

---

Inteligencia Artificial

# Aprendizaje Estadístico

---

Jorge Luis Guevara Diaz

[www.jorge.sistemasyservidores.com](http://www.jorge.sistemasyservidores.com)

---

# Que veremos?

- Aprendizaje Bayesiano
- Aprendizaje MAP y aprendizaje ML
- Aprendizaje en redes bayesianas
  - Aprendizaje de parámetros ML con datos completos
  - Regresión lineal

---

# Aprendizaje Estadístico

- Conceptos claves

- Datos

- Son la evidencia, instancia de algunas o de todas las variables aleatorias del dominio

- Hipótesis

- Teoría probabilística de cómo el dominio trabaja

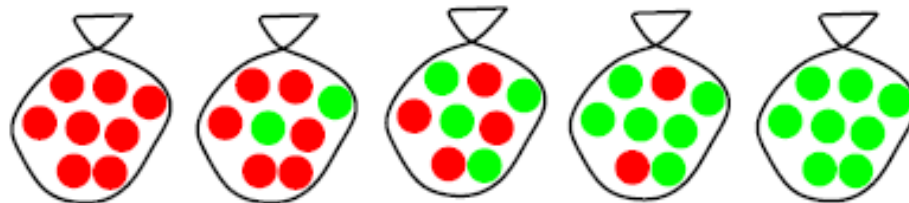
- Ejemplo

- Caramelos de dos sabores : cherry y lima
    - Los dos tienen la misma envoltura
    - Son vendidos en paquetes de los cuales se conocen 5 tipos
-

# Aprendizaje Estadístico

## ■ Paquetes de 5 tipos de caramelos

- 10% son  $h_1$  : 100% cherry
- 20% son  $h_2$  : 75% cherry + 25% lima
- 40% son  $h_3$  : 50% cherry + 50% lima
- 20% son  $h_4$  : 25% cherry + 75% lima
- 10% son  $h_5$  : 100% lima



Si realizamos la siguiente observación 

Que tipo de paquete podrá ser?

Cual será el sabor del siguiente caramelo?

# Aprendizaje Bayesiano

- Se ve el aprendizaje como la actualización bayesiana de una distribución de probabilidad sobre el espacio de hipótesis
- $H$  es la variable hipótesis, y los valores  $h_1, h_2, \dots$  son los valores a priori de  $P(H)$
- $D_j$  es la  $j^{\text{th}}$  salida de la variable aleatoria  $D$
- Los datos de entrenamiento son  $d = d_1, \dots, d_N$
- Dado los datos, cada hipótesis tiene una probabilidad a posteriori
  - $P(h_i | d) = \alpha P(d | h_i) P(h_i)$        $P(d | h_i) = \text{likelihood}$
- Las predicciones usan un promedio del likelihood ponderado sobre las hipótesis
  - $P(X | d) = \sum_j P(X | d, h_j) P(h_j | d)$

---

# Aprendizaje Bayesiano

- Las hipótesis son intermediarias entre los datos y las predicciones
- Las cantidades claves son:
  - La probabilidad a priori de las hipótesis  $P(h_i)$
  - El likelihood de los datos bajo la hipótesis  $P(d|h_i)$

---

# Aprendizaje Bayesiano

- Ejemplo:

- $P(H) = [0.1, 0.2, 0.4, 0.2, 0.1]$

- Likelihood?  $P(\mathbf{d}|h_i) = \prod_j P(d_j|h_i)$

- Suponer que el paquete es del tipo  $h_5$

- $d = d_1 \dots d_{10} = 10$  caramelos de lima

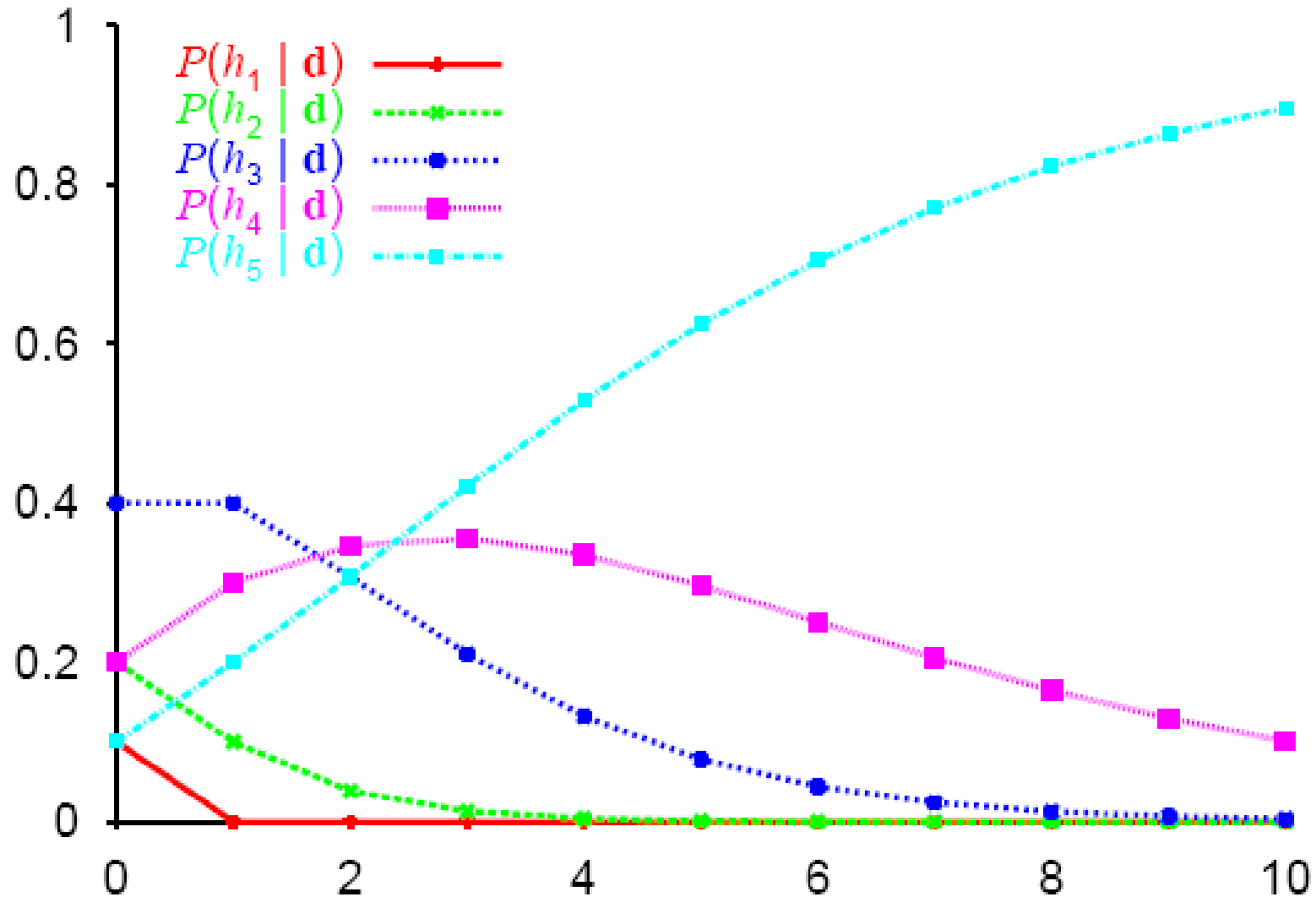
- $P(d|h_3) = 0.5^{10}$

# Aprendizaje Bayesiano

## ■ Ejemplo:

- $P(H) = [0.1, 0.2, 0.4, 0.2, 0.1]$
- Para  $P(h_1 \setminus d = \{\text{lima}\}) = P(d = \{\text{lima}\} \setminus h_1) * p(h_1)$ 
  - $\alpha * 0 * 0.1 = 0$
- Para  $P(h_2 \setminus d = \{\text{lima}\}) = P(d = \{\text{lima}\} \setminus h_2) * p(h_2)$ 
  - $\alpha * 0.25 * 0.2 = 0.05$
- Para  $P(h_3 \setminus d = \{\text{lima}\}) = P(d = \{\text{lima}\} \setminus h_3) * p(h_3)$ 
  - $\alpha * 0.5 * 0.4 = 0.2$
- Para  $P(h_4 \setminus d = \{\text{lima}\}) = P(d = \{\text{lima}\} \setminus h_4) * p(h_4)$ 
  - $\alpha * 0.75 * 0.2 = 0.15$
- Para  $P(h_5 \setminus d = \{\text{lima}\}) = P(d = \{\text{lima}\} \setminus h_5) * p(h_5)$ 
  - $\alpha * 1.00 * 0.1 = 0.1$
- $\alpha = 0 + 0.05 + 0.2 + 0.15 + 0.1 = 0.5$
- $P(H \setminus d = \{\text{lima}\}) = [0, 0.1, 0.4, 0.3, 0.2]$
- $P(H \setminus d = \{\text{lima}, \text{lima}\})?$

# Aprendizaje Bayesiano: probabilidad a posteriori de las hipótesis

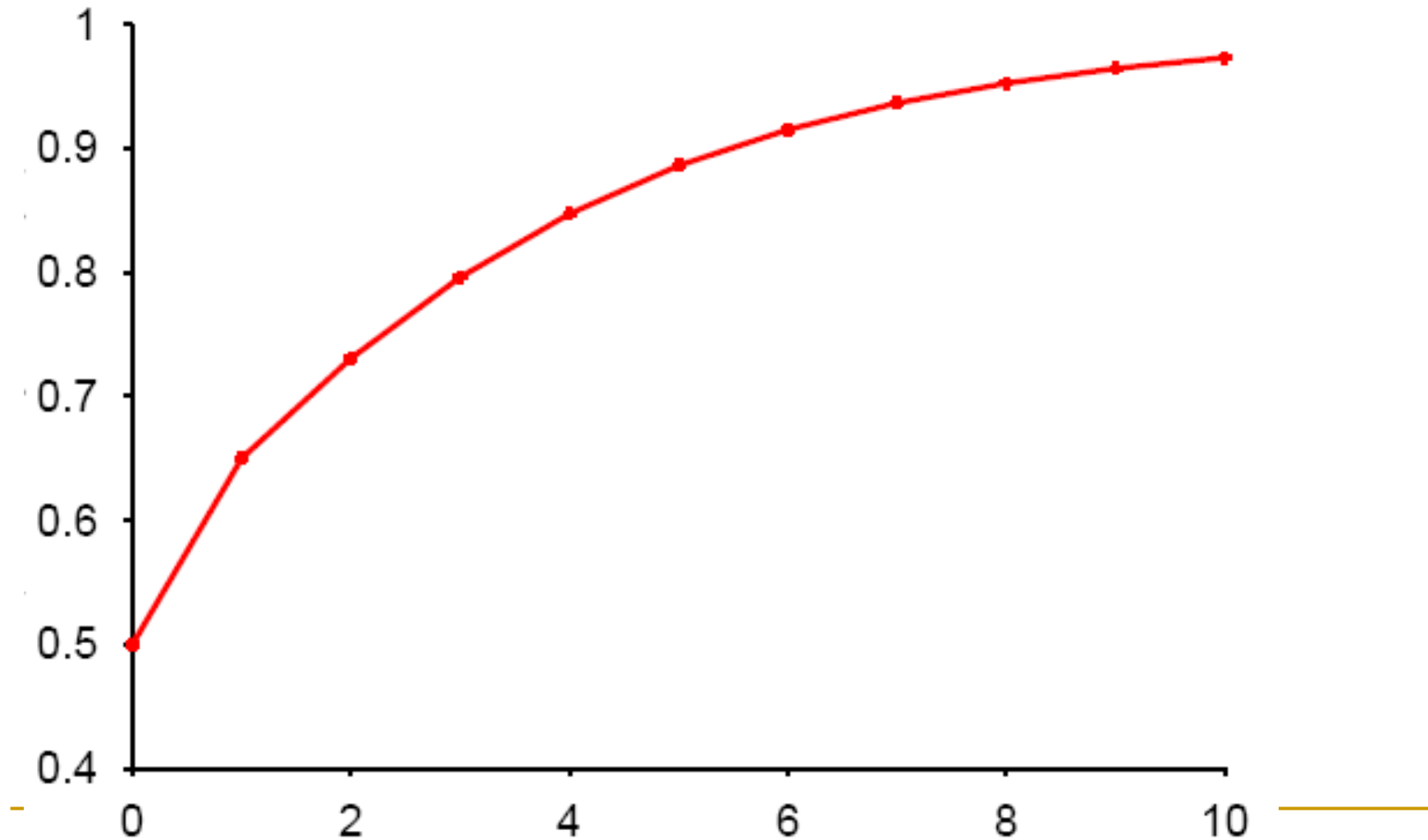


# Aprendizaje Bayesiano

$$P(X|d) = \sum_i P(X|h_i)P(h_i|d)$$

- $P(d_{n+1}=\{\text{lima}\}|d=\{\text{lima}\}) =$ 
  - $\sum_i P(d_{n+1}=\{\text{lima}\}, h_i)P(h_i|d=\{\text{lima}\})$
- $P(d_{n+1}=\{\text{lima}\}|d=\{\text{lima}\}) =$ 
  - $0*0 + 0.25*0.1 + 0.5*0.4 + 0.75*0.3 + 1*0.2 = 0.65$
- $P(d_{n+1}=\{\text{lima}\}|d=\{\text{lima, lima}\}) = ?$

# Aprendizaje Bayesiano: probabilidad de la predicción



# Algoritmo aprendizaje MAP

- Sumar todo el espacio de hipótesis es intratable
- Aprendizaje MAP (maximun a posteriori) escoge  $h_{MAP}$  que maximiza  $P(h_i|d)$ 
  - Maximizar  $P(d|h_i)P(h_i)$  ó
  - Minimizar  $-\log P(d|h_i) - \log P(h_i)$
- Ejemplo
  - $h_{MAP} = h_5$
  - $P(X|d) = P(X=lima|d) \approx P(X=lima|h_{MAP}) = P(X=lima|h_5) = 1$

---

# Algoritmo de aprendizaje ML

- Si se asume una distribución uniforme sobre el espacio de hipótesis, entonces **MAP** se reduce a escoger  $h_{ML}$  que maximiza el likelihood  $P(d|h_i)$  (**ML = maximum likelihood**)

---

# Aprendizaje ML de parametros en redes bayesianas con datos completos

- Buscar los valores numéricos para los parámetros de un modelo probabilístico fijo
- Ejemplo:
  - Encontrar las probabilidades condicionales en una red bayesiana con una estructura dada

Los datos son completos, si cada muestra contiene valores para cada variable en el modelo probabilístico

---

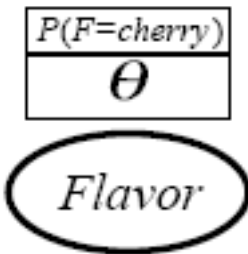
# Aprendizaje ML de parametros en redes bayesianas con datos completos

- Paquete de caramelos de un nuevo fabricante: proporcion desconocida de limas y cherrys
- Parámetro  $\Theta$  : proporcion de caramelos cherrys
- $1 - \Theta$  : proporcion de caramelos de lima
- Hipótesis continuas :  $h_{\Theta}$
- Asumir que el espacio de hipótesis es igualmente distribuido a priori: [algoritmo ML](#)

# Aprendizaje ML de parametros en redes bayesianas con datos completos

- Se extraen **N** caramelos : **c** cherrys y **l** limas

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^N P(d_j|h_\theta) = \theta^c \cdot (1 - \theta)^\ell$$



- Maximizando

$$L(\mathbf{d}|h_\theta) = \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^N \log P(d_j|h_\theta) = c \log \theta + \ell \log(1 - \theta)$$

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{\ell}{1 - \theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c + \ell} = \frac{c}{N}$$

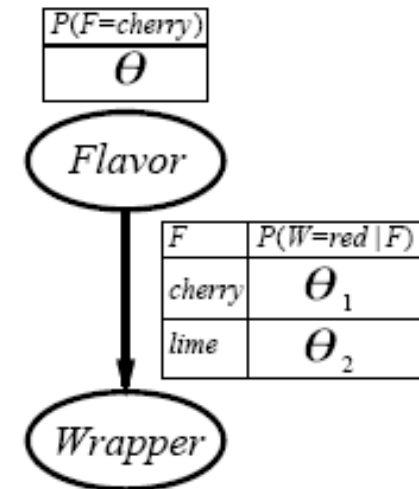
- Con pocos datos disponibles ciertos conteos pueden dar 0

# Aprendizaje ML de parametros en redes bayesianas con datos completos

- Multiples datos

$$P(\mathbf{d}|h_{\theta,\theta_1,\theta_2}) = \theta^c(1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}$$

$$\begin{aligned} L &= [c \log \theta + \ell \log(1-\theta)] \\ &+ [r_c \log \theta_1 + g_c \log(1-\theta_1)] \\ &+ [r_\ell \log \theta_2 + g_\ell \log(1-\theta_2)] \end{aligned}$$



# Aprendizaje ML de parametros en redes bayesianas con datos completos

- Tomando derivadas con respecto a cada parámetro

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c + g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell + g_\ell}$$

- Con datos completos los parámetros son aprendidos separadamente

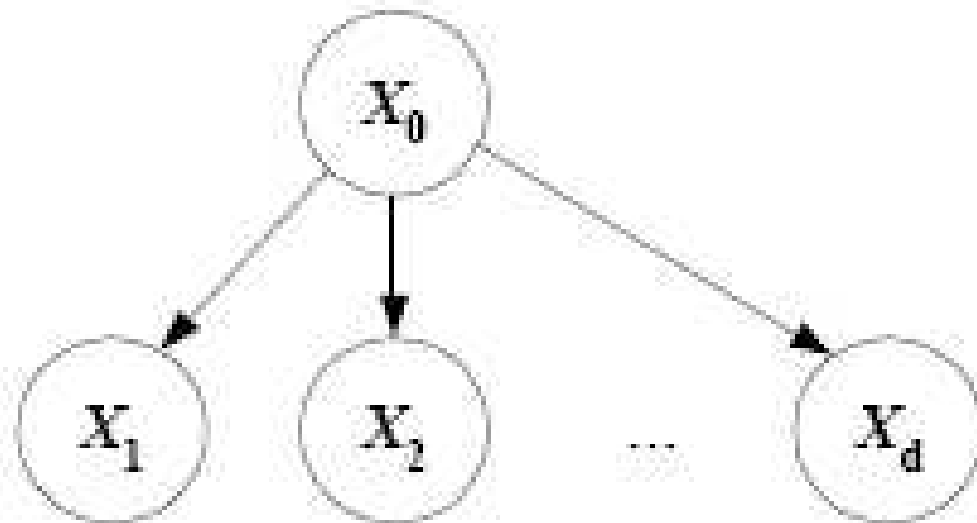
---

# Modelos Naive Bayes

- Es el modelo más común de redes bayesianas usados en machine learning
- Variable de clase  $X_0$  es la clase a ser predecida y va en el nodo raíz
- Variables atributos  $X_i$  son las hojas
- Se llama naive bayes, pues asume que los atributos son **condicionalmente independientes** de cada otro dada la variable de clase  $C$

---

# Modelos Naive Bayes



---

# Referencias Bibliográficas

- Capitulo 20 Artificial Intelligence: A Modern Approach, Russell and Norvig
- <http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/Uncertainty>