Razonamiento Basado en Probabilidades

Introducción

- **En la mayor parte de aplicaciones de S.E. la incertidumbre es lo común:**
  - Por ejemplo, en un diagnóstico médico la pregunta típica es:

    Dado que el paciente presenta un conjunto de síntomas, ¿cuál de las enfermedades posibles es la que tiene el paciente?
  - Esta situación implica cierto grado de incertidumbre puesto que:
    - Los hechos o datos pueden no ser conocidos con exactitud.
      - Hay un cierto grado de incertidumbre en la información asociada a cada paciente (subjetividad, imprecisión, ausencia de información, errores, etc.) Por ejemplo, un paciente no está seguro de haber tenido fiebre la noche anterior.
    - El conocimiento no es determinista.
      - Las relaciones entre las enfermedades y los síntomas no son deterministas. Por ejemplo, podemos encontrar dos pacientes con los mismos síntomas y enfermedades diferentes.

La teoría de la Probabilidad

- **Medida de la probabilidad**
  - Para medir la incertidumbre se parte de un conjunto S (**El espacio muestral**), en el que se incluyen todos los posibles resultados de un cierto experimento.
  - Una vez definido este conjunto, el objetivo consiste en asignar a todo subconjunto S un numero real que mida el grado de incertidumbre sobre su realización.
  - Para obtener medidas con significado claro y práctico, se imponen ciertas propiedades intuitivas adicionales que definen una clase de medidas que se conocen como **medidas de probabilidad.**

Medida de probabilidad

- **Definición 1. Medidas de Probabilidad**. Una función p que proyecta los sub-conjuntos de $S \supseteq A$ en el intervalo [0, 1] se llama medida de la probabilidad si satisface los siguientes axiomas:
  - **Axioma 1 (Normalización)**: $p(S) = 1$.
  - **Axioma 2 (Aditividad)**: Para cualquier sucesión infinita $A_1, A_2, \ldots$ de subconjuntos disjuntos de S, se cumple la igualdad:

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

- **Ejemplo probabilidades: Lanzamiento de un dado no trucado**
  - S = {1, 2, 3, 4, 5, 6}
    - p(S) = 1
    - p({1})= 1/6
    - p({1, 3}) = p({1})+ p({3}) = 1/3.

Distribución de probabilidad

- **Sea $\{X_1, \ldots X_n\}$ un conjunto de variables aleatorias discretas y $\{x_1, \ldots x_n\}$ el conjunto de sus posibles realizaciones.**
  - Por ejemplo, si X_i es una variable binaria x_i puede ser 0 o 1.
- **Se denomina función de** probabilidad conjunta **a**

  $p(x_1, \ldots x_n) = p(X_1 = x_1, \ldots X_n = x_n)$
- **Se denomina función de** probabilidad marginal **de la i-ésima variable mediante la fórmula:**

  $p(x_i) = p(X_i = x_i) = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} p(x_1, \ldots x_n)$
- **El conocimiento de la ocurrencia de un suceso puede modificar las probabilidades de otros sucesos.**
  - Por ejemplo, la probabilidad de que un paciente tenga una enfermedad dada puede cambiar tras el conocimiento de los resultados de un análisis de sangre. Esto conduce al concepto de **probabilidad condicional.**

Probabilidad condicional

- **Definición 2. Probabilidad condicional.** *Sean X e Y dos conjuntos disjuntos de variables tales que p(y) > 0 . Entonces, la probabilidad condicional (función de probabilidad condicionada) de X dado Y=y viene dada por*

  $p(X=x \mid Y=y) = p(x|y) = p(x,y)/p(y)$
  - Esta definición implica que la función de probabilidad conjunta de X e Y puede escribirse como

    $p(x,y) = p(y)p(x|y)$.
  - Se obtiene un caso particular cuando X es una única variable e Y es un subconjunto de variables:

    $p(x_i | x_1, \ldots, x_k) = p(x_i, x_1, \ldots, x_k)/p(x_1, \ldots, x_k)$

Dependencia e Independencia

- **Definición 3. Independencia de dos variables.** *Sean X e Y dos subconjuntos disjuntos del conjunto de variables aleatorias* $\{X_1, \ldots X_n\}$. *Entonces se dice que X es* ***independiente*** *de Y si y solamente si*

  $$p(x|y) = p(x),$$

  para todos los valores posibles de x e y de X e Y; en otro caso, X se dice ***dependiente*** *de Y.*

  Nótese que al si x e y son valores posibles de X e Y, entonces $p(x) > 0$ y $p(y) > 0$.

  - Si X es independiente de Y, entonces nuestro conocimiento de Y no afecta nuestro conocimiento sobre X, es decir, Y no tienen información sobre X.
  - Si X es independiente de Y entonces $p(x,y) = p(x)\ p(y)$.
  - Una propiedad importante de la relación de independencia es su simetría:

    $p(y|x) = p(x,y)\ /\ p(x) = p(x)\ p(y)\ /\ p(x) = p(y)$

Ejemplo: variables

- Sea A una persona elegida al azar de la población.
  - Sin conocer si la persona es fumadora, la probabilidad de que se trate de una mujer es p(A=mujer) = 0.50
  - Si se sabe que las persona es fumadora, esta probabilidad cambia de 0.50 a p(A=mujer | A = f) = 0.64.
  - Por lo tanto, se tiene que p(A=mujer | A = f) ° p (A = mujer) ; por lo que las variables *Sexo* y *Fumador* son dependientes.
  - ¿*Estado Civil* contiene información relevante sobre el *Sexo* ?

Independencia de variables

- **Definición 4. Independencia de un conjunto de variables.** *Las variables aleatorias* $\{X_1, \ldots X_n\}$ *se dice que son independientes si y sólo si*

  $$p(x_1, \ldots, x_m) = \prod_{i=1}^{m} p(x_i)$$

  para todos los valores posibles $x_1, \ldots, x_m$ *de* $X_1, \ldots, X_m$ *. En otro caso se dice que son dependientes.*

  - *Si* $X_1, \ldots, X_m$ *son condicionalmente independientes dado otro subconjunto* $Y_1, \ldots, Y_n$ *, entonces*

    $$p(x_1, \ldots, x_m | y_1, \ldots, y_n) = \prod_{i=1}^{m} p(x_i | y_1, \ldots, y_n)$$

  - Una implicación importante de la **independencia** es que no es rentable obtener información sobre variables independientes, pues es **irrelevante**. Es decir, independencia significa irrelevancia.

Ejemplo : Distribuciones de probabilidad

- Considérese la función de probabilidad conjunta de tres variables binarias:

x	y	z	p(x,y,z)
0	0	0	0.12
0	0	1	0.18
0	1	0	0.04
0	1	1	0.16
1	0	0	0.09
1	0	1	0.21
1	1	0	0.02
1	1	1	0.18

Í Función de probabilidad marginal de X:

$$p(X=0) = \sum_{y=0}^{1} \sum_{z=0}^{1} p(0,y,z) = 0.12+0.18+0.04+0.16 = 0.5$$

$$p(X=1) = \sum_{y=0}^{1} \sum_{z=0}^{1} p(1,y,z) = 0.09+0.21+0.02+0.18 = 0.5$$

Í Función de probabilidad conjunta de X e Y

$$p(X=0, Y=0) = \sum_{z=0}^{1} p(0,0,z) = 0.12+0.18 = 0.3$$

$$p(X=0, Y=1) = \sum_{z=0}^{1} p(0,1,z) = 0.04+0.16 = 0.2$$

$$p(X=1, Y=0) = \sum_{z=0}^{1} p(1,0,z) = 0.09+0.21 = 0.3$$

$$p(X=1, Y=1) = \sum_{z=0}^{1} p(1,1,z) = 0.18+0.02 = 0.2$$

Ejemplo: variables

- Considérense las siguientes características de las personas de una población
  - Sexo = {hombre, mujer}
  - Fumador = {si (f), no (¬ f)}
  - Estado civil = {casado (c), no casado (¬ c)}
  - Embarazo = {si (e), no (¬ e)}
- Función de probabilidad conjunta

		h	h	m	m
		f	¬ f	f	¬ f
c	e	0.00	0.00	0.01	0.05
	¬ e	0.02	0.18	0.04	0.10
¬ c	e	0.00	0.00	0.01	0.01
	¬ e	0.07	0.23	0.10	0.18

- Por ejemplo, el 50% de las personas son mujeres, y el

  (0.01+0.04+0.01+0.1)/ ((0.01+0.04+0.01+0.1)+ (0.00+0.02+0.00+0.07)) = 64%

  de los fumadores son mujeres

Dependencia e Independencia condicional

- **Definición 5: dependencia e independencia condicional:** Sean X, Y, Z tres conjuntos disjuntos de variables, entonces X se dice condicionalmente independiente de Y, dado Z, si y sólo si $p(x \backslash z, y) = p(x \backslash z)$ para todos los valores posibles de *x*, *y* , y *z* de X, Y y Z; En otro caso X e Y se dicen condicionalmente dependientes dado Z.
  - Cuando X e Y son condicionalmente independientes dado Z, se escribe I(X, Y|Z), y se escribe D(X,Y|Z) si X e Y son condicionalmente dependientes de Z.
  - La definición de independencia condicional lleva en sí la idea de que una vez conocida Z, el conocimiento de Y no altera la probabilidad de X.

Teorema de Bayes

- **Teorema de Bayes**

$$p(x_i \mid x_1, \ldots, x_k) = \frac{p(x_i)\, p(x_1, \ldots, x_k \mid x_i)}{\sum_{x_i} p(x_i) p(x_1, \ldots, x_k \mid x_i)}$$

- Supóngase que un paciente puede estar sano, o tiene una de m-1 enfermedades posibles $\{E_1, \ldots, E_{m-1}\}$. Por simplicidad, sea E una variable aleatoria que puede tomar uno de m posibles valores $\{e_1, \ldots, e_m\}$, donde e_m significa que el paciente está sano.
- Supongase también que se tienen n síntomas $\{S_1, \ldots, S_n\}$. Ahora dado que el paciente tiene un conjunto de síntomas $\{s_1, \ldots, s_n\}$, se desea calcular la probabilidad de que el paciente tenga la enfermedad E_i. Aplicando el teorema de Bayes:

$$p(e_i \mid s_1, \ldots, s_k) = \frac{p(e_i)\, p(s_1, \ldots, s_k \mid e_i)}{\sum_{e_i} p(e_i) p(s_1, \ldots, s_k \mid e_i)}$$

Teorema de Bayes

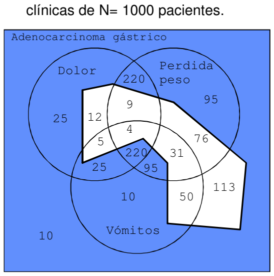
$$p(e_i \mid s_1, \ldots, s_k) = \frac{p(e_i)\, p(s_1, \ldots, s_k \mid e_i)}{\sum_{e_i} p(e_i)\, p(s_1, \ldots, s_k \mid e_i)}$$

- La probabilidad $p(e_i)$ se llama **probabilidad marginal**, **inicial** o "**a priori**" de la enfermedad E= e_i puesto que puede ser obtenida antes de conocer los síntomas.
- La probabilidad $p(e_i \mid s_1, \ldots, s_k)$ es la **probabilidad posterior**, **condicional** o "**a posteriori**" de la enfermedad E = e_i, puesto que se calcula después de conocer los síntomas $S_1 = s_1, \ldots, S_k = s_k$.
- La probabilidad $p(s_1, \ldots, s_k \mid e_i)$ se conoce por el nombre de verosimilitud de que un paciente con la enfermedad E = e_i tenga los síntomas $S_1 = s_1, \ldots, S_k = s_k$.

- **Se puede utilizar el teorema de Bayes para actualizar la probabilidad** "a posteriori" **usando ambas, la probabilidad** "a priori" **y la verosimilitud.**

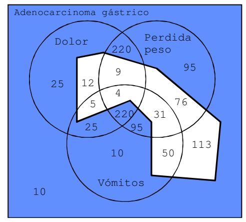
Ejemplo : Adenocarcinoma gástrico

- Un centro médico tienen una base de datos consistente en las historias clínicas de N= 1000 pacientes.



- Hay 700 pacientes (región sombreada) que tienen la enfermedad adenocarcinoma gástrico (G), y 300 no la tienen.
- Tres síntomas: Dolor (D), pérdida de peso (P) y vómitos (V) están ligados a la enfermedad.
- Cuando un paciente nuevo llega a la consulta, hay una probabilidad 700/1000 = 70% de que el paciente tenga G.

Ejemplo : Adenocarcinoma gástrico



=> Pueden hacerse las siguientes afirmaciones:

- Probabilidad a priori: 440 de 1000 pacientes vomitan => p(v) = 0.44
- Verosimilitud: El 50% de los pacientes que tienen la enfermedad vomitan.

  p(v|g) = card(v,g)/card(g)=350/750=0.5

  mientras que sólo el 30% de los paciente que no tienen la enfermedad vomitan

  p(v| ¬ g)=card(v, ¬ g)/card(¬ g)= 90/300=0.3
- Verosimilitud: El 45% de los pacientes que tienen la enfermedad vomitan y pierden peso,

  p(v,p|g) = card(v,p,g)/card(g)=315/750=0.45

  mientras que sólo el 12% de los que no tienen la enfermedad vomitan y pierden peso

  p(v,p| ¬ g) = card(v,p, ¬ g)/card(¬ g)=35/300=0.12

Ejemplo : Adenocarcinoma gástrico

- Puesto que la probabilidad inicial de que el paciente tenga adenocarcinoma gástrico, p(g)=0.7, no es suficientemente alta para hacer un diagnóstico **(tomar una decisión ahora implica una probabilidad 0.3 de equivocarse)**, el doctor decide examinara al paciente para obtener **más información**:
  - Supóngase que los resultados muestran que el paciente tiene los síntomas vómitos (V=v) y pérdida de peso (P=p). ¿Cuál es ahora la probabilidad de que el paciente tenga la enfermedad?
  - Tras observar que V=v la probabilidad a posteriori es

$$p(g|v) = \frac{p(g)\, p(v|g)}{p(g)\, p(v|g) + p(\neg g)\, p(v|\neg g)} = \frac{0.7 \times 0.5}{(0.7 \times 0.5) + (0.3 \times 0.3)} = 0.795$$

  - Tras observar que V=v y P = p la probabilidad a posteriori es

$$p(g|v,p) = \frac{p(g)\, p(v,p|g)}{p(g)\, p(v,p|g) + p(\neg g)\, p(v,p|\neg g)} = \frac{0.7 \times 0.45}{(0.7 \times 0.45) + (0.3 \times 0.12)} = 0.9$$

Ejemplo : Adenocarcinoma gástrico

- Nótese que cuando se aplica el teorema de Bayes sucesivamente, la probabilidad "a posteriori" calculada en una etapa dada es la misma que la probabilidad "a priori" en la siguiente.
  - Por ejemplo, la probabilidad "a posteriori" que se ha calculado en el primer paso anterior, puede ser utilizada como probabilidad "a priori" en la siguiente:

$$p(g|v,p) = \frac{p(g|v)\, p(p|g,v)}{p(g|v)\, p(p|g,v) + p(\neg g|v)\, p(p|\neg g,v)}$$

$$= \frac{0.795 \times 0.9}{(0.795 \times 0.9) + (0.205 \times 0.389)} = 0.9$$

Teorema de Bayes como clasificador

Simplificación del teorema

$$P(h/D)=\frac{P(D/h)P(h)}{P(D)}$$

- **P(h) : probabilidad a priori de la hipótesis h**
- **P(D): probabilidad a priori de los datos de entrenamiento D**
- **P(h/D): probabilidad de h dado D**
- **P(D/h): probabilidad D dado h**

Algoritmo de aprendizaje de la probabilidad máxima a posteriori MAP

- **Para cada hipótesis de h ∈ H, calcule la probabiliada a posteriori**

$$P(h/D)=\frac{P(D/h)P(h)}{P(D)}$$

- **Escoja la hipótesis h_{MAP} de mayor probabilidad a posteriori:**

$$h_{MAP}=\arg\max_{h\in H} P(h/D)$$

Elección de hipótesis

- **Generalmente se desea la hipótesis mas probable observada en los datos de entrenamiento**
- **Hipótesis de mayor probabilidad a posteriori**

$$\begin{aligned} h_{MAP} &= \arg\max_{h\in H} P(h/D) \\ &= \arg\max_{h\in H} \frac{P(D/h)P(h)}{P(D)} \\ &= \arg\max_{h\in H} P(D/h)P(h) \end{aligned}$$

- **Hipótesis de máxima verosimilitud**

$$h_{ML}=\arg\max_{h_i\in H} P(D/h_i)$$

Clasificación mas probable de una nueva instancia

- **Dada una nueva instancia x, cual es su clasificación mas probable?**
  - h_{MAP} (x) no es la clasificación mas probable
- **Considere**
  - $P(h_1/D) = 0.4$, $P(h_2/D) = 0.3$ e $P(h_3/D) = 0.3$
  - Dada una nueva instancia de x suponga que
  - $h_1(x) = +$, $h_2(x) = -$ y $h_3(x) = -$
  - La clasificacion mas probable de x es -

Aplicación del Teorema de Bayes

- **Sea**
  - M = enfermedad meningitis
  - S = Dolor de cabeza
- **Un doctor sabe**
  - P(S/M) = 0.5
  - P(M) = 1 / 50000
  - P(S) = 1 / 20

$$\mathbf{P}(M/S)=\frac{\mathbf{P}(S/M)\mathbf{P}(M)}{\mathbf{P}(S)}$$

$$=\frac{0.5*(1/50000)}{1/20}=0.0002$$

- La probabilidad de una persona tener meningitis dado que ella está com dolor de cabeza es 0,0002% o 1 en 5000.

Clasificador Bayesiano optimo

- **Un nuevo ejemplo puede ser clasificado como $v_j \in V$, la probabiliadad de clasificación correcta v_j**

$$P(v_j/D)=\sum_{h_i\in H} P(v_j/h_i)P(h_i/D)$$

Clasificación bayesiana optima

$$\arg\max_{v_j\in V} \sum_{h_i\in H} P(v_j/h_i)P(h_i/D)$$

Ejemplo clasificador Bayesiano

- **Dado**
  - $P(h_1/D) = 0.4$, $P(-/h_1) = 0$, $P(+/h_1) = 1$
  - $P(h_2/D) = 0.3$, $P(-/h_1) = 1$, $P(+/h_1) = 0$
  - $P(h_3/D) = 0.3$, $P(-/h_3) = 1$, $P(+/h_1) = 0$
- **Por lo tanto**

$$\sum_{h_i \in H} P(+/h_i)P(h_i/D) = 0.4 \qquad \sum_{h_i \in H} P(-/h_i)P(h_i/D) = 0.6$$

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j/h_i)P(h_i/D) = -$$

Clasificador Bayesiano ingenuo

- **Una función de clasificación f: X->V donde cada instancia en X es descrita por los atributos {a_1, ..., a_n}**

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\text{argmax}} P(v_j / a_1, \cdots, a_n) \\ &= \underset{v_j \in V}{\text{argmax}} \frac{P(a_1, \cdots, a_n / v_j)P(v_j)}{P(a_1, \cdots, a_n)} \\ &= \underset{v_j \in V}{\text{argmax}} P(a_1, \cdots, a_n / v_j)P(v_j) \end{aligned}$$

- **Suposición ingenua**
- **Clasificador bayesiano ingenuo**

$$P(a_1, \cdots, a_n / v_j) = \prod_i P(a_i / v_j)$$

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i / v_j)$$

Ejemplo

Dia	Tiempo	Temp.	Humed.	Viento	Jugar
D1	Sol	Caliente	Alta	Debil	No
D2	Sol	Caliente	Alta	Fuerte	No
D3	Cubierto	Caliente	Alta	Debil	Si
D4	Lluvia	Normal	Alta	Debil	Si
D5	Lluvia	Frio	Normal	Debil	No
D6	Lluvia	Frio	Normal	Fuerte	No
D7	Cubierto	Frio	Normal	Fuerte	Si
D8	Sol	Normal	Alta	Debil	No
D9	Sol	Frio	Normal	Debil	Si
D10	Lluvia	Normal	Normal	Debil	Si
D11	Sol	Frio	Alta	Fuerte	?

- **P(Si) = 5/10 = 0.5**
- **P(No) = 5/10 = 0.5**
- **P(Sol/Si) = 1/5 = 0.2**
- **P(Sol/No) = 3/5 = 0.6**
- **P(Frio/Si) = 2/5 = 0.4**
- **P(Frio/No) = 2/5 = 0.4**
- **P(Alta/Si) = 2/5 = 0.4**
- **P(Alta/No) = 3/5 = 0.6**
- **P(Fuerte/Si) = 1/5 = 0.2**
- **P(Fuerte/No) = 2/5 = 0.4**
- **P(Si)P(Sol/Si) P(Frio/Si)**
- **P(Alta/Sm) P(Fuerte/Si) = 0.0032**
- **P(No)P(Sol/No)P(Frio/No)**
- **P(Alta/No) P(Fuerte/No) = 0.0288**
- $\Rightarrow$ Jogar_Tenis(D11) = No