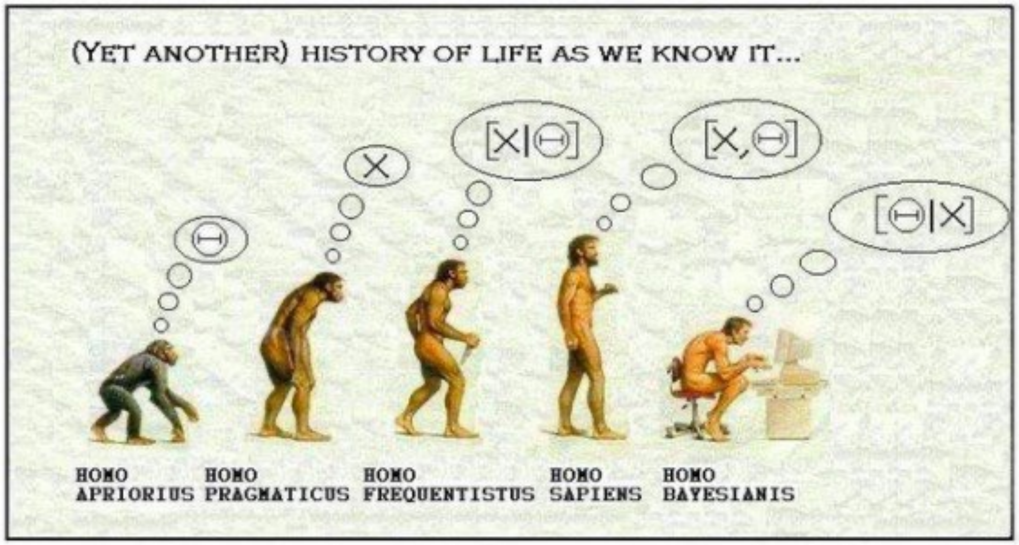


NAÏVE BAYES

UNIVERSIDAD ICESI

PROBABILIDAD BAYESIANA

La estimación de la probabilidad de un evento, o un resultado potencial, debe basarse en la evidencia dada por múltiples ensayos u oportunidades para que ocurra el evento

Los métodos bayesianos proporcionan información sobre cómo la probabilidad de estos eventos puede ser estimada a partir de los datos observados

Los principios básicos de probabilidad se usan transversalmente en el algoritmo **Naïve Bayes**

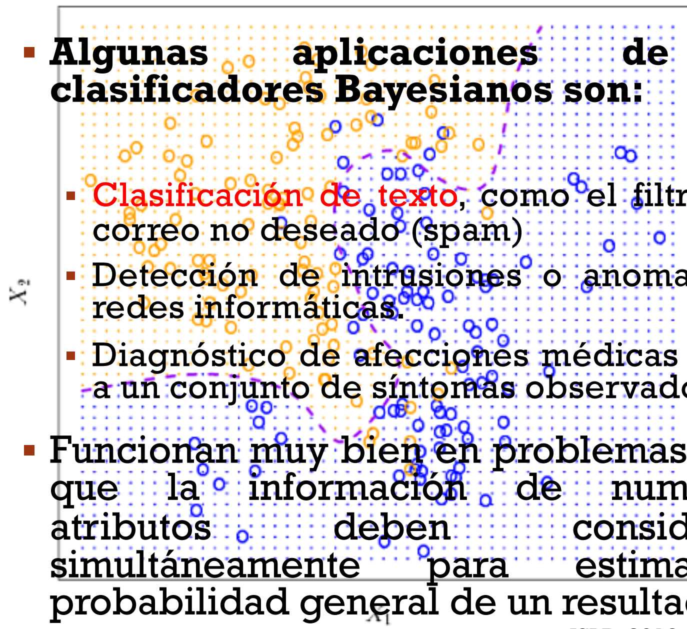
UNIVERSIDAD ICESI

CLASIFICADORES BAYESIANOS

- **Los clasificadores bayesianos** asignan cada observación a la clase j más probable, dados los valores observados de sus variables predictivas:

$$argmax_j \ p(Y = y_j | X = x_{observados})$$

- Si se conocen las distribuciones de probabilidad, el clasificador resultante da la frontera de separación óptima en términos de error
- No siempre se tienen las probabilidades condicionales necesarias.
- **Naïve Bayes** es un algoritmo basado en el Teorema de Bayes



- **Algunas aplicaciones de los clasificadores Bayesianos son:**
  - Clasificación de texto, como el filtrado de correo no deseado (spam)
  - Detección de intrusiones o anomalías en redes informáticas.
  - Diagnóstico de afecciones médicas debido a un conjunto de síntomas observados.
- Funcionan muy bien en problemas en los que la información de numerosos atributos deben considerarse simultáneamente para estimar la probabilidad general de un resultado

ISLR, 2013

NAIVE BAYES (BAYES INGENUO)

Probabilidad a posteriori

Probabilidad a priori

Verosimilitud

Teorema de Bayes: $p(y_j|x_1, x_2, \dots, x_n) = \frac{p(y_j, x_1, x_2, \dots, x_n)}{p(x_1, x_2, \dots, x_n)} = \frac{p(y_j) * p(x_1, x_2, \dots, x_n|y_j)}{p(x_1, x_2, \dots, x_n)}$

Verosimilitud marginal

- El denominador es solo usado para propósitos de normalización (suma de probabilidades = 1)

$$p(x_1, x_2, \dots, x_n) = \sum_j p(y_j) * p(x_1, x_2, \dots, x_n|y_j)$$

- Por ello solo nos fijamos en el numerador:

$$p(y_j, x_1, x_2, \dots, x_n) = p(y_j) * p(x_1|y_j) * p(x_2|x_1, y_j) * p(x_3|x_2, x_1, y_j) * \cdots * p(x_n|x_{1:n-1}, y_j)$$

- Si asumimos ingenuamente (**naïvely**) que todas las variables predictivas x_i son independientes condicionalmente con respecto a la clase y_j [1] entonces el numerador se simplifica a:

$$p(y_j) * p(x_1|y_j) * p(x_2|y_j) * p(x_3|y_j) * \cdots * p(x_n|y_j)$$
$$= p(y_j) \prod_{i=1}^{n} p(x_i|y_j)$$

1 optimalidad de NB On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 1997; 29:103-130.

NAÏVE BAYES (BAYES INGENUO)

- La regla de clasificación es:

$$\operatorname{argmax}_j \; p(y_j) \prod_{i=1}^{n} p\left(x_i \middle| y_j\right)$$

- **Sólo necesitamos especificar :**
  - → Las probabilidades a priori de cada clase
  - → Las distribuciones de probabilidad de las variables predictivas para cada clase (condicionadas a la clase)
- Esta información se constituye en los **parámetros** del modelo, y en el caso de variables categóricas se obtienen a partir de tablas de frecuencias (conteos)

UNIVERSIDAD ICESI

NAIVE BAYES

Ejemplo: Un banco quiere predecir si un cliente va a adquirir un CDT.

Creamos un clasificador Naïve Baye a partir de los datos históricos para calcular las probabilidades posteriores de cada clase: subscribed=yes and subscribed=no.

{Single, Married, Divorced}

{Subscribed Yes, Subscribed No}

$$p(y_j|x_1,\dots,x_n) = argmax_j\ p(\boxed{y_j}) \prod_{i=1}^{n} p(\boxed{x}|\boxed{y_j})$$

Marital	Subscribed=yes
Single	35%
Married	53%
Divorced	12%

Marital	Subscribed=no
Single	28%
Married	61%
Divorced	11%

Subscribed=yes	12%

Subscribed=no	88%

¿Debería el banco ofrecerle un CDT al cliente con la información siguiente?

Job=Management
Marital=Married
Education=Secondary
Default=no
Housing=yes
Loan=no
Contact=Cellular
Outcome=Success

Suponga que se disponen de las probabilidades condicionales para todas las variables predictivas (ya ilustradas para el estado civil "Marital")

NAIVE BAYES

Ejemplo: Un banco quiere predecir si un cliente va a adquirir un CDT.

Creamos un clasificador Naïve Bayes a partir de los datos históricos para calcular las probabilidades posteriores para cada clase: subscribed=yes and subscribed=no.

$$p(y_j|x_1, \dots, x_n) = argmax_j \; p(y_j) \prod_{i=1}^{n} p(x_i|y_j)$$

Marital	Subscribed=yes
Single	35%
Married	53%
Divorced	12%

Marital	Subscribed=no
Single	28%
Married	61%
Divorced	11%

Subscribed=yes	12%

Subscribed=no	88%

¿Debería el banco ofrecerle un CDT al cliente con la información siguiente?

	Subscribed=yes	Subscribed=no
Job=Management	22%	21%
Marital=Married	53%	61%
Education=Secondary	46%	51%
Default=no	99%	98%
Housing=yes	35%	57%
Loan=no	90%	85%
Contact=Cellular	85%	62%
Outcome=Success	15%	1%
Priors	12%	88%

Numerador	0,000255914	0,000169244
Proba posterior	60%	40%

NAÏVE BAYES (BAYES INGENUO)

- ¿Qué pasa si algunos de los valores de las variables predictivas tienen frecuencia nula con respecto a las categorías de la clase? ¿cuáles serían sus probabilidades aposteriori asociadas?
- Para evitar este problema, se utilizan métodos de **suavización**.
  - Por ejemplo, al contar las frecuencias de ocurrencia de cada valor se les agrega un valor pequeño, $\boldsymbol{\varepsilon} > \mathbf{0}$, evitando que alguna probabilidad sea cero:

$$P(casado|cliente\ potencial) = \frac{Conteo(casado,\ cliente\ potencial) + \varepsilon}{Conteo(cliente\ potencial) + N(x) * \varepsilon}$$

- El método de suavización de **Laplace** se aplica usualmente con ε=1, otro valor puede ser 1/n donde n es el número de datos de entrenamiento.

NAÏVE BAYES (BAYES INGENUO)

- Cuando las variables predictivas no son categóricas (e.g. numéricas), es necesario establecer una distribución de probabilidad:
  1. Se puede discretizar (en compartimentos) la variable convirtiéndola en categórica.
  2. Se establece una distribución de probabilidad empírica utilizando KNN,

$$P(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in \aleph_0} I(y_i = j)$$

  3. Se supone que se trata de un tipo de distribución de probabilidad y se utiliza su función de densidad.
     - Por ejemplo, si se supone la variable sigue una **distribución normal** condicionada a la categoría objetivo, se puede calcular la media μ y desviación estándar σ a partir de los datos históricos, y utilizar la función de densidad:

$$P(edad|cliente\ potencial) = \frac{1}{\sigma_{edad|cliente}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{edad-\mu_{edad|cliente}}{\sigma_{edad|cliente}}\right)^2}$$

UNIVERSIDAD ICESI

NAÏVE BAYES (BAYES INGENUO)

Pros:

- **Simple, rápido** y muy **efectivo**, permite atributos tanto categóricos como numéricos
- Estima efectivamente **las probabilidades condicionales** con respecto a los valores de la categoría objetivo
- Trabaja bien con atributos categóricos, con **valores faltantes** y con ruido
- Resistente al **overfitting**, sobretodo si se incluye un suavizador (e.g. Laplace)
- Trabaja bien con muestras de entrenamiento pequeñas y también con grandes

Contras:

- Sólo se puede utilizar para **clasificación**
- Se basa en **suposiciones** muy fuertes
- **Muy sensible** a atributos correlacionados (considera varias veces los mismos efectos)
- Las probabilidades estimadas son menos confiables que las clases predichas