad conjunta con eventos **independientes**

$$P(A, B) = P(A)*P(B)$$

$$P(A, B, C) = P(A)*P(B)*P(C)$$

# Probabilidades condicionales y conjuntas

*Ejemplo*



Tengo 4 bolitas negras y 2 blancas en una tómbola,

- Al sacar una bolita al azar, ¿Cuál es la probabilidad de que salga una blanca?

- Al sacar dos bolitas al azar, ¿Cuál es la probabilidad de que ambas salgan blancas?

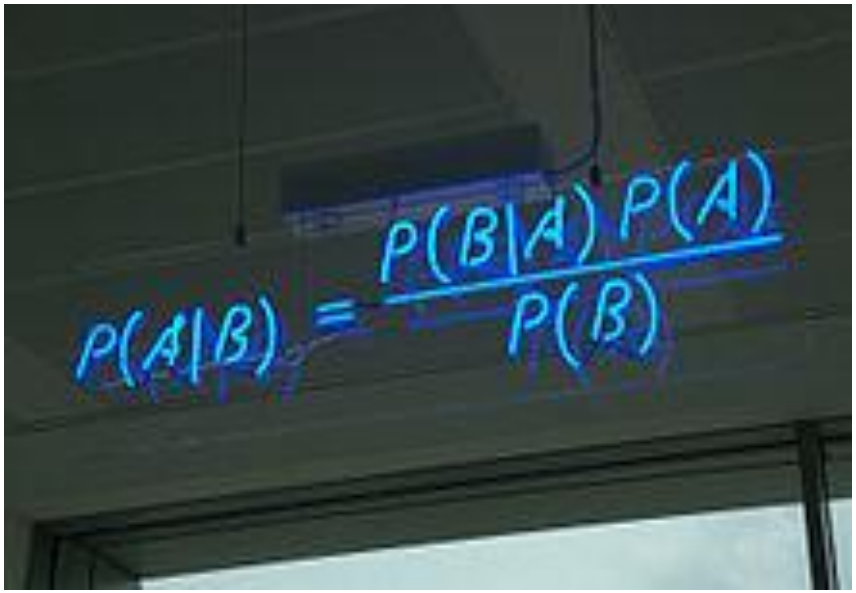
Probabilidad conjunta
con eventos **dependientes**

$$P(A, B) = P(A|B)*P(B) = P(B|A)*P(A)$$

$$P(A, B, C) = P(A|B,C)*P(B|C)*P(C)$$

# Teorema de Bayes

Thomas Bayes (1701 – 1761)



# Teorema de Bayes

- P(A = sí): Probabilidad del evento A sea “sí”
- P(A=sí|B=sí): Probabilidad de que el evento A sea “sí” DADO QUE el evento B fue “sí”
- Por simplicidad, usamos P(A) = P(A=“sí”)

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(A \mid B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

# Teorema de Bayes

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

**Prior**: Distribución de probabilidad a priori. El conocimiento de la probabilidad o incerteza de la clase antes de observar o condicionar los datos.

**Likelihood**: La probabilidad del evento bajo cierta clase o categoría, condicionada por los datos.

**Evidence**: Suma de las probabilidades del evento bajo todas las clases.

**Posterior**: Distribución de probabilidad condicional, que representa la probabilidad del evento condicionado luego de observar los datos.

# Noción del Teorema de Bayes

- << La riqueza hace la felicidad >>
- ¿Son felices los ricos? P(feliz = sí | rico = sí)
- … yo sé que de la gente feliz, 20% es rica.


- $$P(feliz \mid rico) = \frac{P(rico|feliz) * P(feliz)}{P(rico)}$$
- 20% no es tanto … por lo cual podemos concluir que la riqueza no hace la felicidad. ¿o no?

# Noción del Teorema de Bayes

- “La riqueza hace la felicidad”
- ¿Son felices los ricos? P(feliz = sí | rico = sí)
- … yo sé que de la gente feliz, 20% es rica.


**Supongamos:**
A: gente feliz = 40% de la población
B: gente rica = 10% de la población

$$P(feliz \mid rico) = \frac{P(rico|feliz) * P(feliz)}{P(rico)}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

# Noción del Teorema de Bayes

- "La riqueza hace la felicidad"
- ¿Son felices los ricos? P(feliz = sí | rico = sí)
- … yo sé que de la gente feliz, 20% es rica.


**Supongamos:**

A: gente feliz = 40% de la población

B: gente rica = 10% de la población

C: P(rico| feliz) = 20%

$$P(feliz \mid rico) = \frac{P(rico|feliz) * P(feliz)}{P(rico)}$$

$$P(feliz \mid rico) = \frac{P(rico|feliz) * P(feliz)}{P(rico)}$$

$$P(feliz \mid rico) = \frac{0.2 * 0.4}{0.1}$$

= 0.8 = **80%**

# ¿Y por qué se llama “Naïve” ?

- Naïve significa “**ingenuo**”
- Es “ingenuo” por que asume independencia de los eventos*

- * en realidad, asume independencia condicional

| Manzana | Carne | Pastel | ¿Alergia? |
|---|---|---|---|
| No | Sí | No | Sí |
| No | Sí | Sí | Sí |
| No | Sí | No | Sí |
| Sí | Sí | Sí | Sí |
| Sí | Sí | No | No |
| No | No | Sí | No |
| Sí | No | No | No |
| No | No | No | No |

- ¿Cuál es la probabilidad de haber consumido el alimento Manzana, dado que hubo alergia, es decir P(Manzana=Si|Alergia=Si)? ¿Cuál es la probabilidad de haber consumido el alimento pastel, dado que no hubo alergia, es decir, P(Pastel=Si|Alergia=No) ?

$$P(M|A) = (P(A|M) * P(M)) / P(A) = (1/3 * 3/8 )/ 1/2 = 1/4$$

$$P(P|\text{No } A) = (P(\text{No } A|P) * P(P)) / P(\text{No } A) =(1/3 * 3/8)/ 1/2 = 1/4$$

# Volviendo: Ejemplo de Clasificación

- Consideremos un auto SUV, color rojo, doméstico. ¿La probabilidad de que la roben es mayor o menor de que no la roben?

| Example No. | Color | Type | Origin | Stolen? |
|---|---|---|---|---|
| 1 | Red | Sports | Domestic | Yes |
| 2 | Red | Sports | Domestic | No |
| 3 | Red | Sports | Domestic | Yes |
| 4 | Yellow | Sports | Domestic | No |
| 5 | Yellow | Sports | Imported | Yes |
| 6 | Yellow | SUV | Imported | No |
| 7 | Yellow | SUV | Imported | Yes |
| 8 | Yellow | SUV | Domestic | No |
| 9 | Red | SUV | Imported | No |
| 10 | Red | Sports | Imported | Yes |

| Example No. | Color | Type | Origin | Stolen? |
|---|---|---|---|---|
| 1 | Red | Sports | Domestic | Yes |
| 2 | Red | Sports | Domestic | No |
| 3 | Red | Sports | Domestic | Yes |
| 4 | Yellow | Sports | Domestic | No |
| 5 | Yellow | Sports | Imported | Yes |
| 6 | Yellow | SUV | Imported | No |
| 7 | Yellow | SUV | Imported | Yes |
| 8 | Yellow | SUV | Domestic | No |
| 9 | Red | SUV | Imported | No |
| 10 | Red | Sports | Imported | Yes |

P(Robo | Red, SUV, Domestic) = Posterior

Rojo SUV Domestic

P(Robo) = (p(Robo)p(Color|Robo)*p(Tipo|Robo)*p(Origen|Robo))/N

Prior Likelihood

N = p(Robo)p(Color|Robo)*p(Tipo|Robo)*p(Origen|Robo)
+
p(No Robo)p(Color|No Robo)*p(Tipo|No Robo)*p(Origen|No Robo)

Evidence

| Example No. | Color | Type | Origin | Stolen? |
|---|---|---|---|---|
| 1 | Red | Sports | Domestic | Yes |
| 2 | Red | Sports | Domestic | No |
| 3 | Red | Sports | Domestic | Yes |
| 4 | Yellow | Sports | Domestic | No |
| 5 | Yellow | Sports | Imported | Yes |
| 6 | Yellow | SUV | Imported | No |
| 7 | Yellow | SUV | Imported | Yes |
| 8 | Yellow | SUV | Domestic | No |
| 9 | Red | SUV | Imported | No |
| 10 | Red | Sports | Imported | Yes |

P(Robo | Red, SUV, Domestic)

Rojo SUV Domestic

P(Robo) = (p(Robo)p(Color|Robo)*p(Tipo|Robo)*p(Origen|Robo))/N

N = p(Robo)p(Color|Robo)*p(Tipo|Robo)*p(Origen|Robo)

+

p(No Robo)p(Color|No Robo)*p(Tipo|No Robo)*p(Origen|No Robo)

P(Robo) = (5/10 * 3/5 * 1/5 * 2/5)/((5/10 * 3/5 * 1/5 * 2/5) + (5/10 *2/5 * 3/5 *3/5))

P(Robo) = 0.25

| Manzana | Carne | Pastel | ¿Alergia? |
|---|---|---|---|
| No | Sí | No | Sí |
| No | Sí | Sí | Sí |
| No | Sí | No | Sí |
| Sí | Sí | Sí | Sí |
| Sí | Sí | No | No |
| No | No | Sí | No |
| Sí | No | No | No |
| No | No | No | No |

| Manzana | Carne | Pastel | ¿Alergia? |
|---|---|---|---|
| Sí | No | Sí | ?? |
| Sí | Sí | No | ?? |

- Basado en los datos de la Tabla 1, usando un clasificador Naive Bayes, clasifique los siguientes dos casos dados los datos de la Tabla 1.

# Ejemplo de Clasificación Numérico

- ¿Qué ocurrirá en el caso con datos numéricos y no categóricos?

# Ejemplo de Clasificación Numérico

- ¿Qué ocurrirá en el caso con datos numéricos y no categóricos?
- R: En general se asume que los datos distribuyen como una gaussiana (puede ser otra distribución) y se calculan sus parámetros acorde a los datos.
- Luego, se utilizan para utilizar la función de densidad para calcular un estimado de la probabilidad del dato a predecir.

# Referencias

- Material de Tom M. Mitchell, CMU:

http://www.cs.cmu.edu/~awm/15781/slides/NBayes-9-27-05.pdf

http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf