Aprendizaje Automático

Víctor de Juan

UAM - 15/16 C1

14 de junio de 2016 12:49

Apuntes UAM
Doble Grado Mat.Inf.

Código en Github

Índice general

[0] Documento compilado el 14 de junio de 2016 a las 12:49

Capítulo I

Introducción

Aristóteles en el siglo IV a.c. dice que el razonamiento deductivo se analiza aplicando 2 silogismos fuertes.

Silogismos fuertes

Definición I.1 **Silogismos fuertes**. *Premisa:* $A \implies B$

1 Si observamos A, deducimos B.

2 Si observamos $\neg B$*, deducimos que* $\neg A$*.*

Silogismos débiles

Definición I.2 **Silogismos débiles**. *Si vemos la calle mojada,* $\implies$ *lo más probable es que haya llovido.*

I.1. Introducción a la probabilidad

Vamos a repasar conceptos de probabilidad para poder utilizar los silogismos débiles.

Dominio

Definición I.3 **dominio**. *Todos los posibles resultados de un experimento aleatorio.*

Variable aleatoria

Definición I.4 **Variable aleatoria**. *Variable que identifica el experimento aleatorio*

Suceso

Definición I.5 **Suceso**. *Valor que toma la variable aleatoria al realizar el experimento aleatorio.*

Ejemplo: Tirar un dado.

- Dominio $= \{1, 2, 3, 4, 5, 6\}$
- Variable X representa tirar un dado.
- Posibles sucesos: $X = 1, X = 2, ..., X = 6$

A partir de esto ya podemos calcular probabilidades:

$$P\left(primo\right)$$

$$P\left(par\right)$$

Más repaso de probabilidad: probabilidad condicionada (regla de la suma)

Regla del producto

Definición I.6 **Regla del producto**.

$$P(A,B|I) = P(A|B,I)\cdot P(B|I) = P(B|A,I)\cdot P(A|I)$$

Vamos a ver cómo utilizar esta regla combinándola con los silogismos:

Podemos incluir el primer silogismo fuerte **??** haciendo $I = A \implies B$, entonces tenemos:

$$P(A,B|A \implies B) = ... = 1$$

Y aplicando el segundo silogismo fuerte, tenemos:

$$P(A|\bar{B},A \implies B) = ... = 0$$

Sin embargo, utilizando los silogismos débiles:

$$P(A|B,A \implies B) = \frac{P(A|A \implies B)}{P(B|A \implies B)} \geq P(A|A \implies B)$$

I.2. Toma de decisiones

A la hora de tomar decisiones, podemos utilizar 2 criterios. El criterio de máxima verosimilitud y el de ...

Criterio MAP

Definición I.7 **Criterio MAP**. *Maximiza la probabilidad a posteriori.*

$$H_i = \max_i\{p(H_i|D)\} = \max_i\{\frac{P(D|H_i)\cdot P(H_i)}{P(D)}\} = \max_i\{P(D|H_i)\cdot P(H_i)\}$$

Donde $\max_i$ *devuelve la hipótesis en la que se encuentra el máximo.*

Criterio MV

Definición I.8 **Criterio MV**. *Maximiza la máxima verosimilitud.*

$$H_i = \max_i\{P(D|H_i)\}$$

La diferencia entre los 2 criterios es la corrección y por tanto la complejidad. **MAP es más correcto**, pero **MV** es **más fácil.**

Vamos a verlo con un ejemplo.

Ejemplo: *Tenemos monedas justas y monedas trucadas. Las monedas trucadas tienen un 75 % de salir cara.*

Hemos obtenido cara en un lanzamiento. ¿Cuál es la probabilidad de que sea una moneda justa?

$D = c, H = t|j$

$$P(j|c) = \frac{P(c|j) \cdot P(j)}{P(c)} = \frac{0.5 \cdot 0.5}{P(c)}$$
$$P(t|c) = \frac{P(c|t) \cdot P(j)}{P(c)} = \frac{0.75 \cdot 0.5}{P(c)}$$

Como $P(j|c) + P(t|c) = 1$, *tenemos que* $P(c) = 0.5 \cdot 0.5 + 0.75 \cdot 0.5 = 0.625$, *entonces:*

$$P(j|c) = 0.4$$
$$P(t|c) = 0.6$$

Según el criterio MAP diríamos que es más probable que la moneda esté trucada.

Por algún motivo, calculamos:

$$P(j|x) = \frac{P(x|j) \cdot P(j)}{P(x)} = \frac{0.5 \cdot 0.5}{P(x)} = \frac{0.25}{P(x)}$$
$$P(t|x) = \frac{P(x|t) \cdot P(t)}{P(x)} = \frac{0.25 \cdot 0.5}{P(x)} = \frac{0.125}{P(x)}$$

Normalizando obtenemos:

$$P(j|x) = \frac{2}{3}$$
$$P(t|x) = \frac{1}{3}$$

I.2.1. Tasas de error

Vamos a estudiar las tasas de error siguiendo con el ejemplo de las monedas.

Existen 4 modelos posibles: *Tenemos:*

- *Espacio de atributos* $= \{c, x\}$[1]
- *Clases posibles*[2] $= H \in \{j, t\}$

Vamos a construir los 4 modelos posibles:

[1] cara o cruz

[2] Clases que puede asignar nuestra hipótesis H

- $h_1(c) = j$ *y* $h_1(x) = t$
- $h_2(c) = t$ *y* $h_2(x) = j$
- $h_3(c) = j$ *y* $h_3(x) = j$
- $h_4(c) = t$ *y* $h_4(x) = t$

Con lo calculado anteriormente (I.2 y I.2), observamos que el modelo h_2 es elegido según el criterio MAP.

Pero podemos plantearnos, ¿qué modelo es mejor? Pues el que menos se equivoque. Vamos entonces a calcular las probabilidades de error de los modelos.

Dejuan ha hecho a mano estos cálculos porque no tenía ni pies ni cabeza lo que había copiado.

Cálculo de errores

h_1: $P(error) = p(t,c) + p(j,x) = P(j|c) \cdot P(c) + P(j|x) \cdot P(x) = ... = 50\,\%$
También podíamos haber calculado (como estaba mal escrito en la versión anterior de los apuntes). A gusto del consumidor un método o el otro.

$$P(error) = 1 - P(acierto) = 1 - (p(j,c) + p(t,x)) = ...$$

h_2: $P(error) = p(j,c) + p(t,x) = P(j|c) \cdot P(c) + P(t|x) \cdot P(x) = ... = 50\,\%$

h_3: $P(error) = p(t,c) + p(t,x) = P(t|c) \cdot P(c) + P(t|x) \cdot P(x) = ... = 62.5\,\%$

h_4: $P(error) = p(j,c) + p(j,x) = P(j|c) \cdot P(c) + P(j|x) \cdot P(x) = ... = 37.5\,\%$

Matriz de confusión

Definición I.9 **Matriz de confusión**. *Tabla para visualizar el error de un clasificador. En la diagonal encontramos las tasas de acierto, y fuera de la diagonal, las tasas de error.*

En esta matriz, la diagonal son las tasas de acierto y lo que está fuera de la diagonal las tasas de fallo. Con este criterio, comprobamos entonces que efectivamente la tasa de error es $25 + 12.5 = 37.5$, como habíamos calculado anteriormente erróneamente.

Si tomamos h_2, la matriz de confusión sería (tomando la parte real arriba y las predicciones a la izquierda):

$$\begin{pmatrix} ccc & justa & trucada \\ \hline j & A_{11} & A_{12} \\ t & A_{21} & A_{22} \end{pmatrix}$$

Y vamos a ir calculando cada A_{ij}.

¿CUál es la probabilidad de que digamos justa siendo la moneda justa? Para ello pensamos ¿Cuándo decimos nosotros justa? Cuando ha salido una cruz, entonces A_{11} es la probabilidad de que sea justa y de que haya salido cruz, es decir $A_{11} = P(x, j) = P(x|j) \cdot P(j) = P(j|x) \cdot P(x)$. Entonces:

$$\begin{pmatrix} ccc & justa & trucada \\ \hline j & P(x,j) & P(x,t) \\ t & P(c,j) & P(c,t) \end{pmatrix}$$

Pero... Pensemos en el caso de una enfermedad y queremos, a partir de unos datos predecir si va a tener la enfermedad o no. Tenemos una enfermedad muy grave que padecen 10 personas en una muestra de 15000. Un modelo que diga que nadie la tiene, sólo falla en un $\frac{10}{150000} \sim 0$, un error muy bajo. ¿Es este modelo bueno? En casos como este, interesa saber detectar la enfermedad cuando se da y es un error mucho más grave no diagnosticar la enfermedad cuando sí la tiene, que diagnosticarla cuando no la tiene, con lo que, la matriz de confusión ya no es una buena manera de valorar modelos. Por ello construimos la matriz de coste o riesgo.

Matriz de coste (o riesgo)

Definición I.10 **Matriz de coste (o riesgo)**. *Una matriz de coste o riesgo consiste en definir ponderaciones para la gravedad de cada uno de los tipos de error.*

Regla de Bayes simétrica

Relga de Bayes asimétrica

Regla simétrica vs Regla asimétrica La **Regla de Bayes simétrica** utiliza una matriz de coste que es igual que la matriz de confusión, es decir, la importancia de los errores es simétrica. Por consiguiente, el lector avispado podrá imaginarse que la **Relga de Bayes asimétrica** utiliza una matriz de costes distinta de la de confusión, es decir, la importancia de los errores no es simétrica.

En el caso de la enfermedad descrito anteriormente, convendría utilizar una regla asimétrica.

Ejemplo: *Vamos a ver un ejemplo de costes con el caso de las monedas.*

Sea R el coste si pasamos una trucada por justa y R' el coste si pasamos una justa por trucada.

En el caso de las monedas, suponemos $R = 3R'$, con lo que el cálculo de los costes quedaría así:

h_1: $P(error) = R' \cdot P(j, c) + R \cdot P(t, x) = ... = R' \cdot 0.625$

h_2: $P(error) = R' \cdot P(j, x) + R \cdot P(t, c) = ... = R' \cdot 1.37$

h_3: $P(error) = R' \cdot P(j, x) + R \cdot P(t, x) = ... = R' \cdot 1.5$

h_4: $P(error) = R' \cdot P(j,c) + R \cdot P(t,c) = ... = R' \cdot 0.5$

Entonces, es h_4 la hipótesis que minimiza el coste (con la ponderación elegida), aunque no es la que minimiza el error en términos generales (como hemos visto antes).

Ejercicio 2.lentillas:

Este es un ejemplo de problema multidimensional.

Aquí tenemos los datos de donde sacar las probabilidades:

EDAD	LESIÓN	ASTIGM.	PROD_LAGRIM	DIAGNOSTICO
joven	miopía	no	reducida	no
joven	miopía	no	normal	blandas
joven	miopía	sí	reducida	no
joven	miopía	no	normal	duras
joven	hipermetropía	no	reducida	no
joven	hipermetropía	no	normal	blandas
joven	hipermetropía	sí	reducida	no
joven	hipermetropía	sí	normal	duras
mediana	miopía	no	reducida	no
mediana	miopía	no	normal	blandas
mediana	miopía	sí	reducida	no
mediana	miopía	no	normal	duras
mediana	hipermetropía	no	reducida	no
mediana	hipermetropía	no	normal	blandas
mediana	hipermetropía	sí	reducida	no
mediana	hipermetropía	sí	normal	no
mediana	miopía	no	reducida	no
mediana	miopía	no	normal	no
mediana	miopía	sí	reducida	no
mediana	miopía	no	normal	duras
mediana	hipermetropía	no	reducida	no
mediana	hipermetropía	no	normal	blandas
mediana	hipermetropía	sí	reducida	no
mediana	hipermetropía	sí	normal	no

- *Hipótesis: no llevar lentillas (n), lentillas duras (d) o lentilla blandas (b)*
- *Datos o atributos*

  edad: joven (j); mediana (m); avanzada (a)

  lesión (l): miopía (m); hipermetropía (h)

  Astigmatismo(a): sí (s), no (n)

  Producción de lágrimas (p): normal (n), reducido (r)

a) *Llega un paciente con $(l = m, p = n)$. ¿Diagnóstico?*

b) *Llega un paciente con $(l = h, p = n)$. ¿Diagnóstico?*

c) *Llega un paciente con $(l = m, p = r)$. ¿Diagnóstico?*

d) *Llega un paciente con* $(l = h, p = r)$*. ¿Diagnóstico?*

Apartado a)

$$P(d = n|l = m, p = n) = \frac{P(l = m, p = n|d = n)P(d = n)}{P(l = m, p = n)}$$

Probabilidades a priori

- $P(d = n) = \frac{15}{24}$
- $P(d = d) = \frac{4}{24}$
- $P(d = b) = \frac{5}{24}$

Verosimilitues:

- $P(l = m, p = n|d = n) = \frac{1}{15}$
- $P(l = m, p = n|d = d) = \frac{3}{4}$
- $P(l = m, p = n|d = b) = \frac{2}{5}$

Por el criterio de máxima verosimilitud, diagnosticaríamos lentillas duras, ya que ese diagnóstico es el que tiene mayor verosimilitud.

Por el criterio de máxima probabilidad a posteriori (MAP) tendríamos:

$$P(d = n|l = m, p = n) = \frac{P(l = m, p = n|d = n)P(d = n)}{P(l = m, p = n)} = \frac{\frac{1}{15}\frac{15}{24}}{...}$$

$$P(d = d|l = m, p = n) = \frac{P(l = m, p = n|d = n)P(d = n)}{P(l = m, p = n)} = \frac{\frac{3}{4}\frac{4}{24}}{...} \iff MAP$$

$$P(d = b|l = m, p = n) = \frac{P(l = m, p = n|d = n)P(d = n)}{P(l = m, p = n)} = \frac{\frac{2}{5}\frac{5}{24}}{...}$$

Por el criterio MAP, el diagnóstico será lentillas duras.

Apartado b)

$(l = h, p = n)$ Calculamos las probabilidades a posteriori (para el criterio MAP):

$$P(d = n|l = h, p = n) = ... = \frac{1}{3} = 33\,\%$$

$$P(d = d|l = h, p = n) = ... = \frac{1}{6} = 16\,\%$$

$$P(d = b|l = h, p = n) = ... = \frac{3}{6} = 50\,\%$$

La manera "chapucera" de hacer este cálculo es restringir la tabla a aquellos que tengan hipermetropía y una producción de lágrimas normal, y obtenemos una subtabla con 6 filas. De estas 6 filas contamos cuántos tienen diagnóstico blandas y encontramos 3. De las 6 filas, encontramos a 2 que tengan diagnóstico duras y a 1 que tenga diagnóstico blanda.

Calculamos las verosimilitudes:

$$P(l = h, p = n|d = n) = ... = \frac{2}{15} = 6\,\%$$

$$P(l = h, p = n|d = b) = ... = \frac{3}{5} = 60\,\%$$

$$P(l = h, p = n|d = d) = ... = \frac{1}{4} = 25\,\%$$

La manera "chapucera" de hacer este cálculo es restringir por diagnóstico. Nos quedamos con los que tengan diagnóstico no. De esas 15 filas, contamos los que tienen hipermetropía y producción normal de lágrimas.

Apartado c)

$(l = m, p = r)$ Calculamos las probabilidades a posteriori (para el criterio MAP):

$$P(d = n|l = h, p = n) = ... = 100\,\%$$
$$P(d = d|l = h, p = n) = ... = 0\,\%$$
$$P(d = b|l = h, p = n) = ... = 0\,\%$$

Calculamos las verosimilitudes:

$$P(l = m, p = r|d = n) = ... =$$

$$P(l = m, p = r|d = b) = ... =$$
$$P(l = m, p = r|d = d) = ... =$$

Apartado d)

$(l = h, p = r)$ Calculamos las probabilidades a posteriori (para el criterio MAP):

$$P(d = n|l = h, p = r) = ... = 100\,\%$$
$$P(d = d|l = h, p = r) = ... = 0\,\%$$
$$P(d = b|l = h, p = r) = ... = 0\,\%$$

Calculamos las verosimilitudes:

$$P(l = h, p = r|d = n) = ... =$$

$$P(l = h, p = r|d = b) = ... =$$
$$P(l = h, p = r|d = d) = ... =$$

I.2.2. Naive Bayes

Dimensión del espacio de atributos de lentillas: $3 \cdot 2 \cdot 2 \cdot 2 = 24$ ejemplos distintos, los mismos que en la tabla. En ejemplos con un espacio de atributos así reducido, podemos dedicarnos a contar para hallar las probabilidades, pero en caso de tener 10 atributos binarios, tendríamos un espacio de 1024 donde ya es más difícil ponerse a contar.

A este fenómeno se le denomina **MALDICIÓN DE LA DIMENSIONALIDAD**. Oh my god.

Maldición de la dimensionalidad

Definición I.11 **Maldción de la dimensionalidad**. *Básicamente, lo que dice esta maldición que nos ha caido es que:*

El número de ejemplos necesario para cubrir el espacio de atributos de forma uniforme crece exponencialmente con el número de atributos.

I.2.2.1. Primera puntualización

Solución inocente: Suponemos que todos los atributos son independientes. Con esto conseguimos que $P(l = m, p = n|d = d) = P(l = m|d = d) \cdot P(p = n|d = d)$

Obtención y explicación del clasificador Dado un vector de atributos (datos) $\overline{x} = (x_1, x_2, ..., x_n)$, queremos obtener la clase más probable entre todas las posibilidades H_i, es decir, hay que calcular:

$$P(H_i|\overline{x}) \forall i = 1, ..., k$$

Aplicando el teorema de Bayes, hay que calcular:

$$P(H_i|\overline{x} = \frac{P(\overline{x}|H_i) \cdot P(H_i)}{P(\overline{x})} = \frac{P(x_1, ..., x_n|H_i) \cdot P(H_i)}{P(x_1, ..., x_n)}$$

Lo que dice el método Naive Bayes es utilizar la ingenuidad[3] de que los atributos son independientes entre sí **dada la clase**, es decir:

$$\frac{P(x_1, ..., x_n|H_i) \cdot P(H_i)}{P(x_1, ..., x_n)} = \frac{P(x_1|H_i) \cdot ... \cdot P(x_n|H_i) \cdot P(H_1)}{P(x_1, x_n)}$$

Independiente dada la clase quiere decir que sólo podemos tener $P(x_1, ..., x_n|H_i) = \prod P(x_i|H_i)$ y no $P(x_1, ..., x_n) = \prod P(x_i)$, ya que en el segundo caso no están condicionados a la clase.

Entonces en el denominador seguimos teniendo $P(x_1, ..., x_n)$, que no nos gusta. Para calcularlo, podemos utilizar:

$$P(x_1, ..., x_n) = \sum_i \prod_j P(x_j|H_i) \cdot P(H_i)$$

[3] Naive se traduce al español por ingenuo o inocente

Ahora ya podemos construir la regla de clasificación con Naive Bayes:

Naive Bayes

Definición I.12 **Naive Bayes**.

$$argmax_{H_i} \prod_{j=1}^{n} P(x_j|H_i) \cdot P(H_i)$$

Donde $argmax$ significa: "el argumento que maximiza ... "

Complejidad $O(k \cdot n)$*, donde n es el número de atributos y k el número de hipótesis.*

Sobre este método, al parecer las $P(h_j|H_i$ y $P(H_i)$ son _____ fáciles de estimar de los datos y son estimaciones relativamente buenas.

Ejemplo: *¿Cuánto de fiable puede ser este método? Vamos a ver un ejemplo con el problema de las lentillas.*

$$P(l = m, p = n|d = n) \simeq P(l = m|d = n) \cdot P(p = n|d = n) = \frac{7}{15} \cdot \frac{3}{15} = \frac{7}{75}$$

Por otro lado, sabemos:

$$P(l = m, p = n|d = n) = \frac{1}{15}$$

*Y **obviamente**:*

$$\frac{1}{15} = \frac{5}{15} \neq \frac{7}{75}$$

Para realizar el cálculo hay que obtener las tablas en entrenamiento que cruzan los atributos con las hipótesis. Estas tablas son las que se construyen durante el entrenamiento para agilizar el cáculo. En el fondo, estas tablas son las $P(x_j|H_i)$ de la fórmula

lesión	n	b	d
miopía	7	2	3
hipermetropía	8	3	1

prod	n	b	d
reducida	12	0	0
normal	3	5	4

Vamos a calcular el resto de cosas para comparar cuánto nos desviamos al suponer independencia:

$$P(d = n|l = m, p = n) = \frac{P(l = m|d =?)P(p = n|d =?)P(d =?)}{P(...)} = \frac{1}{P(...)} \cdot (...) = 22\,\% \neq 17\,\%$$

$$P(d = b|l = m, p = n) = \frac{P(l = m|d =?)P(p = n|d =?)P(d =?)}{P(...)} = \frac{1}{P(...)} \cdot (...) = 31\,\% \neq 33\,\%$$

$$P(d = d|l = m, p = n) = \frac{P(l = m|d =?)P(p = n|d =?)P(d =?)}{P(...)} = \frac{1}{P(...)} \cdot (...) = 47\,\% \neq 50\,\%$$

No está muy desviado y obtenemos la misma decisión. Podría ocurrir, que saliera una decisión distinta al simplificar el algoritmo ignenuamente (naively).

I.2.2.2. Segunda puntualización:

Por ejemplo, vamos a cambiar uno de los datos anteriores:

$$P(d=n|l=m,p=r)=\frac{P(l=m|d=n)P(p=r|d=n)P(d=n)}{P(...)}=\frac{\frac{7}{15}\frac{12}{15}\frac{15}{24}}{P(...)}=\frac{\frac{7}{30}}{P(...)}=1$$
$$P(d=b|l=m,p=r)=\frac{P(l=m|d=b)P(p=r|d=b)P(d=b)}{P(...)}=\frac{\frac{3}{4}\frac{0}{4}\frac{4}{24}}{P(...)}=0$$
$$P(d=d|l=m,p=r)=\frac{P(l=m|d=d)P(p=r|d=d)P(d=d)}{P(...)}=...=0$$

Problema: Como es habitual en la asigatura, vamos a llevar el caso al extremo. Vamos a suponer que tenemos 150 atributos. Si en uno de los atributos hay un 0, toda la hipótesis se descarta. Lo que nos interesa es intentar evitar los 0 sin inventarnos demasiado los resultados, asique aplicamos la corrección de Laplace.

Corrección de Laplace

Definición I.13 **Corrección de Laplace**. *En casos en que aparecen valores nulos en las tablas se suma 1 a toda la tabla.*

Interpretación: La idea intuitiva es que pueden faltarnos datos y que tal vez no es categórico ese 0. Supongamos que tiro una moneda 3 veces y salen 3 caras. Sólo debería aparecer un 0 si la moneda tuviera 2 caras. Si no, simplemente nos falta información. Como sólo tener 2 caras en una moneda es absurdo, nos inventamos un experimento positivo y otro negativo para compensar esa posible falta de información.

Con esta modificación, reescribimos la tabla de producción de lágrimas:

prod	n	b	d
reducida	13	1	1
normal	4	6	5

Y recalculamos las probabilidades de diagnóstico con esta corrección:

$$P(d=n|l=m,p=r)=\frac{P(l=m|d=n)P(p=r|d=n)P(d=n)}{P(...)}=\frac{\frac{7}{15}\frac{13}{17}\frac{15}{24}}{P(...)}=\frac{0.22}{P(...)}$$
$$P(d=b|l=m,p=r)=\frac{P(l=m|d=b)P(p=r|d=b)P(d=b)}{P(...)}=\frac{\frac{3}{4}\frac{1}{6}\frac{4}{24}}{P(...)}=\frac{1}{48}\cdot\frac{1}{P(...)}$$
$$P(d=d|l=m,p=r)=\frac{P(l=m|d=d)P(p=r|d=d)P(d=d)}{P(...)}=\frac{1}{84}\frac{1}{P(...)}$$

¿Podría pasar que al aplicar la corrección de Laplace variase la hipótesis elegida? ¡Claro que sí!

I.2.2.3. Atributos cuantitativos

Ejemplo: Naive Bayes para atributos cuantitativos *En este caso, vamos a estudiar un problema de clasificación de plantas en función de la anchura.*

Estos son los datos:

Anchura de pétalos	*setosa*	*vesicolor*	*virginica*
0-0.2	$\frac{1}{10}$		
0.2-0.4	$\frac{18}{25}$		
0.4-0.6	$\frac{4}{25}$		
0.6-0.8	$\frac{4}{50}$		
0.8-1.0			
1-1.2		$\frac{1}{5}$	
1.2-1.4		$\frac{18}{50}$	
1.4-1.6		$\frac{17}{50}$	$\frac{3}{50}$
1.6-1.8		$\frac{4}{50}$	$\frac{1}{25}$
1.8-2.0		$\frac{1}{50}$	$\frac{8}{25}$
2.0-2.2			$\frac{6}{25}$
2.2-2.4			$\frac{11}{50}$
2.4-2.6			$\frac{3}{25}$
2.6-2.8			
2.8-3.0			

La solución es suponer que la variable sigue una distribución normal dada la clase.

Estudiamos $p(v_1, ..., v_n|\mu\sigma^2)$ suponiendo que su distribución es una normal.

¿Qué valores de μ y σ maximizan la verosimilitud de la muestra? Para ello derivamos respecto a μ y σ e igualamos a 0.

Tras muchas cuentas que no van a ser reflejadas aquí, obtenemos que

$$\mu = \frac{1}{N}\sum_{i=1}^{N} v_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - v_i)^2$$

Ha caido en un examen esas cuentas largas.

Volviendo al problema de las plantas, tenemos la siguiente tabla con la información sobre la anchura de los pétalos:

	set	ver	vir
μ	0.246	1.326	2.026
σ^2	0.0111	0.0391	0.0754

El objetivo es poder clasificar una nueva planta que nos venga de acuerdo a esa información, suponiendo normalidad dada la clase.

Ejemplo: *Tenemos:* $p(set) = p(vir) = p(ver) = \frac{1}{3}$

Suponemos $D \equiv ap = 1.5$ *¿A qué clase pertenece?¿Qué tipo de planta es?*

Vamos a clacular las probabilidades

$$p(set|ap = 1.5)$$

Si recurriéramos a la tabla de los datos I.2.2.3, diríamos que $p(set|ap = 1.5) = 0$ *ya que no tenemos ninguna planta con esas características. Pero como estamos suponiendo normalidad:*

$$p(set|ap = 1.5) = \frac{P(ap = 1.5|set)P(set)}{P(ap = 1.5)} =$$

Y aquí entra en juego la normalidad (dada la calse)

$$= p(ap = 1.5|set) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{2 \cdot \pi \cdot 0.0111} \cdot e^{\left(\frac{-(1.5-0.246)^2}{2 \cdot 0.0111}\right)} \simeq 10^{-31}$$

Ejemplo: *Vamos a utilizar Naive-Bayes con 2 atributos, uno discreto (longitud de sépalos) y otro continuo (anchura de pétalos).*

Anchura de pétalos

	set	*ver*	*vir*
μ	*0.246*	*1.326*	*2.026*
σ^2	*0.0111*	*0.0391*	*0.0754*

Longitud de sépalos

	set	*ver*	*vir*
pequeño (≤5)	*20*	*1*	*1*
mediano (5<...<6)	*30*	*25*	*7*
grande (6<...<7)	*0*	*23*	*30*
muy grande (≥ 7)	*0*	*1*	*12*

Nos piden calcular $P(set|ls = m, ap = 1.9)$. *Vamos a suponer independencia ya que aplicamos Naive-Bayes*

$$P(set|ls = m, ap = 1.9) = \frac{P(ap = 1.9|set) \cdot P(ls = m|set) \cdot P(set)}{P(...)} = \frac{1.23 \cdot 10^{-53} \cdot \frac{30}{50} \cdot \frac{1}{3}}{P(...)} \simeq 0$$

$$P(ver|ls = m, ap = 1.9) = \frac{2.99 \cdot 10^{-2} \frac{25}{50} \frac{1}{3}}{P(...)} = \frac{4.98 \cdot 10^{-3}}{P(...)}$$

$$P(vir|ls = m, ap = 1.9) = \frac{1.3 \cdot \frac{7}{50} \frac{1}{3}}{P(...)} = \frac{6.1 \cdot 10^{-2}}{P(...)}$$

Sin necesidad de normalizar, elegiremos la clase vir.

En el ejemplo anterior suponíamos independencia entre la longitud del sépalo y la anchura del pétalo, pero esta suposición no es muy acertada, ya que no sabemos (y es digno de suponer que sí) que haya una correlación entre los 2 atributos.

I.3. Modelos Gráficos (Redes de Bayes)

En esta sección vamos a ver cómo saber si las variables son independientes (y entonces Naive-Bayes mola) o si son dependientes y no deberíamos utilizarlo. La clave es la

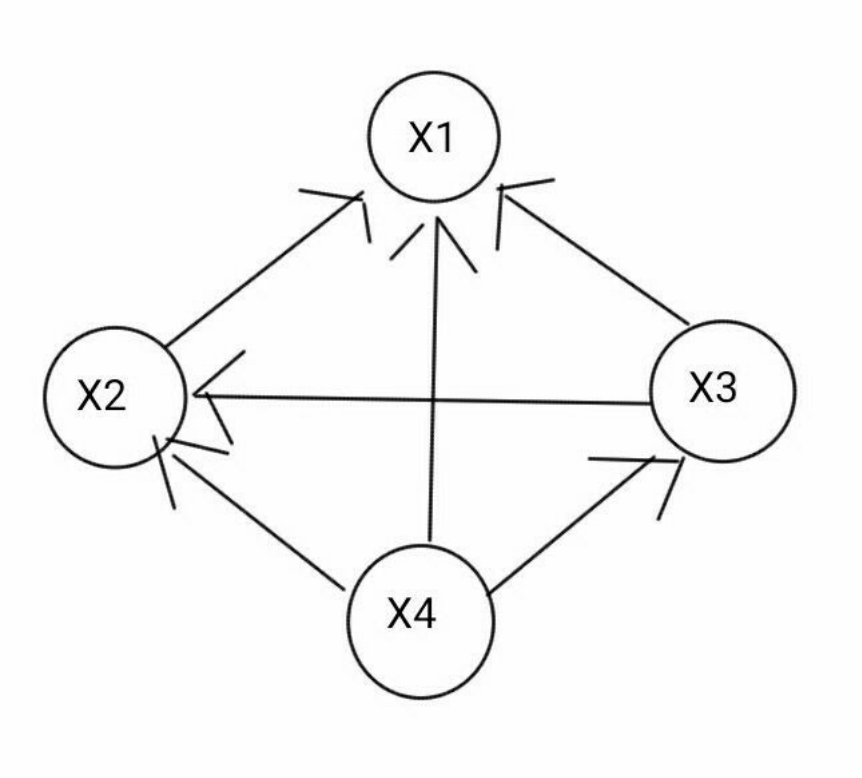
Regla de la cadena

Definición I.14 **Regla de la cadena**. *Aplicación de la regla del producto múltiples veces.*

Consideramos las variables aleatorioas x_1, x_2, x_3, x_4:

$$P(x_1, x{-}2, x_3, x_4) = P(x_1|x_2, x_3, x_4)P(x_2, x_3, x_4) = ... = P(x_1|x_2, x_3, x_4)P(x_2|x_3, x_4)P(x_3|x_4)P(x_4)$$

Vamos a representar gráficamente la regla de la cadena mediante un grafo **acíclico** dirigido que representa las relaciones entre las variables.



Red de Bayes

Definición I.15 **Red de Bayes**.

$$P(x_1, ..., x_n) = \prod_{i=1}^{N} p(x_i|pa_1)$$

donde pa_i es el conjunto de padres de la variable x_i, con $P(x_1|pa_i) = p(x_i)$ en caso de no tener padres la variables x_i.

Observación: En caso de ser independientes, las variables no tendrían padres , debido a que el grafo no tendría aristas.

I.3.1. Modelo gráfico para Naive-Bayes

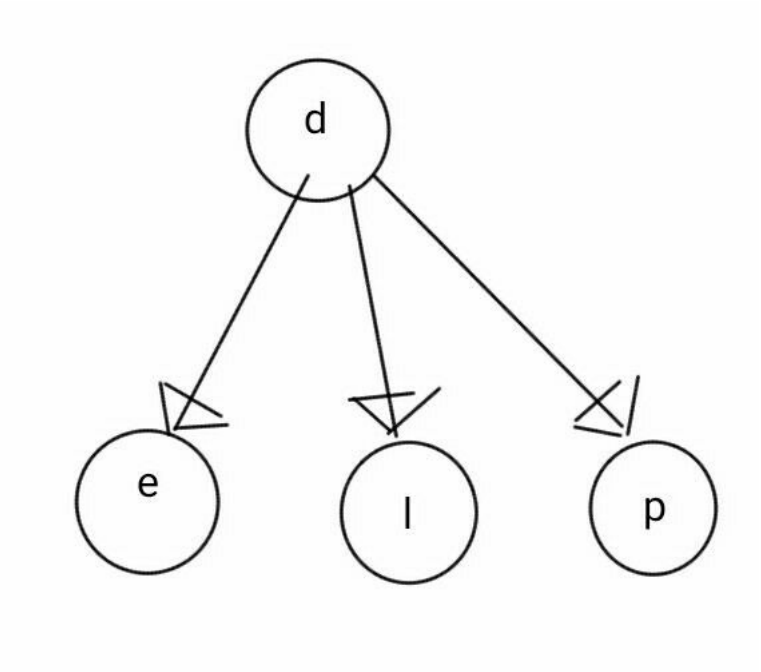
Recordando el problema de las lentillas:

$$P(d|e, l, p) = \frac{P(e, l, p|d) \cdot P(d)}{P(e, l, p)}$$

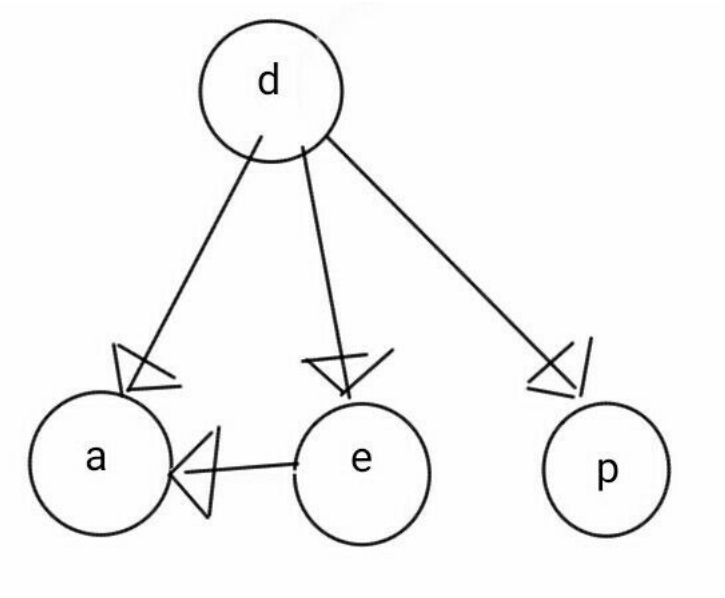
Con la regla de la cadena:

$$P(e, l, p|d) \cdot P(d) = P(e, l, p, d) = P(e, l, p|d) \cdot P(d) = P(e|d)P(l|d)P(p|d)P(d)$$

El grafo correspondiente a este caso:



¿Y si este grafo nos parece muy sencillo? Podemos suponer que la edad influye en el astigmatismo, asique añadimos una arista al grafo, y escribimos la fórmula:



$$P(d|a,e,p) = \frac{P(d,a,e,p)}{P(a,e,p)} =$$

Vamos a descomponer $P(d,a,e,p)$ utilizando el grafo.

- d no depende de nadie, es huérfano sin padres.
- e depende de d.
- a depende de 2 cosas, con lo que $P(a|e,d)$.
- ...

Con esta información, obtenemos:

$$P(d|a,e,p) = \frac{P(d,a,e,p)}{P(a,e,p)} = \frac{P(a|e,d)P(e|d)P(p|d)P(d)}{P(e,a,p)}$$

Seguimos el próximo día.

Vamos a ver los posibles diagnósticos:

$$P(d=n|e=j,p=na=s) = \frac{P(a=s|d=n,e=j)P(e=j|d=n)P(p=n|d=n)P(d=n)}{P(...)} =$$

$$= \frac{\frac{2}{4}\frac{4}{15}\frac{3}{15}\frac{15}{24}}{P(...)} = ... = 17\,\%$$

$$P(d=d|e=j,p=na=s) = \frac{\frac{2}{2}\frac{2}{4}\frac{4}{4}\frac{4}{24}}{P(...))} = ... = 83\,\%$$

$$P(d=b|e=j,p=na=s) = ... = 0$$

I.4. Vecinos próximos

I.4.1. Introducción y motivación del método

Hasta ahora la verosimilitud la hemos calculado de diferentes maneras:

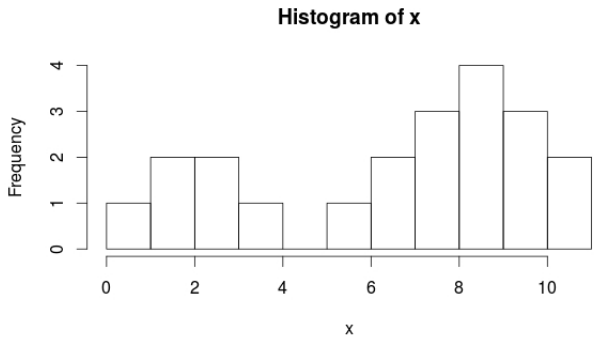
Discretos Con frecuencias

Continuos Discretizando y usando frecuencias (histograma). Esta estimación es no paramétrica y la probabilidad de que un punto caiga en una caja es el número de puntos en esa caja dividido del total.

Estimando la densidad de probabilidad con una normal. La estimación es paramétrica y se calcula utilizando la integral de la función de densidad, es decir:

$$P(1.4 \leq x < 1.5) = \int_{1.4}^{1.5} Pdf(x)dx$$

¿Qué podemos hacer para estimar paramétricamente una muestra como la siguiente?



Aquí vemos claramente que no podemos estimar utilizando una normal, ya que ni se le parece. Podríamos tomar 2 normales, o tendríamos que intentar generalizar la estimación por histograma a D dimensiones.

En el caso de estimar generalizando el histograma, tendríamos $P(x \in R) = \frac{K}{N}$, donde K es el número de puntos en la región y N es el número total de datos.

Si intentamos exportar este modelo a más dimensiones, generalizar la estimación por histograma, necesitamos muchos más datos de los que podemos obtener para estimar así, debido a la **MALDICIÓN DE LA DIMENSIONALIDAD**.

Para que la estimación generalizada por histograma sea una buena estimación, tenemos que tener el espacio suficientemente cubierto de puntos, es decir n tiene que ser más grande cuanto mayor es la dimensión.

Si tenemos R suficientemente pequeña, podemos intentar estimarlo con integrales:

$$P(\bar{x} \in R) = \int_R Pdf(\bar{x})d\bar{x} \simeq pdf(\bar{x}) \cdot V$$

Donde V es el volumen de R.

Igualando, obtenemos

$$P(\bar{x} \in R) = \frac{K}{N} = pdf(\bar{x}) \cdot V \implies pdf(\bar{x}) = \frac{K}{NV}$$

Esta última fórmula se puede usar de 2 formas:

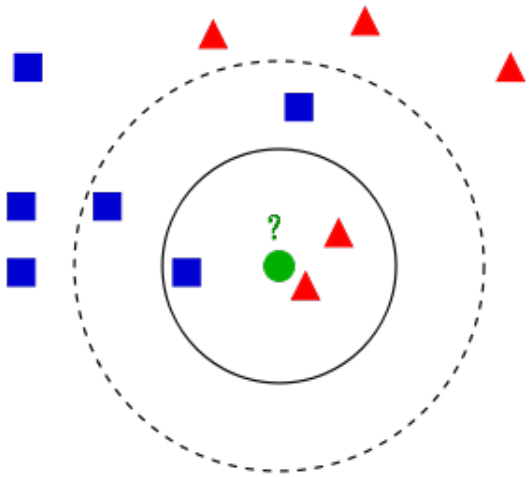
- Si fijamos K y obtenemos el V necesario paar incluir K puntos, clasificamos por el algoritmo de vecinos próximos.
- Si fijamos V y calculamos K obtenemos métodos de núcleos (SVM) que son los más eficientes que hay y que no vamos a ver.

Vecinos próximos

Definición I.16 **Vecinos próximos**. *Fijamos K (el número de vecinos)*

Se calcula una esfera alrededor del punto de interés hasta que incluya K puntos.

Ejemplo: *Tenemos 2 clases, cuadrados azules y triángulos verdes. La línea continua corresponde a $K = 3$, ya que hemos agrandado la esfera hasta incluir 3 puntos.*



¿Y cómo asignar la clase? Vamos a ir poco a poco:

- ***Verosimilitud*** $P(\overline{x}|Clase) = \frac{K_c}{N_c V}$
- ***Priori*** $P(Clase) = \frac{N_c}{N}$
- ***Evidencia*** $P(\overline{x}) = \frac{K}{NV}$

Con estos datos podemos calcular la probabilidad a posteriori:

$$P(Clase|\overline{x}) = \frac{\dots}{\dots} = \text{ se cancela todo ... } = \frac{K_c}{K}$$

Con $K = 3$ puntos, la clase asignada (a priori) sería "triángulo rojo", ya que es la clase con mayor probabilidad a posteriori $\left(\frac{2}{3}\right)$.

En cambio, si tomamos $K = 5$ (la linea discontinua), asignaríamos (a priori) la clase "cuadrado azul", ya que es la clase mayoritaria e la esfera.

I.4.2. Regla de clasificación $K - NN$

Algoritmo

Entrenamiento: Guardar los datos de entrenamiento.

Clasificar:

1. Por cada ejemplo $\overline{z}$ en entrenamiento se calcula la distancia de $\overline{z}$ a $\overline{x}$ usando la métrica.
2. Se seleccionan los K más cercanos.
3. Se devuelve la clase más común en esos K

Complejidad: $O(N \cdot \#\{ \text{ atributos } \})$

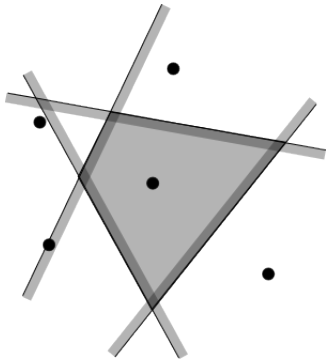
I.4.3. Problemas de K-NN

I.4.3.1. Normalizar

Los cambios de unidades pueden afectar a la clasificación, con lo que necesitamos **normalizar los datos**. Normalizamos los datos normalizando, es decir, $y = \frac{x-\mu}{\sigma} \forall x$ y trabajamos con los y.

Poligonos de Thiessen

Definición I.17 **Poligonos de Thiessen**. *Tomamos un punto y haciendo mediatrices con los K vecinos cercanos, creamos los polígonos.*



A la hora de clasificar, comprobamos a qué polígono pertenece.

I.4.3.2. Elección de K

Selección de K óptimo:

- Calcular distancias entre todos los pares de puntos.
- Para cada ejemplo:

  Se ordena el resto de eemplos por distancias.p

  Se hacen tantas clasificaciones como valores de K estemos evaluando.
- Se calcula el error y nos quedamos con el que menor error tenga.

I.5. Clasificadores lineales

Son aquellos que dan lugar a fronteras de decisión lineales. En 2 dimensíones es una recta, en n dimensiones es un hiperplano.

Clasificador lineal

Definición I.18 **Clasificador lineal**. *Sea h un clasificador lineal:*

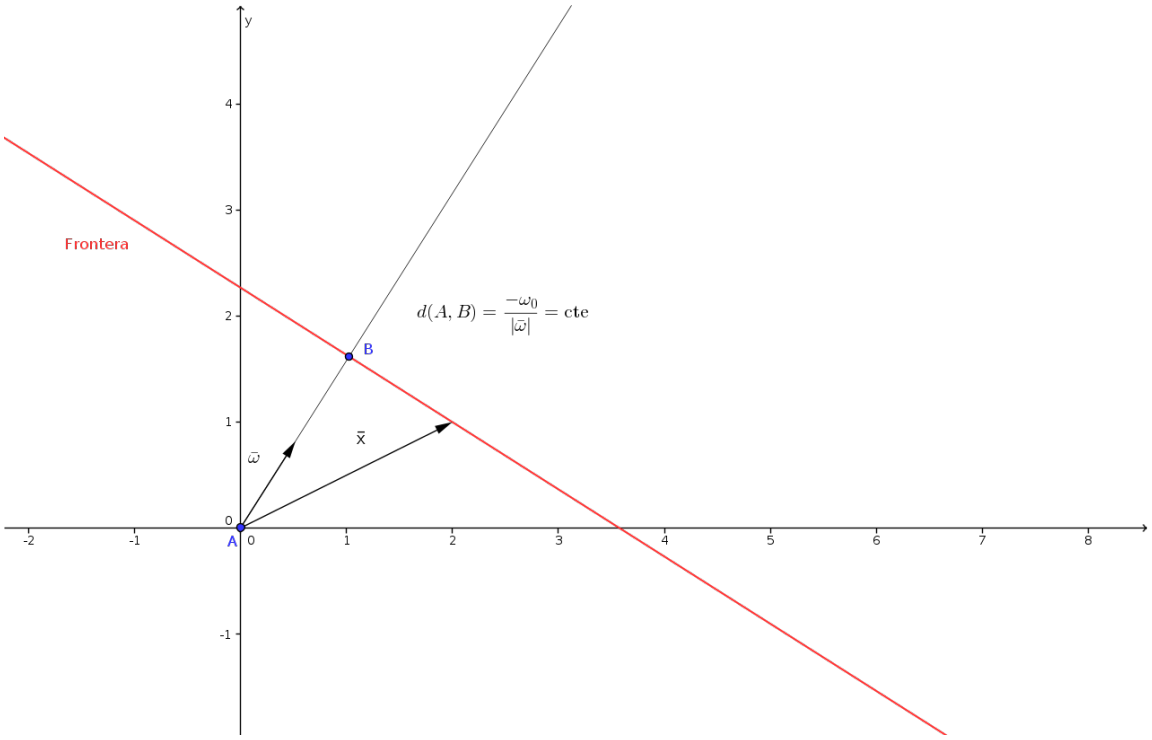
$$h(\mathbf{x}) = \begin{cases} c_1 & \langle \omega, \mathbf{x} \rangle + \omega_0 > 0 \\ c_2 & \langle \omega, \mathbf{x} \rangle + \omega_0 < 0 \end{cases}$$

Frontera

Llamamos ***Frontera*** *al hiperplano* $\langle \omega, \mathbf{x} \rangle + \omega_0$.

Ejemplo: *Vamos a construir la frontera de manera genérica. Para ello, tenemos:*

$$\langle \omega \mathbf{x} \rangle + \omega_0 = 0 \rightarrow |\mathbf{x}| \cos(\alpha) = \frac{-\omega_0}{|\omega|}$$



Veamos como surgen las fronteras lineales de forma "natural" haciendo suposiciones sencillas sobre la forma funcional de los datos.

Función sigmoidal

Definición I.19 **Función sigmoidal**.

Consideramos un problema de dos clases C_1, C_2. *Podemos escribir la probabilidad a posterior:*

$$P(C_1|\mathbf{x}) = \frac{P(\mathbf{x}|C_1)P(C_1)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|C_1)P(C_1)}{\sum P(\mathbf{x}|C_i)P(C_i)}$$

Dividimos numerador y denoiador por $P(\mathbf{x}|C_1)(C_1)$

$$\frac{1}{1+c} = \frac{1}{1+e^{-a}} = \sigma(a)$$

donde $a = ln(c)$ *y* $c = \dfrac{P(\mathbf{x}|C_1)P(C_1)}{P(\mathbf{x}|C_2)P(C_2)}$

¿Qué hemos ganado con esto? Hemos reescrito Bayes como una sigmoidal.

Esto permite clasificar de la siguiente manera:

$$P(\mathbf{x}|C_1)P(C_1) > P(\mathbf{x}|C_2)P(C_2) \implies \textit{elegimos } C_1$$

Esto es equivalente a "si a > 0, entonces elegimos C_1, sino, elegimos C_2"

Suponemos que la veroimilitud sigue una distribución normal en n dimensiones, siendo $\boldsymbol{\mu_k}$ el vector de medias y Σ la matriz de covarianzas. Todos los productos entre vectores son productos escalares. Ante cualquier tipo de duda (de que no cuadran las dimensiones, etc...) aplicar el producto escalar.

$$p(\mathbf{x}|C_x = \frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left\{\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu_k})\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right\}$$

I.5.1. Frontera de decisión con hiperplano (2 clases)

Suponiendo:

- tenemos 2 clases.
- $\Sigma_1 = \Sigma_2 = I$

Vamos a construir la frontera de decisión con hiperplano para 2 clases:

$$a = \ln\left(\frac{P(\mathbf{x}|C_1)P(C_1)}{P(\mathbf{x}|C_2)P(C_2)}\right) = \ln\left\{\exp\left\{\frac{\mathbf{x}-\boldsymbol{\mu_1}}{2} - \frac{\mathbf{x}-\boldsymbol{\mu_2}}{2}\right\}\frac{P(C_1)}{P(C_2)}\right\} =$$

$$= \left(\frac{(\mathbf{x}-\boldsymbol{\mu_2})^2}{2} - \frac{(\mathbf{x}-\boldsymbol{\mu_2})^2}{2}\right) + \ln\frac{P(C_1)}{P(C_2)} = 0$$

Si $P(C_1) = P(C_2)$, tenemos:

$$\left(\frac{(\mathbf{x}-\boldsymbol{\mu_2})^2}{2} - \frac{(\mathbf{x}-\boldsymbol{\mu_2})^2}{2}\right) = 0$$

Si $(\mathbf{x}-\boldsymbol{\mu_2})^2 > (\mathbf{x}-\boldsymbol{\mu_1})^2 \implies C_1$. ¿Porué C_1? Porque estamos más lejos de la media $\boldsymbol{\mu_2}$ que de la media $\boldsymbol{\mu_1}$. Es como vecinos próximos si sólo tuvieramos 2 vecinos, la media de 1 y la media del otro.

Si por el contrario, $P(C_1) \neq P(C_2)$, tenemos:

$$\frac{\mathbf{x}^2 + \mu_2^2 - 2\mathbf{x}\mu_2}{2} - \frac{\mathbf{x}^2 + \mu_1^2 - 2\mathbf{x}\mu_1}{2} + \ln\frac{P(C_1)}{P(C_2)} =$$

$$= (\mu_1 - \mu_2)\mathbf{x} + \frac{1}{2}(\mu_2^2 - \mu_1^2) + \ln\frac{P(C_1)}{P(C_2)} = 0 \tag{I.1}$$

Esto es un hiperplano que utilizaremos como frontera de decisión.

En este caso, tenemos $\omega = \boldsymbol{\mu_1} - \boldsymbol{\mu_2}$ y $\omega_0 = \frac{1}{2}(\mu_2^2 - \mu_1^2) + \ln \frac{P(C_1)}{P(C_2)}$

Es decir, el hiperplano será:

$$\langle \omega, \mathbf{x} \rangle + \omega_0 = 0 \implies \langle (\boldsymbol{\mu_1} - \boldsymbol{\mu_2}), \mathbf{x} \rangle + \frac{1}{2}(\mu_2^2 - \mu_1^2) + \ln \frac{P(C_1)}{P(C_2)}$$

I.5.1.1. Ampliamos las suposiciones I

- 2 clases
- Verosimilitud normal en $D-$dimension con $\Sigma_1 = \Sigma_2 = \Sigma$

Entonces, la frontera de decisión:

$$a = 0 = \ln \left(\frac{P(\mathbf{x}|C_1)P(C_1)}{P(\mathbf{x}|C_2)P(C_2)} \right)$$

Desarrollando igual que antes, obtenemos que la frontera de decisión óptima es un hiperplano:

$$\omega = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$
$$\omega_0 = \frac{1}{2}\boldsymbol{\mu_1}\Sigma\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}\Sigma\boldsymbol{\mu_2} + \ln \left(\frac{P(C_1)}{P(C_2)} \right)$$

Ahora vamos a obtener los parámetros de las gausianas por máxima verosimilitud, es decir: $\text{máx}_i\{P(D|H_i)\}$

Tenemos:

$$D = \{(\mathbf{x}_i, y_i) i = 1, ..., N_{train}\}$$
$$H = \{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \Sigma\}$$

Vamos a ello:

$$P(D|H) = P(\{\mathbf{x}_i, y_i\}|\boldsymbol{\mu_1}\boldsymbol{\mu_2}\Sigma) = \prod_i^{N_{train}} P(\mathbf{x}_i y_i|\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \Sigma)$$

Ahora aplicamos el logaritmo, derivamos e igualamos a 0.

Este es el enfoque de regresión logística.

Si te parece que queda colgado este final, a mi no me preguntes porque ha sido así de literal en clase.

I.6. Regresión logística

Aunque la regresión logística se llame regresión logistica, no es un método de regresión. Es un método de clasificador.

La probabilidad a posteriori de C_1 es:

$$P(C_1|\mathbf{x}) = \sigma(a) = \sigma(\langle \boldsymbol{\omega}, \mathbf{x} \rangle + \omega_0)$$

¿Cuántos parámetros habría que estimar en $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \Sigma$? Sabiendo que $P(\mathbf{x}|C_i) = N(\boldsymbol{\mu_i}, \Sigma)$

Tenemos $2D$ por $\boldsymbol{\mu_i}$ más, $\frac{D(DH)}{2}$ por Σ y 2 ($P(C_i)$), es decir demasiados.

En cambio, utilizando la probabilidad a posteriori de C_i, sólo hay que estimar $\boldsymbol{\omega}, \omega_0$, es decir, $D + 1$ parámetro.

Notación: Para simplificar, vamos a escribir:

$$\boldsymbol{\omega}_{new} = (\omega_0, \boldsymbol{\omega})$$

$$\boldsymbol{x}_{new} = (1, \boldsymbol{x})$$

¿Qué hemos ganado con esto? Algo más de compacidad, ya que $P(C_1|\mathbf{x}) = \sigma(\langle \boldsymbol{\omega}_{new}, \mathbf{x}_{new} \rangle)$

Ejemplo: *Disponemos de N ejemplos.*

$$D = \{(\mathbf{x}_i, t_i) i = 1, ..., N\}$$

donde $t_i = \begin{cases} 1 & x_i \in C_1 \\ 0 & x_i \in C_2 \end{cases}$

En regresión logística,

- $H = \boldsymbol{\omega}$
- $D = t_1, ..., t_N$
- $I = \boldsymbol{x_1}, ..., \boldsymbol{x_N}$

Objetivo: *maximizar por MV para encontrar los valores de $\boldsymbol{\omega}$.*

La verosimilitud es:

$$P(t_1, ..., t_N|\boldsymbol{\omega}, \boldsymbol{x}_1, ..., \mathbf{x}_N) = \prod^{N} P(t_i|\boldsymbol{\omega}, \mathbf{x}_i)$$

Con la codificación de t_i, podemos escribir:

$$P(t_i|\boldsymbol{\omega}, \mathbf{x}_i) = \sigma\left(\langle \boldsymbol{\omega}, \mathbf{x}_i \rangle\right)^{t_i} \cdot \left(1 - \sigma(\langle \boldsymbol{\omega}, \mathbf{x}_i \rangle)\right)^{(1-t_i)}$$

Juntando y llamando a $\sigma_i = \sigma(\boldsymbol{\omega}, \mathbf{x}_i)$*:*

$$= ... = \prod^{N} \sigma_i^{t_i}(1-\sigma_i)^{(1-t_i)}$$

Primero aplicamos el $-\ln$ *para simplificar el cálculo y derivamos:*

... ...

Conclusión: *10 minutos después sigue haciendo cuentas.*

En la clase siguiente *se tomó la decisión de mejor optimizar ejemplo a ejemplo.*

Ejemplo: *Sea* $\omega = \{-1, \frac{1}{3}, \frac{2}{3}\}$ *y queremos clasificar los datos* $x_1 = (0,0)$ *y* $x_2 = (3, 1.5)$*. Los coeficientes se han calculado tomando como clase de referencia o clase principal* C_1*.*

x_1

$$\langle \omega, (1, x_1) \rangle = \langle \left(-1, \frac{1}{3}, \frac{2}{3}\right), (1,0,0) \rangle = -1$$

Y ahora aplicamos la sigmoidal:

$$P(C_1|x_1) = \frac{1}{1+e^{-(-1)}} = 0.27$$

Comentamos que $P(C_2|x_1) = 1 - P(C_1|x_1)$

x_2

$$\langle \omega, (1, 3, 1.5) \rangle = ... = 1$$

Y ahora aplicamos la sigmoidal:

$$P(C_1|x_2) = \frac{1}{1+e^{-1}} = 0.73$$

Observación*: Es interesante ver que son los complementarios. ¿A qué se debe esto? A que están a la misma distancia de la recta frontera.*

I.6.1. Algoritmo de la regresión logística

Para aumentar la verosimilitud de t_i del ejemplo $\mathbf{x_i}$ hay que mover la recta ω en sentido opuesto al gradiente $(\sigma_i - t_i)\mathbf{x_i}$ y proporcional a una constante de aprendizaje $\eta = \omega + \omega - \eta(\sigma_i - t_i)\mathbf{x_i}$

Algoritmo Definimos 2 parámetros: η y $epocas$ (estos parámetros se darán como argumento)

1 Generamos un ω aleatorio con componentes entre -1 y 1.

2 for (ie=1;ie<epocas;ie++) for(i = 1; i < N; i++)

[2.1] calculamos $\sigma_i = \sigma(\langle \omega, \mathbf{x}_i \rangle)$

[2.2] Actualizamos $\omega = \omega - \eta(\sigma_i - t_i)\mathbf{x_i}$

3 return ω

El coste del entrenamiento del algorimo es $O(epocas \cdot N \cdot \#atributos$

Ejemplo:

x_1	x_2	t	
0	*0*	*0*	C_2
1	*1*	*0*	C_2
0	*1*	*0*	C_2
1	*1*	*1*	C_1

Tomamos $\eta = 1$ *y arbitrariamente* $\omega = \{-0.3, 0, 0.4\}$

La frontera de decisión entonces es:

$$-0.3 + 0x_1 + 0.4x_2 = 0 \rightarrow x_w = \frac{0.3}{0.4}$$

Ejemplo: $\mathbf{x} = (1, 0, 0)$

$$\langle \omega, \mathbf{x} \rangle = -0.3$$

$$\sigma(-0.3) = \frac{1}{1 + e^{-(-0.3)}} = 0.43$$

Está bien claisificado, ya que da C_2 *(la principal es* C_1*).*

$$\omega = \omega - \eta(\sigma - t)\mathbf{x} = (-0.3, 0, 0.4) - 1(0.43 - 0)(1, 0, 0) = (0.73, 0, 0.4)$$

Vamos a hacer otra iteración del bucle. Ahora toca $\mathbf{x} = (1, 1, 0)$*:*

$$\langle \omega, \mathbf{x} \rangle = 0.73$$

...

$$\omega = (-1.06, -0.33, -0.4)$$

Aquí dejamos los datos de las siguientes iteraciones:

$\mathbf{x} = (1, 0, 1) \rightarrow \omega = (-1.39, -0.33, 0.06)$ $\mathbf{x} = (1, 1, 1) \rightarrow \omega = (-0.55, 0.51, 0.9)$

I.6.1.1. Regresión logísitca por MAP

En vez de maximizar la verosimilitud de t dado $\mathbf{x}$, vamos a maximizar la probabilidad a posteriori. Teníamos:

$$P(t_i, ..., t_n)|\omega, \mathbf{x_1}, ..., \mathbf{x_n}) = \prod \sigma_i^{t_i}(1-\sigma_i)^{t_i-1}$$

Suponiendo un a priori normal para el módulo de ω tenemos $p(\omega) = N(0, \sigma)$.[4] Entonces, la probabilidad a posteriori queda:

$$P(\omega, \mathbf{x_1}, ..., \mathbf{x_n}) = \alpha \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-|\omega|^2}{2\sigma^2}\right) \prod \sigma_i^{t_i}(1-\sigma_i)^{t_i-1}$$

Observación: En la fórmula hay 2 σ distintas. σ_i es la sigmoidal y σ es la varianza del módulo de ω.

Aplicamos $-ln$ y derivamos la verosimilitud y la regla de actualización queda:

$$\omega = \omega - R\left(\sigma_i - t_i\right)\mathbf{x_i} + \frac{\omega}{N\sigma^2}$$

Algoritmo: Completar

[4] Como ω se calcula con valores iniciales aleatorios entre $-a$ y a. Lo que no tengo ni idea es de la varianza...

Capítulo II

Algoritmos genéticos

- Población de individuos que luchan por recursos del entorno.
- Los individuos que mejor se adaptan al entorno son los que sobreviven.
- Existen mecanismos de adaptación al entorno: reproducciño, mutçión.
- AG.. utilizan estas ideas para resolver problemas de optimización:
  Problema $\equiv$ entonro.
  Función de ajusto o fitness (mide adaptación al entorno).
  Conjunto de soluciones (Población)
  Cada individuo codigica una solución (material genético).
  Mecanismos de adaptación aleatorios (cruce, selección, mutación).

Este es el esquema básico de algorimos genéticos:

Algoritmo

1 Inicialización aleatoria de la población P.

2 Mientras no se cumpla la condición de terminación:
  P' =seleccion_progenitores(P)
  P' =recombinación(P')
  P' =mutación(P')
  P =seleccion(P, P') return best(P')

Pero esto es demasiado vago. ¿Qué es la selección, la recombinación, la mutación...? ¿Cuál es la condición de terminación? Estas son las preguntas que cada algoritmo concreto irá respondiendo.

Crommosoma **Notación:** Doficicación de la solución en cromosomas. Un **crommosoma** es una cadena de genes. Los valores que toman los genes se denominan alelos. Por ejemplo: si las soluciones son enteros $[0, 31]$, se podrían codificar con 5 bits (genes) cuyos valores son 0 o 1. El cromosoma es el array de 5 genes.

Definir la función de fitnes que diga lo bueno o mala que es una solución. Ha de ser mayor cuanto mejor sea la solución.

Inicialiazción y determinación del tamaño de la población. Mayot cuanto maás complejo es el problema.

Determinar operadores de la evolución:

- Selección de progenitores: se selecciona aleatoriamente con reemplzado de la población tantos progenitores como individos. Normalmente se hace selección proporcinal a fitness.
- Recombinación: Generalmente se obtienen 2 vástafos de cada 2 progenitores. Se aplica con probabilidad p_c. p_c es alta $\sim [0.7, 1]$. Importante: los vástagos deben seguir siendo soluciones válidas.

  Algunos operadores de cruce:

  Cruce en un punto. A la izquierda los progenitores y a la derecha los vástagos:

$$\left.\begin{matrix}(1\,1 & 0\,1\,1)\\(0\,1 & 1\,0\,1)\end{matrix}\right\} \rightarrow \begin{matrix}(1\,1\,1\,0\,1)\\(0\,1\,0\,1\,1)\end{matrix}$$

  Cruce en n puntos: el anterior cortamos en un único punto (después del segundo gen). Ahora podemos cortar en n puntos como pueda ser intuitivo.

  Cruce uniforme: Vamos a torar una moneda por cada gen para determinar a qué vástago va.
- Mutación: Se realizan modificaciones aleatorias de genes en la población con probabilidad p_m.

  Flip aleatorio de un bit con $p_m \sim \frac{1}{N}$ siendo N el tamaño de la población por el número de genes.
- Selección: Copia de P' en P a veces con eltismo: copia las 2 mejores a P.
- Condición de parada:

  Nº máximo de generaciones.

  Nº máximo de generaciones sin mejorar.

Ejemplo: *La función de fitnes es $f(x) = x^2$.*

Los datos son:

Inicial	*x*	*f*	$\frac{f_i}{\sum f_i}$
01101	*13*	*169*	*0.14*
11000	*24*	*576*	*0.49*
01000	*8*	*64*	*0.06*
10011	*19*	*361*	*0.31*

Ahora seleccionamos proporcional a fitness (la primera columna han sido aleatorios)

Selección	*prob.fitness*		*Pto cruce*	*Vástagos*	*Mutación*	*x*	*fitness*
1	*01101*		*4*	*01100*	*01100*	*12*	*144*
2	*11000*			*11001*	*11001*	*25*	*625*
2	*11000*		*2*	*11011*	*11011*	*27*	*729*
4	*10011*			*10000*	*10100*	*18*	*324*

II.1. Teoría de esquemas:

Los algoritmos genéticos funcionan descubriendo y recombinando buenos bloques de soluciones de forma paralela.

Formalizaremos el concepto de bloque como esquema.

Para codificación binaria un esquema va a ser una cadena de $0, 1, *$, donde $*$ representa cualquier valor posible.

Ejemplo:

- $H = 1 * * * 0$*: Todas las cadenas de longitud 6 que empiezan por 1 y terminan por 0.*

Orden de un esquema

Definiciones *Definición II.1* ***Orden de un esquema****. Número de bits definidos*

Distancia de un esquema

Definición II.2 ***Distancia de un esquema****. Distancia entre los bits definidos más lejanos*

Ejemplo:

Orden $O(1 * * * 0) = 2$

Distancia $d(*1 * *0) = 3$

Distancia $d(10) = 1$

Distancia $d(* * * \underbrace{1 * 0 * 1 * 0} * * *) = 6$

Combinatoria básica:

- ¿Cuántos cromosomas existen de longitud l? 2^l
- ¿Cuántos esquemas existen de longitud l? $(2 + 1)^l$

  ¿Cuántos esquemas cumple el cromosoma: 101? 2^l, ya que cada bit puede tomar 2 valores: $*$, o el correspondiente de 101

Proposición II.1. *Una población de N individuos contiene instancias desde 2^l, hasta $N2^l$. En este caso, se están evaluando $O(2^l)$ esquemas.*

Demostración. A buen entendedor, pocas palabras bastan. □

En el algoritmo, la inicialización es aleatoria. Por ello, en promedio habrá $\frac{N}{2}$ esquemas que empiecen por 1 y $\frac{N}{2}$ esquemas que empiecen por 0.

Funciones de t

Definición II.3 **Funciones de t.**

- $n_H(t)$ *es el número de instancias del esquema H en tiempo t.*
- $f_i(t)$= *fitness del individuo i en tiempo t.*
- $\bar{f}(t)$ *fitness promedio de la población en tiempo t.*
- $\bar{f}_H(t)$ *fitness promedio de las instancias de H en tiempo t.*

Queremos calcular el valor esperado de $n_H(t+1)$. Para ello, vamos a dar 3 pasos:

1. Selección de progenitores proporcional a fitness:

$$\frac{f_i(t)}{\sum f_i(t)}$$

Entonces, el número esperado de veces que se selecciona el individuo i es

$$N\frac{f_i(t)}{\sum f_i(t)} = \frac{f_i(t)}{\bar{f}(t)}$$

Además, el número de individuos del esquema H seleccionados S_s es:

$$S_s(t) = \sum \frac{f_i(t)}{\bar{f}(t)} = \sum \frac{n_H(t) f_i(t)}{n_H(t)\bar{f}(t)} = \frac{\bar{f}_H(t)}{\bar{f}(t)} n_H(t)$$

Ejemplo: *Cruce en un punto con probabilidad p_c.*

Un esquema sobrevive si al menos 1 de sus vástagos pertenece al esquema.

Una cota inferior de supervivencia de cruce es:

$$S_c = 1 - p_c \frac{d(H)}{l-1}$$

Ejercicio 1.1:

Buscar el máximo de $f(x) = (x-63)^2$ en $[0, 63]$ con algoritmos genéticos, utilizando:

- *Selección proporcional a fitness*
- *Cruce en 1 punto con probabilidad* $p_c = 1$.
- *Mutación con* $P_m = \frac{1}{24}$

Apartado a)

Completa la tabla de la población inicial.

cromosomas	*x*	*fitness*
100110		
110001		
110110		
011001		

Apartado b)

Esquemas: $H_1 = 1 * 0 * **$ *y* $H_2 = 0 * * * * * *$

Cómo sobrevive cada esquema a cada uno de los pasos del algoritmo en 1 iteración.

a) Población inicial:

cromosomas	x	fitness
100110	38	625
110001	49	196
110110	54	81
011001	25	1444

b)

$$S_s(H) = n_H(t)\frac{\overline{f_H})t}{\overline{f}(t)} \text{ para } H \in \{H_1, H_2\}$$

Vamos a calcular cada término

$$\overline{f}(t) = \frac{1444 + 81 + 196 + 625}{4} = 586.5$$
$$\overline{f}_{H_1}(t) = \frac{625 + 196 + 81}{3} = 300.7$$
$$\overline{f}_{H_2}(t) = \frac{1444}{1} = 1444$$

Ahora ya podemos calcular la supervivencia (el número esperado de individuos) para cada esquema:

$$S_s(H_1) = 3 \cdot 300.7\frac{1}{586.5} = 1.54$$
$$S_s(H_2) = 1 \cdot 1444\frac{1}{586.5} = 2.46$$

Conclusiones: Esto significa que en media, el cuarto individuo va a salir en media 2.4 veces y los otros 3, en media 1.56 veces.

No es casualidad que $1.54 + 2.46 = 4$, número de individuos.

Pero no es lo mismo el número esperado de individuos de un esquema en caso de que haya cruces o en caso de que no. Si hay mutaciones o no, etc

Aplicamos la fórmula de supervivencia tras el cruce:

$$S_c = 1 - p_c \frac{d(H)}{l-1} \rightarrow \begin{cases} S_c(H_1) = 1 - 1\frac{2}{5} = 0.6 \\ S_c(H_2) = 1 - 1\frac{0}{5} = 1 \end{cases}$$

Y ahora, calculamos la supervivencia tras la mutación, con una fórmula máginay novedosa:

$$S_M = (1 - P_M)^{o(H)} \rightarrow \begin{cases} S_M(H_1) = (1 - \frac{1}{24})^2 = 0.92 \\ S_M(H_2) = 1 - 1\frac{0}{5} = 1 \end{cases}$$

Y podríamos preguntarnos, porque somos muy quisquillosos: ¿cuál es la probabilidad total de supervivencia del esquema? El producto de las 3. Pero... 2.46 no es una probabilidad...

Sería[1]

$$S_T(H_1) = \frac{1.54}{4} \cdot 0.6 \cdot 0.92 = 0.21$$

$$S_T(H_2) = \frac{2.46}{4} \cdot 1 \cdot 0.96 = 0.59$$

Como vemos, tiene mucho sentido que el esquema 2 tenga más probabilidades de supervivencia, ya que genera individuos con in fitnees muy alto.

Ejercicio 1.2:

Queremos resolver el problema del viajante con algoritmos genéticos.

Para este problema, definir los cromosomas, el ftness el cruce y la mutación.

Lo primero es definir el cromosoma. **No tienen por qué ser bits** es una información importante para esto.

Los **genes** son arrays de enteros entre 1 y n, siendo n el número de nodos que visitar, donde el elemento i del array es el elemento i en ser visitado.

El **fitness** podría ser la distancia. El problema es que queremos minimizar la distancia y la función fitness, cuanto más alta mejor. Es por ello que tomamos $f = \frac{1}{distancia}$

El **cruce** es algo más raro: Sea $32|1465$ y $21|4563$. Vamos a construir cada hijo. El primer hijo toma la parte izquierda del primer padre y reordena su parte derecha

[1] No lo hemos visto en clase, pero tiene sentido

por el orden dado según el otro padre. Es decir, el primer hijo será $32|---$ donde la segunda parte dentrá los números $1, 4, 5, 6$ tal y como vienen en el padre derecho, esto es $1, 4, 5, 6$ (vemos que se han intercambiado el 5 y el 6). El primer hijo será $32|1456$. El segundo hijo será $21|---$ donde la parte derecha serán los números $3, 4, 5, 6$ según el orden del primer padre, es decir: $21|3465$

Ejercicio 1.3: *Vamos a construir el vector ω de regresión logística con algoritmos genéticos.*

Nos piden definir la codificación de los cromosomas, fitness, cruce y mutación

Cromosoma un array de floats.

Fitness % de acierto podría ser una opción el problema es que podemos tener varias rectas con el mismo porcentaje de acierto. Entonces, podemos tomar como fitness el sumatorio de la probabilidad a posteriori de cada punto (o algo así).

Cruce uniforme

Cruce Podemos tomar el cruce de siempre, ya que el orden no influye como en el caso anterior. El cruce de siempre se denomina **Cruce uniforme**. Otro posible cruce sería dar sólo 1 hijo a partir de 2 padres con la media de los valores.

Mutación Una posible mutación puede ser cambiarle el signo. Esto es una mutación demasiado heavy. Otra posibilidad es añadirle ruido aleatorio, por ejemplo con una $N(0, 1)$.

Capítulo III

Clustering

Agrupar elementos que son parecidos.

Vamos a ver un ejemplo manual:

Perro	pez volador	besugo
Pinguino	murciélago	dipnoi
gallina	delfín	gorrión

¿Qué podemos clasificar aquí? Según "tierra, mar y aire", o según "mamífero, no mamífero" ... Hay muchas posibilidades para agrupar.

Podemos definir los siguientes atributos:

- Medio.
- Respiración.
- Alimentación.
- Volador o no.

Antes de empezar, tenemos que definir la selección de atributos, la medida de proximidad, el algoritmo de clústering y cómo validar los resultados, es decir, cómo agrupar los agrupamientos.

M-clústering

Definición III.1 **m-clústering**. *Sea un conjunto de datos* $X = \{\{\mathbf{x}_1\}, ..., \{\mathbf{x}_n\}\}$

Definimos un m-clustering de X *como la partición en m-subconjuntos* $c_1, ..., c_m$ *disjuntos no vacíos.*

III.1. Medidas de proximidad entre puntos: métrica

Métrica

Definición III.2 **Métrica**. *https://es.wikipedia.org/wiki/Distancia*

Distancia euclídea:
Distancia Manhattan:

- **Distancia euclídea:** $\sqrt{\sum}$
- **Distancia Manhattan:** $d(x, y) = \sum_{d=1}^{n} |x_d - x'_d|$

Pero también necesitamos definir la distancia entre un punto y un cluster. Podemos definirla como la distancia al centro, la distancia al más cercano o la distancia al más lejano.

También tendríamos que definir la distancia entre 2 clusters, de la misma manera que antes. Media de las distancias, mínima(enlace simple) y máxima (enlace completo).

Apéndice A

Ejercicios

A.1. Hoja 1

Ejercicio 1.1:

Un taxi golpea a una persona de noche y huye. En la ciudad operan dos compañías de taxis. Una verde y otra azul. El 85 % de los taxis son verdes.

a) *¿De qué color es más probable que sea el taxi?*

b) *Un testigo dice que es azul con fiabilidad del 80 %*

c) *Decisión final usando el criterio de máxima verosimilitud y el criterio de máxima probabilidad a posteriori.*

Notación H de hipótesis, D de Datos e I de información.

Apartado a)

Se deja como ejercicio para el lector experimentado.

Apartado b)

P(H=verde) = 85 % P(H=azul) = 15 %

$$P\left(H = azul|D = azul, I\right) = \frac{P(D = a|H = a, I) \cdot P(H = A|I)}{P(D = a|I)}$$

Para calcularlo, necesitamos saber la probabilidad con la que el testigo dice que un taxi es azul.[1]

$$P(D = a|I) = P(D = a|H = a)\cdot P(H = a)+P(D = verde|H = v)\cdot P(H = v) = 0.8\cdot 0.15+0.2\cdot 0.85 = 0$$

Con esto ya tenemos:

$$P\left(H = azul|D = azul, I\right) = \frac{0.8 \cdot 0.15}{0.29} = 0.41$$

[1] Suponemos que el testigo se equivoca igual de azul a verde que de verde a azul

Por el contrario, la probabilidad de que el taxi sea verde (utilizando la información del testigo) es de un 0.59.

Notación Observamos que todas las probabilidades son condicionadas a I, que no influye en nada y omitiremos cuando no sea influyente

Apartado c)

Por el criterio MV - máxima verosimilitud (I.8):

$$\max_i\{P(D=a|H=a), P(D=a|H=verde)\} = \max_i\{0.8, 0.2\} = P(D=a|H=a)$$

Este criterio, al ser más sencillo y omitir parte de la información (que a veces no podremos calcular), nos da una respuesta menos correcta, y en este caso contraria a la obtenida anteriormente.

Ejercicio 1.2: *Las probabilidades pueden representar conexiones lógicas, no casuales. Consideremos una urna con 6 bolas blancas y 6 bolas negras*

a) *En la primera extracción se ha eliminado una bola blanca. ¿Cuál es la probabilidad de extraer una bola blanca en la segunda extracción?*

b) *Por el contrario, supongamos que la segunda extracción ha sido blanca. ¿Cuál es la probabilidad de que la primera fuera blanca también?*

Apartado a)

$P(s=b|p=b) = \frac{5}{11}$

Apartado b)

Como siempre, empezamos definiendo los datos y las hipótesis.

$D \equiv (s=b)$ donde s se refiere a segunda extracción.

$H \equiv (p=b|p=n)$ donde p se refiere a primera extracción.

$$P(p=b|s=b) = \frac{P(s=b|p=b)\cdot P(p=b)}{P(s=b)} = \frac{\frac{5}{11}\cdot\frac{6}{11}}{P(s=b)}$$

Donde $P(s=b) = \frac{6}{12}\cdot\frac{5}{12} + \frac{6}{12}\frac{6}{12}$, con lo que

$$P(p=b|s=b) = \frac{5}{11}$$

$$P(p=n|s=b) = \frac{P(s=b|p=n)\cdot P(p=n)}{P(s=b)} = ... = \frac{6}{11}$$

Por si el lector avispado se lo pregunta, $P(p = n|s = b) + P(p = b|s = b) = 1$ no es casualidad. Sólo hay 2 opciones, o la primera era negra o era blanca, independientemente de toda la información de la que dispongamos.

Ejercicio 1.3:

Según el hombre del tiempo, la probabilidad de lluvia hoy es del 20 %. Estamos en un sótano siun ventanas y no podemos saber qué tiempo hace fuera. Sin embargo, vemos entrar a alguien llevando un paraguas. Sabiendo que la probabilidad de que alguien lleve paraguas y esté lloviendo es del 70 % y sólo del 10 % en caso contrario. ¿Cuál es la probabilidad de que esté lloviendo?

Aclaramos que la probabilidad del 10 % se refiere a llevar paraguas en caso de que no esté lloviendo.

$D = p$, es decir, alguién entra con paraguas.

$H = r$ (rain), es decir, llueve.

$$P(r|p) = \frac{P(p|r) \cdot P(r)}{P(p)} = \frac{0.7 \cdot 0.2}{P(p)}$$
$$P(\bar{r}|p) = \frac{P(p|\bar{r}) \cdot P(\bar{r})}{P(p)} = \frac{0.1 \cdot 0.8}{P(p)}$$

Para saber cuál es más probable, no es necesario calcular el denominador. Por ello concluiomos que es más probable que esté lloviendo (ya que $0.14 > 0.08$).

Para calcular las probabilidades con exactitud, necesitamos saber $P(p)$. Para ello tenemos 2 opciones:

La versión corta, es que $P(r|p) + P(\bar{r}|p) = 1$, con lo que el denominador debe ser $P(p) = 0.7 \cdot 0.2 + 0.1 \cdot 0.8$, con lo que las probabilidades son $63.8\,\%$ y $36.2\,\%$ respectivamente.

Por otro lado, podríamos aplicar la regla del producto para calcularlo:

$$P(p) = P(p|r) \cdot P(r) + P(p|\neg r) \cdot P(\neg r) = 0.7 \cdot 0.2 + 0.1 \cdot 0.8$$

Obteniendo exactamente el mismo resultado.

Ejercicio 1.4: *Tenemos 2 bolsas. La bolsa A tiene 2 bolas negras y 3 blancas. La bolsa B tiene 3 bolas negras y 2 blancas.*

a) *Se elige una de las bolsas al azar, y escogemos al azar una bola. ¿Cuál es la probabilidad de que sea negra?*

b) *Habiendo extraido una bola negra. ¿Cuál es la bolsa más probable?*

c) *Decisión final usando el criterio de máxima verosimiitud y el critero de máxima probabilidad a posteriori.*

Apartado a)

$$P(negra) = P(A, negra) + P(B, negra) = P(negra|A, I) \cdot P(A) + P(negra|B, I) \cdot P(B)$$

$$P(negra) = 0.4 * 0.5 + 0.6 * 0.5 = 0.5$$

Apartado b)

La intuición nos dice que será de la bolsa B, ya que esta tiene más bolas negras, y las 2 bolsas son equiprobables.

Vamos a justificarlo matemáticamente:

P(H=A) = 50 %

D=negra

$$P(A|D = negra) = \frac{P(negra|A) \cdot P(A)}{P(negra)}$$

Ejercicio 1.5: *Hay 2 bolsas: A y B. La bolsa A contiene 2 bolas negras y 3 blancas. La B, 3 negras y 2 blancas. Se selecciona una bolsa al azar, teniendo en cuenta que la bolsa A tiene un 75 % de probabilidad de ser elegida, y se extrae una bola.*

a) *Calcular la probabilidad de que la bola sea negra.*

b) *Si la bola obtenida ha sido negra, calcular qué bolsa es más probable que hayamos elegido.*

c) *Decisión final usando el criterio de máxima verosimilitud y el criterio de máxima probabilidad a posteriori.*

Notación:

n, b se refiere a bola negra o blanca respectivamente.

A, B se refieren a la bolsa de la que se ha extraído la bola.

Apartado a)

$P(n) = P(n|A) \cdot P(A) + P(n|B) \cdot P(B) = \frac{2}{5} \cdot \frac{3}{4} + \frac{3}{5} \cdot \frac{1}{4} = \frac{9}{20}$

Apartado b)

$P(A|n) = \frac{P(n|A) \cdot P(A)}{P(n)} = \frac{\frac{2}{5} \cdot \frac{3}{4}}{\frac{9}{20}} = \frac{2}{3}$

Por otro lado, $P(B|n) = \frac{1}{3}$. Como la bola sólo puede estar en la bolsa A o en la bolsa B, es decir, son casos disjuntos, $P(A|n) + P(B|n) = 1$. Si la bola pudiera estar en 2 bolsas a la vez, tendríamos que hacer el cálculo porque esta regla no se cumpliría.

Apartado c)

Según el criterio MV - máxima verosimilitud (I.8), es más probable que la bolsa sea la B.

$$\max_i\{P(n|A), P(n|B)\} = \max_i\{0.4, 0.6\} \implies P(n|B)$$

A.3. Tema 3

Ejercicio 3.1:

¿Cuál es la verosimilitud de B, C, D, E dado A, siendo este el modelo gráfico?

5 nodos:

- $A \to \{B, C, D\}$
- $B \to \{\}$
- $C \to \{B\}$
- $D \to \{E\}$
- $E \to \{C\}$

$$P(A, B, C, D, E) = \underbrace{P(B|A, C)P(C|A, E)P(D|A)P(E|D)}_{\text{verosimilitud}}\underbrace{P(A)}_{\text{a priori}}$$

Ejercicio 3.2: *¿Cuál es el modelo gráfico asociado a ...?*

$$P(x_1)P(x_2)P(x_3)P(x_4|X_1x_2x_3)P(x_5|x_1x_3)P(x_6|x_4)P(x_7|x_5x_4)$$

Un grafo con 7 nodos en el que:

- $x_1 \to \{x_4, x_5\}$
- $x_2 \to \{x_4\}$
- $x_3 \to \{x_4, x_5\}$
- $x_4 \to \{x_6, x_7\}$
- $x_5 \to \{x_7\}$
- $x_6 \to \{\}$

- $x_7 \rightarrow \{\}$

Ejercicio 3.3:

- $Clase \rightarrow \{A, C, E\}$
- $A \rightarrow \{\}$
- $C \rightarrow \{A\}$
- $E \rightarrow \{\}$

Además, tenemos:

A	*C*	*E*	*Clase*
1	*1*	*0*	*A*
1	*0*	*1*	*A*
0	*0*	*0*	*S*
0	*1*	*1*	*S*
0	*1*	*1*	*A*
1	*1*	*0*	*A*
1	*1*	*0*	*A*
0	*0*	*1*	*S*
1	*0*	*1*	*A*
1	*0*	*0*	*S*

$$P(\text{Clase} = A | A = 1, C = 0, E = 0) =$$

$$\frac{P(A = 1 | C = 1, \text{Clase} = A) P(C = 0 | \text{Clase} = A) P(E = 0 | \text{Clase} = A) P(\text{Clase} = A)}{P(A = 1, C = 0, E = 0)}$$

$$P(\text{Clase} = A | A = 1, C = 0, E = 0) = \frac{\left(\frac{2}{2} \cdot \frac{2}{6} \cdot \frac{3}{6}\right)\left(\frac{6}{10}\right)}{P(...)}$$

Índice alfabético