APRENDIZAJE BAYESIANO

La ausencia de evidencia no es evidencia de ausencia

EL CISNE NEGRO

EL IMPACTO DE LO ALTAMENTE IMPROBABLE

Nassim Nicholas Taleb

PRÓLOGO

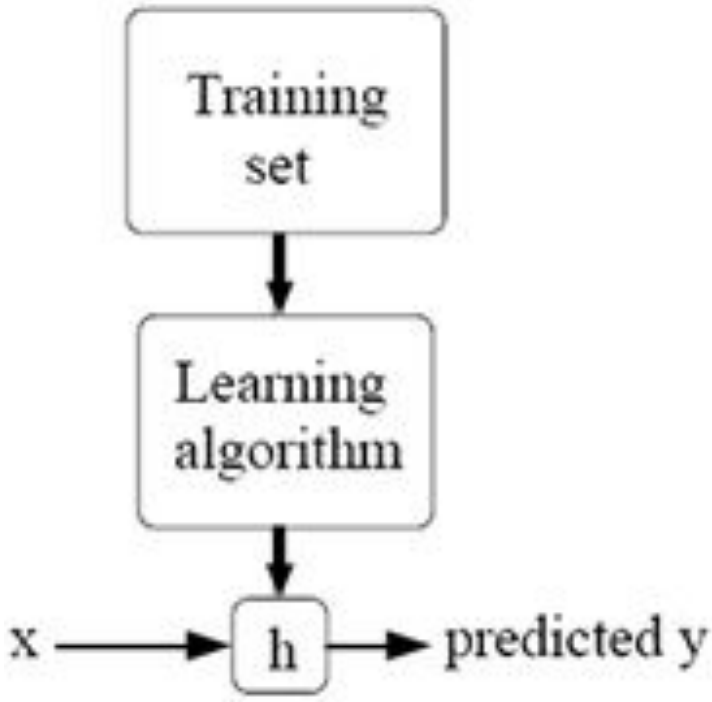
Del plumaje de las aves

Antes del descubrimiento de Australia, las personas del Viejo Mundo estaban convencidas de que todos los cisnes eran blancos, una creencia irrefutable pues parecía que las pruebas empíricas la confirmaban en su totalidad. La visión del primer cisne negro pudo ser una sorpresa interesante para unos pocos ornitólogos (y otras personas con mucho interés por el color de las aves), pero la importancia de la historia no radica aquí. Este hecho ilustra una grave limitación de nuestro aprendizaje a partir de la observación o la experiencia, y la fragilidad de nuestro conocimiento. Una sola observación puede invalidar una afirmación generalizada derivada de milenios de visiones confirmatorias de millones de cisnes blancos. Todo lo que se necesita es una sola (y, por lo que me dicen, fea) ave negra.[1]

Agenda de la semana

- Introducción
- Teorema de Bayes
- Hipótesis MAP
- Clasificador sencillo

- Clasificador de texto
- Clasificador óptimo
- Algoritmos MAP
- Tarea 2



¿De qué hablamos cuando hablamos de “aprender”?

Un sistema es realmente inteligente si es capaz de observar su entorno y aprender de él.

¿Cómo actualizar nuestras creencias cada vez que obtenemos nuevas evidencias?

La auténtica inteligencia reside en adaptarse, tener capacidad de integrar nuevo conocimiento, resolver nuevos problemas, y aprender de errores.

Motivación

- Los modelos basados en árboles de decisión o reglas asumen que hay una división clara entre conceptos (solo una respuesta correcta)
- Para algunos problemas es más interesante tener decisiones difusas (soft)
- Esto se puede modelar usando distribuciones de probabilidad para representar los conceptos

All models are wrong but some are useful

Inferencia estadística

La inferencia estadística es el proceso de utilizar el *análisis de datos* para deducir las propiedades de una distribución de probabilidad subyacente.

¿Los milagros existen?

David Hume (1711-1776) decía que son imposibles:

Evidencia de los milagros ← testimonio externo de otros

Evidencia en contra de los milagros ← se experimenta personalmente todos los días a través de las leyes inmutables de la naturaleza.

La única forma de que la evidencia de los milagros sea suficiente para su prueba es si el testimonio externo es más convincente y confiable que las mismas leyes de la naturaleza.

¿Los milagros existen?

Richard Price (1723-1791) decía que no es posible negar la posibilidad de un milagro basándose en observaciones de normalidad a gran escala, es decir, la falta de milagros.

Calcula la supuesta probabilidad de ver que la marea no llegará a la costa algún día.

Conclusión: aunque improbables, los milagros existen y son el producto de un poder superior.

¿Los milagros existen?

Razonamiento bayesiano:

un enfoque probabilístico para la inferencia

Thomas Bayes
(1702-1761)

ry *. The purpoſe I mean is, to ſhew what reaſon we have for believing that there are in the conſtitution of things fixt laws according to which events happen, and that, therefore, the frame of the world muſt be the effect of the wiſdom and power of an intelligent cauſe; and thus to confirm the argument taken from final cauſes for the exiſtence of the Deity. It will be eaſy to ſee that the converſe problem ſolved in this eſſay is more directly applicable to this purpoſe; for it ſhews us, with diſtinctneſs and preciſion, in every caſe of any particular order or recurrency of events, what reaſon there is to think that ſuch recurrency or order is derived from ſtable cauſes or regulations inna-ture, and not from any of the irregularities of chance.

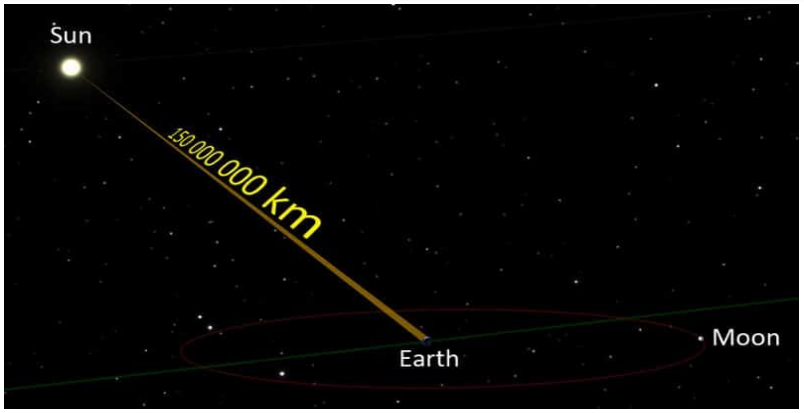
Ver artículo original

Estadísticamente hablando, no es correcto decir que algo tiene probabilidad cero sólo por el hecho de que no fue visto antes

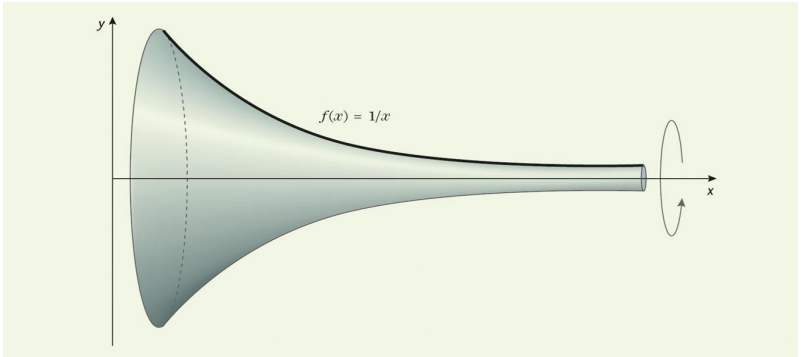
La ausencia de evidencia
no es evidencia de ausencia

La intuición como inferencia estadística

¿Qué altura tendría una hoja de papel doblada 50 veces?



¿Puede existir un cuerpo de volumen finito y superficie infinita?



¿Uruguay -como país- va a seguir existiendo dentro de 20 años?

Interpretación de las probabilidades

El test de Covid tiene una sensibilidad del 85%.

Me hago el test y me da Positivo.

¿Qué probabilidad hay de que tenga Covid efectivamente?

Falacia de tasa base

Ignorar la información que describe a la mayoría de los casos.

Repaso de probabilidades

Eventos independientes

Si la ocurrencia de un evento no tiene ningún efecto sobre la ocurrencia de otro, se dice que los dos eventos son independientes.

Matemáticamente, se dice que dos eventos A y B son independientes si:

$$P(A \cap B) = P(AB) = P(A) * P(B)$$

Por lo general, no es tan fácil identificar eventos independientes.

Repaso de probabilidades

Probabilidad condicional

La probabilidad condicional se define como la probabilidad de un evento A, dado que ya ha ocurrido otro evento B (es decir, A condicional B).

Esto está representado por P (A | B) y podemos definirlo como:

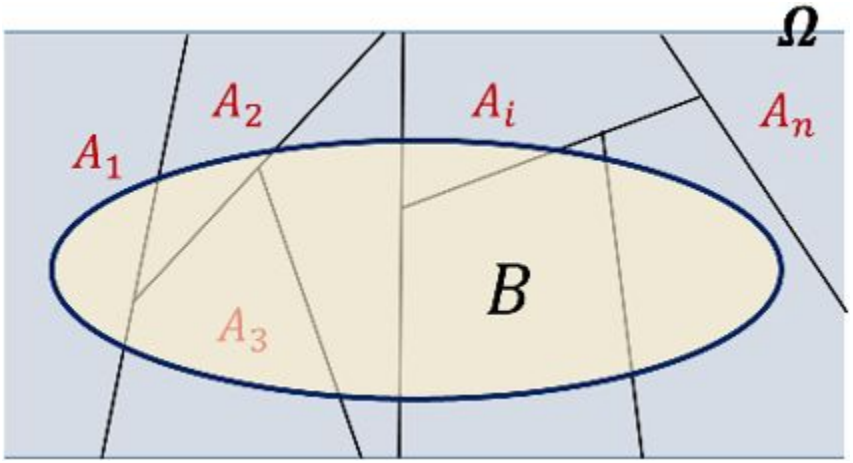
$$P(A \mid B) = P(A \cap B) / P(B)$$

Repaso de probabilidades

Probabilidad total (Laplace)

Si Ai son particiones de un espacio muestral y B un evento cualquiera, entonces:

$$P(B) = P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2) + \cdots + P(B \mid A_n) \cdot P(A_n)$$



Repaso de probabilidades

Ejemplo Covid:

La prevalencia del Covid en la población es de 0.001 (1 cada 1000)

P (Prueba = Positiva | Covid = Verdadero) = 0.85

P(Covid = Verdadero) = 0.001

P(Prueba Positiva & Covid Verdadero) =

P (Prueba = Positiva | Covid = Verdadero) * P(Covid = Verdadero)

= 0.85 *0.001 = 0.085%

Aprender de la experiencia

- Por ejemplo, dado el siguiente conjunto de datos:

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

queremos inferir si un día

<soleado, frío, alta, fuerte>

corresponde Juega = SI o NO.

Aprender de la experiencia

$$P(jugar = si) = \frac{9}{14}$$

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

<soleado, frío, alta, fuerte>

Aprender de la experiencia

$$P(jugar = no) = \frac{5}{14}$$

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

<soleado, frío, alta, fuerte>

Aprender de la experiencia

$$P(tiempo = soleado \mid jugar = si) = \frac{2}{9}$$

$$P(temp = frio \mid jugar = si) = \frac{3}{9}$$

$$P(humedad = alta \mid jugar = si) = \frac{3}{9}$$

$$P(viento = fuerte \mid jugar = si) = \frac{3}{9}$$

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	**Alta**	Suave	Sí
4	Lluvioso	Templado	**Alta**	Suave	Sí
5	Lluvioso	**Frío**	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	**Frío**	Normal	**Fuerte**	Sí
8	Soleado	Templado	Alta	Suave	No
9	**Soleado**	**Frío**	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	**Soleado**	Templado	Normal	**Fuerte**	Sí
12	Nuboso	Templado	**Alta**	**Fuerte**	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

<soleado, frío, alta, fuerte>

Aprender de la experiencia

$$P(tiempo = soleado \mid jugar = no) = \frac{3}{5}$$

$$P(temp = frio \mid jugar = no) = \frac{1}{5}$$

$$P(humedad = alta \mid jugar = no) = \frac{4}{5}$$

$$P(viento = fuerte \mid jugar = no) = \frac{3}{5}$$

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	**Soleado**	Caluroso	**Alta**	Suave	No
2	**Soleado**	Caluroso	**Alta**	**Fuerte**	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	**Frío**	Normal	**Fuerte**	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	**Soleado**	Templado	**Alta**	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	**Atla**	**Fuerte**	No

<soleado, frío, alta, fuerte>

Aprender de la experiencia

- Para los días en que Juega = SI

$$P(soleado|si) \,.\, P(frio|si) \,.\, P(alta|si) \,.\, P(fuerte|si) \,.\, P(si) = 0{,}0053$$

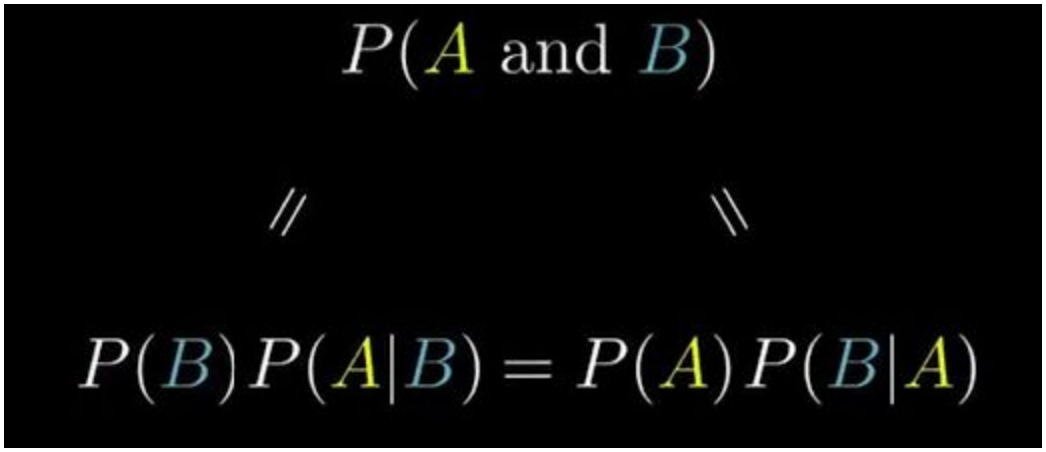
- Para los días en que Juega = NO

$$P(soleado|no) \,.\, P(frio|no) \,.\, P(alta|no) \,.\, P(fuerte|no) \,.\, P(no) = 0{,}0206$$

- Entonces se podría inferir que la mejor respuesta es NO

Formula de Bayes

Dado que P(A ∩ B) = P(B ∩ A)



Se desprende la fórmula:

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

que modela probabilísticamente las relaciones entre las probabilidades condicionales (o causas y efectos)

Teorema de Bayes

Sea H un espacio de hipótesis y D un conjunto de datos de entrenamiento

Dada h (una hipótesis de H) notamos:

- P(h) a la probabilidad inicial de la ocurrencia de h
  - Refleja un conocimiento inicial sobre la validez de h
  - Se llama **probabilidad previa o a priori**
- P(D) a la probabilidad previa de que se observen los datos de entrenamiento D. Es la evidencia.

Teorema de Bayes

- P(D|h) a la probabilidad de observar datos D dado que se cumple la hipótesis h

- P(h|D) a la probabilidad de la hipótesis h dado que son conocidos los datos D
  - Refleja un nivel de confianza de h despues de conocer los datos de entrenamiento D
  - Se llama **probabilidad a posteriori de h**

Teorema de Bayes

El teorema de Bayes proporciona un método para calcular la probabilidad posterior P(h|D) a partir de la probabilidad previa P(h)

$$P(h \mid D) = \frac{P(D \mid h) * P(h)}{P(D)}$$

$$P(h \mid D) = \frac{P(D \mid h) * P(h)}{P(D \mid h) * P(h) + P(D \mid \neg h) * P(\neg h)}$$

Hipótesis MAP y ML

- La ‘mejor’ hipótesis es la hipótesis más probable dados los datos: buscamos obtener la hipótesis h de H que maximiza P(h|D).
- Estas hipótesis son llamadas *Maximum A Posteriori* (MAP)

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

- El teorema de Bayes da una forma de obtener esta probabilidad:

$$\begin{aligned} h_{map} &\equiv argmax_{h \in H} P(h|D) \\ &\equiv argmax_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &\equiv argmax_{h \in H} P(D|h)P(h) \end{aligned}$$

- Si asumo que las hipótesis son equiprobables:

$$h_{ml} \equiv argmax_{h \in H} P(D|h)$$

Ejemplo

- Ejemplo: H={cancer, no cancer}

P(cancer) = 0,008 P(¬cancer)=0,992

P(test ⊕ | cancer) = 0,98 P(test ⊗ | cancer) =0,02

P(test ⊕ | ¬cancer) = 0,03 P(test ⊗ | ¬cancer) =0,97

- Si el test da positivo, ¿Qué deberíamos diagnosticar? ¿Qué hipótesis es MAP?

P(test ⊕| cancer) * P(cancer) = 0,98 * 0,008 = 0,0078

P(test ⊕|¬cancer) * P(¬cancer) = 0,03 * 0,992 = 0,0298

- El diagnóstico más probable es que el paciente está sano.

Teorema de Bayes - Ejemplo

- ¿Cuál es la probabilidad de ese resultado?

$$P(\neg cancer \mid test \oplus) = \frac{P(test \oplus \mid \neg cancer) * P(\neg cancer)}{P(test \oplus)}$$

$$= \frac{0.03 \times 0.992}{0.0078 + 0.0298}$$

$$= 79.15\%$$

Características de los métodos bayesianos

- Cada caso de entrenamiento cambia la probabilidad estimada de que una hipótesis sea correcta (soporta ruido).
- El conocimiento previo puede ser utilizado para determinar la probabilidad de una hipótesis.
- Pueden dar predicciones probabilísticas.
- Pueden clasificar nuevas instancias combinando distintas hipótesis.
- Algunos algoritmos tienen un costo alto.

¿Qué caracteriza a un sistema “inteligente”?

Este método provee una regla para actualizar nuestras creencias cada vez que obtenemos nuevas evidencias.

Sólo podemos considerar que un sistema es realmente inteligente si es capaz de observar su entorno y aprender de él.

La auténtica inteligencia reside en adaptarse, tener capacidad de integrar nuevo conocimiento, resolver nuevos problemas, y aprender de errores.

Dificultades del método

Requiere conocer valores iniciales de muchas probabilidades.

Cuando estas no son conocidas hay que estimarlas:

- conocimiento previo
- hipótesis
- distribuciones

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

35% -> 21%

Buscar en Inumet cantidad promedio de días de lluvia en Uruguay

Dificultades del método

El costo computacional para determinar la hipótesis más probable puede ser muy alto

- Lineal con el número de hipótesis candidatas
- Depende de la cantidad de posibles valores de cada variable

$$argmax_{h \in H} P(D \mid h) P(h)$$

Clasificador sencillo (Naive Bayes)

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

Simplificación:

independencia de atributos dada una hipótesis

P((a1,a2,...,aN) | H) = P (a1 | H) * .... * P(aN | H)

Clasificador sencillo

- Consideremos instancias de la forma <$a_1 ... a_n$>, y una función objetivo f que toma valores sobre un conjunto finito V.
- Buscamos:

$$v_{MAP} = argmax_{v_j \in V} P(v_j \mid a_1 \ldots a_n)$$

$$= argmax_{v_j \in V} \frac{P(a_1 \ldots a_n \mid v_j) \, . \, P(v_j)}{P(a_1 \ldots a_n)}$$

$$= argmax_{v_j \in V} P(a_1 \ldots a_n \mid v_j) \, . \, P(v_j)$$

- Utilizando al clasificador sencillo:

$$v_{NB} = argmax_{v_j \in V} \prod_i P(a_i \mid v_j) \, . \, P(v_j)$$

Clasificador sencillo

- Queremos inferir si un día

<soleado, frío, alta, fuerte>

corresponde Juega = SI o NO.

- Haciendo cuentas:

$$P(soleado|si).P(frio|si).P(alta|si).P(fuerte|si).P(si) = 0{,}0053$$

$$P(soleado|no).P(frio|no).P(alta|no).P(fuerte|no).P(no) = 0{,}0206$$

- Normalizando a 1 estas probabilidades: con un 79.5% de seguridad puedo afirmar que la respuesta es negativa.

Característica de Naive Bayes

No hay una búsqueda explícita en el espacio de posibles hipótesis.

Sólo necesitamos conocer P(Hj) y P(Ai | Hj)

Lo calculamos usando lo que tenemos en los datos de entrenamiento

Acumulación de experiencias y certezas

La regla de Bayes es útil para actualizar creencias a partir de nuevas evidencias, por lo que no están permitidos los argumentos de probabilidad 0 o 1.

Por más evidencia favorable que se acumule respecto de una creencia con probabilidad inicial menor a 1, jamás asignará a una creencia a posteriori el valor de certeza (1) ni de imposibilidad (0).

Estimación de probabilidades cuando hay pocos datos

PERO

- ¿Cómo calcularía $P(tiempo = nuboso \mid jugar = no)$?
- ¿Cómo clasificaría a

<nuboso, frío, alta, fuerte> ?

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

Estimación de probabilidades cuando hay pocos datos

- No siempre es buena la aproximación utilizando la frecuencia

$$p = \frac{|evento|}{|oportunidades|}$$

- Cuando no hay ejemplos para algunos casos, la probabilidad asignada es cero, con lo cual "anula" todo el término del clasificador.

Estimación de probabilidades cuando hay pocos datos

- m-estimador: $\frac{e + m \cdot p}{n + m}$

donde:

  - p es la estimación a priori de la probabilidad buscada
  - m es el "tamaño equivalente de muestra".

- La fórmula puede verse como aumentar la muestra con m ejemplos, distribuidos según p.
- Por ejemplo, tomando valores equiprobables:

$$p = \frac{1}{|valores|}$$

Estimación de probabilidades cuando hay pocos datos

$$P(tiempo = nuboso \mid jugar = no)$$

Calculado como frecuencia = 0 / 5 = 0

Calculado como m-estimador: $\frac{e + m \cdot p}{n + m}$

con

e=0

p= ⅓

n=5

m=2

P(nuboso | no) = (0 + 2 * ⅓) / 7 = 0.095

#	Tiempo	Temperatura	Humedad	Viento	Juega
1	Soleado	Caluroso	Alta	Suave	No
2	Soleado	Caluroso	Alta	Fuerte	No
3	Nuboso	Caluroso	Alta	Suave	Sí
4	Lluvioso	Templado	Alta	Suave	Sí
5	Lluvioso	Frío	Normal	Suave	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nuboso	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Suave	No
9	Soleado	Frío	Normal	Suave	Sí
10	Lluvioso	Templado	Normal	Suave	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nuboso	Templado	Alta	Fuerte	Sí
13	Nuboso	Caluroso	Normal	Suave	Sí
14	Lluvioso	Templado	Atla	Fuerte	No

Atributo	Datos (n=5)	Frecuencia	(m=1)-estimator	(m=2)-estimator	(m=5)-estimator
soleado	3	0,6	0,556	0,524	0,467
nuboso	0	0	0,056	0,095	0,167
lluvioso	2	0,4	0,389	0,381	0,367

APRENDIZAJE BAYESIANO

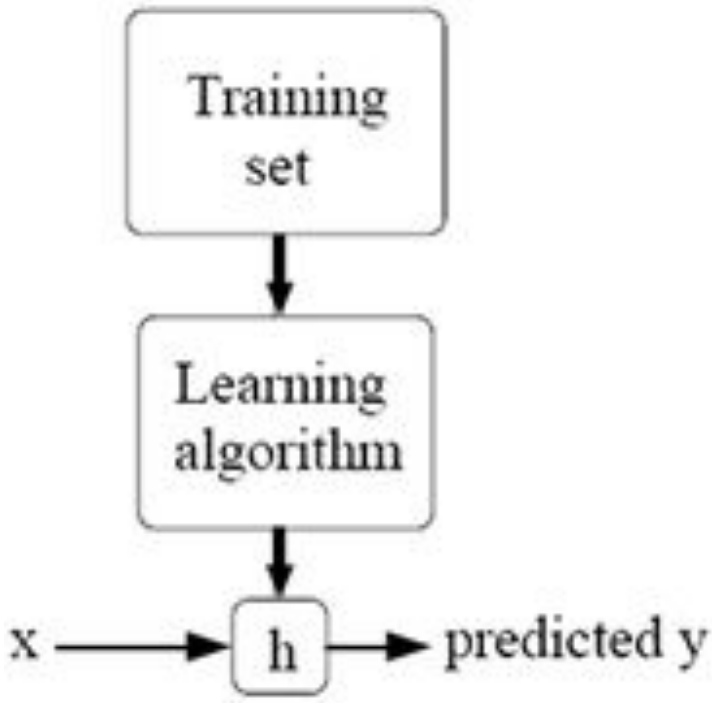
Frase típica del grupo de Whatsapp de padres

Mi hijo se olvidó de la mochila en la escuela.

Agenda de la semana

- Introducción
- Teorema de Bayes
- Hipótesis MAP
- Clasificador sencillo

- Resumen de clase anterior
- Tarea 2
- Clasificador de texto
- Clasificador óptimo
- Algoritmos MAP



RESUMEN DE CLASE ANTERIOR

- Probabilidad condicional
- Fórmula de Bayes
- Hipótesis MAP
- Naive Bayes
- m-estimador

Tarea 2: Aprendizaje bayesiano

Problema

Considere el problema de predicción de palabras que todos utilizamos diariamente en servicios como el buscador de Google o los teclados de los celulares.

Se propone implementar una solución a dicho problema utilizando el algoritmo Naive Bayes, asumiendo independencia en el orden de aparición de las N últimas palabras de la frase que se viene escribiendo (N es un hiperparámetro de la implementación).

Para entrenar el modelo cada grupo deberá descargar de su propio Whatsapp[1] el contenido del grupo que mas conversaciones tenga. Dicho contenido no debe ser parte de la entrega; es sólo a los efectos de que las pruebas que hagan tengan sentido para ustedes, y para estandarizar el formato de carga de datos. Con dichos datos se podrán aproximar las probabilidades requeridas por el método.

Para probarlo deberan utilizar un simulador de cliente implementado en un cuaderno provisto[2]. Este simulador permite crear frases *palabra por palabra*, mostrar las palabras recomendadas (resultado del algoritmo a implementar), aceptar la recomendación (con ENTER) o ignorarla (escribiendo una nueva palabra para la frase), y recomenzar con una frase nueva (con PUNTO).

La entrega deberá ser realizada en una copia de dicho cuaderno, donde deberán:

1. Modificar la función `recomendacion_bayesiana` con su implementación del algoritmo descrito.
2. Permitir cargar el archivo de entrenamiento con un CSV provisto por Whatsapp y considerar solo el texto de los mensajes (ignorando las columnas fecha y autor)
3. Comparar el desempeño del algoritmo variando el hiperparámetro N (con valores 1, 2, 3 y 4)
4. Reentrenar el modelo al finalizar cada frase (cada vez que se ingresa un PUNTO). Esto puede ser computacionalmente pesado pero muestra la capacidad del método para adaptarse a nueva evidencia.

Este algoritmo será intenso en uso de CPU. Se deberán utilizar las estructuras de datos más adecuadas provistas por el lenguaje Python para que la implementación sea lo más limpia y eficiente posible.

Entregables

El informe a entregar debe ser una copia del cuaderno provisto, donde se incluirá:

- el código escrito para resolver el problema
- un informe con las pruebas realizadas y los resultados obtenidos

Fecha límite de entrega

Miércoles 13 de setiembre (inclusive).

Frase típica del grupo de Whatsapp de padres

Mi hijo se olvidó de la mochila en la escuela.

Algunos tips

- Preprocesar los datos de entrenamiento

```
cat _chat.txt |
tr -cs '[:alpha:]' '\n' |
tr 'A-Z' 'a-z' |
sort |
uniq -c |
sort -nr |
```

- Reducir el volumen cruzando con diccionarios en español

Algunos tips

- Utilizar estructuras de datos adecuadas (diccionarios, listas, sets)

```
P[h] = 0.0034          P(h)   probabilidad previa
PD[h][d] = 0.0021      P(d|h) evidencia para h

P = {
  "mochila" : {
    "perdió" : 0.005,
    "la" : 0.045,
    "nunca" : 0.00009
       :
  },
  "buzo" : {
    "perdió" : 0.0049,
    "rompió" : 0.01
       :
  }
    :
```

Algunos tips

- Aprovechar servicios de las estructuras de datos

  - Construcción dinámica de diccionarios

    ```
    P["mochila"]["perdió"] = 0.005
    ```

  - Valores por defecto al buscar claves que no existen

    ```
    P["mochila"].get("se",P_nada)
    ```

    en vez de

    ```
    P["mochila"]["se"]
    ```

  - Sublistas
    - `frase = ["mi", "hijo", "se", "olvidó", "de", "la"]`
    - `frase[2:4]`
    - `frase[-5:]`

```
D = ["mi", "hijo", "se", "olvidó", "de", "la"]

Horizonte = 4
h_MAP = ""
p_MAP = 0

for h in P:

  prob = P[h]
  for d in D[-Horizonte:]:
    prob = prob * PD[h].get(d, P_nada)

  if prob > p_MAP:
    h_MAP , p_MAP = h , prob

print(h_MAP)
```

CLASIFICADOR DE TEXTO

Clasificador de Texto

- Queremos clasificar textos en categorías.
- Por ejemplo:
  - *Me dio error al pagar con la tarjeta.* → Pagos
  - *No encuentro el producto que busco.* → Soporte
  - *El producto estaba defectuoso.* → Envíos

Clasificador de Texto

- Podría entrenar a un clasificador bayesiano y utilizarlo...

$$v_{NB} = argmax_{v_j \in V} P(v_j) \cdot \prod_i P(a_i = w_k \mid v_j)$$

- Por ejemplo, para «El paquete llegó roto.»

$$P(a_1 = el \mid pagos) \cdot P(a_2 = paquete \mid pagos) \cdot P(a_3 = llego \mid pagos) \cdot P(a_4 = roto \mid pagos) \cdot P(pagos)$$

$$P(a_1 = el \mid soporte) \cdot P(a_2 = paquete \mid soporte) \cdot P(a_3 = llego \mid soporte) \cdot P(a_4 = roto \mid soporte) \cdot P(soporte)$$

$$P(a_1 = el \mid envios) \cdot P(a_2 = paquete \mid envios) \cdot P(a_3 = llego \mid envios) \cdot P(a_4 = roto \mid envios) \cdot P(envios)$$

Clasificador de Texto

- Se asume que las palabras son independientes entre sí.

- Con esa simplificación, de todas formas, lo anterior no es práctico.

$$P(a_i = w_k \mid v_j)$$

- Con un vocabulario de 50 mil palabras, textos de a los sumo 40 palabras, y 3 categorías tengo $3X40X50000 = 6.10^6$ parámetros

Clasificador de Texto

- Agregamos otra suposición (irreal): la aparición de las palabras es independiente de su posición y de la clasificación del mensaje.

$$P(w_k \mid v_j)$$

- Por ejemplo, para «El paquete llegó roto.» y «Llegó roto el paquete.» se calculan las mismas probabilidades:

$$P(el \mid pagos) \cdot P(paquete \mid pagos) \cdot P(llego \mid pagos) \cdot P(roto \mid pagos) \cdot P(pagos)$$

$$P(llego \mid pagos) \cdot P(roto \mid pagos) \cdot P(el \mid pagos) \cdot P(paquete \mid pagos) \cdot P(pagos)$$

- La cantidad de parámetros se reduce a 150 mil.

Clasificador de Texto

- Notamos
  - v1, v2 y v3 a los tres valores de la función objetivo.
  - wk a cada una de las palabras del vocabulario

- Debemos calcular: $$v_{NB} = argmax_{v_j \in V} P(v_j) . \prod_i P(w_k | v_j)$$

- Estimamos:

$$P(v_j) = \frac{|documentos_{vj}|}{|documentos|}, \forall v_j \in Categorias$$

$$P(w_k | v_j) = \frac{n_{kj} + 1}{n_j + |vocabulario|}, \forall w_k \in Vocabulario, v_j \in Categorias$$

Clasificador de Texto

- Para clasificar un nuevo documento:

$$v_{NB} = argmax_{v_j \in V} P(v_j) \cdot \prod_{a_i \in words} P(a_i \mid v_j)$$

words ← palabras del documento que están en el vocabulario

Probabilidades logarítmicas

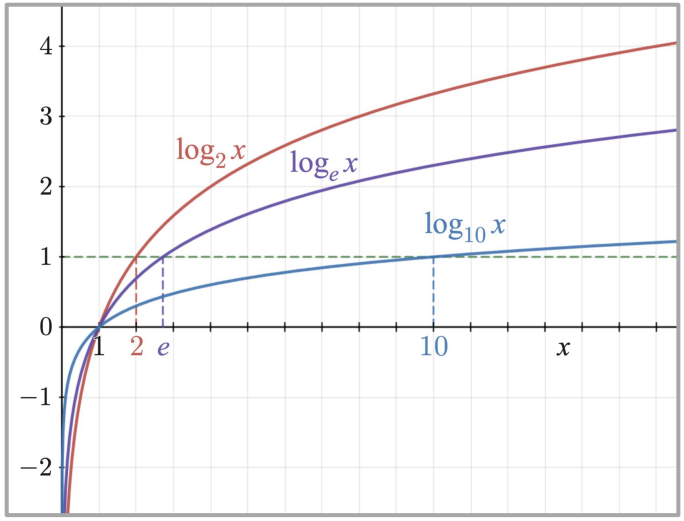
- Las productorias de muchas probabilidades pueden generar números muy pequeños e inestabilidad numérica
- Propiedades de los logaritmos:
  - Convierten productos en sumas

$$\log(\mathrm{P}(E) \cdot \mathrm{P}(F)) = \log \mathrm{P}(E) + \log \mathrm{P}(F)$$

  - Amplían el rango de valores

$$0 \leq \mathrm{P}(E) \leq 1$$
$$-\infty \leq \log \mathrm{P}(E) \leq 0$$

  - Si a < b entonces Log(a) < Log(b)



Probabilidades logarítmicas

- Naive Bayes usando probabilidades logarítmicas:

$$\text{argmax} \quad \left[\prod_{j=1}^{m} P(F_j \mid H_i)\right] P(H_i)$$

$$\text{argmax} \quad \left[\sum_{j=1}^{m} \ln P(F_j \mid H_i)\right] + \ln P(H_i)$$

Clasificador de Texto

- Usualmente se aplican otras simplificaciones, como limitar el vocabulario y eliminando palabras vacías (*stop words*)
- ¿Qué sucede si una palabra no se da en alguna clase de mi conjunto de entrenamiento?
- ¿Qué sucede si una palabra de un nuevo mensaje no se había visto antes?

TF-IDF

- Método estadístico para evaluar la importancia de las palabras para un documento dentro de un conjunto de documentos
- TF = Term Frequency: cantidad de veces que aparece la “t” en el documento “d” $tf\,(t, d) = log\,(1 + freq\,(t, d))$
- IDF = Inverse Term Frequency: prevalencia de la palabra “t” en el universo de documentos “D” (tamaño N)

$$idf\,(t, D) = log\,(\ \frac{N}{count\,(d \in D : t \in d)}\ )$$

entonces

$$tf\,idf\,(t, d, D) = tf\,(t, d)\ .\ idf\,(t, D)$$

Clasificador de Texto

- Modelo sencillo
- Se calcula rápido
- Funciona relativamente bien

¿Qué pasa con emails cortos?

¿Qué pasa con las palabras repetidas?

Clasificador bayesiano óptimo

- Sabemos determinar la(s) hipótesis más probable(s) dado un conjunto de entrenamiento.

- Pero, ¿la **hipótesis** más probable es la que nos da la **clasificación** más probable de una nueva instancia?

Clasificador bayesiano óptimo

- La clasificación más probable se obtiene combinando los resultados de todas las hipótesis ponderada por sus probabilidades posteriores:

$$P(v|D) = \sum_{h \in H} P(v|h)P(h|D)$$

- Clasificación bayesiana óptima:

$$argmax_{v \in V} \sum_{h \in H} P(v|h)P(h|D)$$

Clasificador bayesiano óptimo

- Por ejemplo, sea un espacio con tres hipótesis y una instancia x que queremos clasificar:

  $P(h_1|D) = 0,4$ $P(h_2|D)=0,3$ $P(h_3|D) = 0,3$

  $h_1(x)= \oplus$ $h_2(x)= \otimes$ $h_3(x)= \otimes$

- ¿Qué valor devuelve una hipótesis MAP?

- ¿Cuál es la clasificación más probable?

Clasificador bayesiano óptimo

$P(h_1|D) = 0{,}4 \quad P(h_2|D)=0{,}3 \quad P(h_3|D) = 0{,}3$

$P(\otimes|h_1)= 0 \quad P(\otimes|h_2)= 1 \quad P(\otimes|h_3)= 1$

$P(\oplus|h_1)= 1 \quad P(\oplus|h_2)= 0 \quad P(\oplus|h_3)= 0$

$$P(\oplus|D) = \sum_{h \in H} P(\oplus|h)\, P(h|D) = 0{,}4$$
$$P(\otimes|D) = \sum_{h \in H} P(\otimes|h)\, P(h|D) = 0{,}6$$

$$\text{argmax}_{v \in \{\oplus, \otimes\}} \sum_{h \in H} P(v|h).P(h|D) = \otimes$$

Clasificador bayesiano óptimo

- Un clasificador bayesiano óptimo es cualquier sistema que clasifique las instancias de esta manera (independiente de la implementación).
- Estos sistemas **maximizan la probabilidad de clasificar correctamente nuevas instancias** dado el conjunto de entrenamiento y las probabilidades a priori de las hipótesis.
- La desventaja de estos métodos son muy costosos (debemos calcular la probabilidad a posteriori de todas las hipótesis).
- Si bien el clasificador bayesiano óptimo es costoso, lo podemos usar como parámetro para evaluar otros métodos.

Bayes y el Aprendizaje Conceptual

- ¿Cómo se puede aplicar el teorema de Bayes en el Aprendizaje Conceptual?
- Podemos buscar h_{MAP} en H.
- Algoritmo Fuerza-Bruta

  1. Para cada hipótesis h de H calculamos: $$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

  2. Damos como salida $$h_{map} \equiv argmax_{h \in H} P(h|D)$$

- No es una buena opción cuando H es un espacio muy grande o infinito.

Bayes y el Aprendizaje Conceptual

- Para aplicar el algoritmo debemos calcular P(D|h) y P(h)
- Estas probabilidades se pueden elegir de acuerdo a los conocimientos previos que tengamos sobre el espacio de búsqueda
- Por ejemplo, supongamos que: $$P(D \mid h) = \begin{cases} 1 \text{ si } d_i = h(x_i) \forall d_i \in D \\ 0 \text{ en caso contrario} \end{cases}$$
  - El concepto objetivo está en nuestro espacio H.
  - A priori todas las hipótesis son equiprobables: $$P(h) = \frac{1}{|H|}$$
  - El conjunto D no tiene ruido

Bayes y el Aprendizaje Conceptual

- Bajo estos supuestos, ¿cómo estimamos P(D)?

$$P(D) = \sum_{h \in H} P(D|h_i) \,.\, P(h_i) = \sum_{h \in VS} 1.P(h_i) + \sum_{h \notin VS} 0.P(h_i) = \sum_{h \in VS} \frac{1}{|H|} = \frac{|VS_{H,D}|}{|H|}$$

- Si considero una hipótesis inconsistente con los datos, su probabilidad es nula $P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{0 * P(h)}{P(D)} = 0$

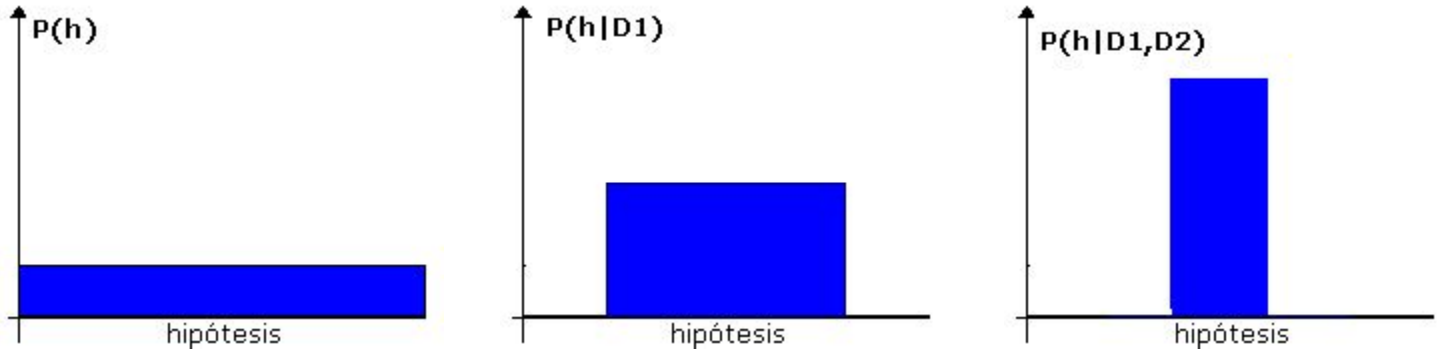
- En cambio, si es consistente:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = 1.\frac{1}{|H|} \cdot \frac{|H|}{|VS_{H,D}|} = \frac{1}{|VS_{H,D}|}$$

- Por lo tanto, bajo los supuestos establecidos, toda hipótesis consistente con el conjunto de entrenamiento es MAP.

Bayes y el Aprendizaje Conceptual

- A medida que agrego datos al entrenamiento, disminuye $VS_{H,D}$ y aumenta la probabilidad de cada hipótesis consistente.



- Todo algoritmo (consistente) da como resultado una hipótesis MAP, si se cumple a priori una distribución uniforme sobre H.

Bayes y el Aprendizaje Conceptual

- Find-S y Candidate-Elimination no manejan ningún tipo de probabilidad y sin embargo son algoritmos MAP porque generan hipótesis consistentes con los datos.

- ¿Existen otras condiciones bajo las cuales Find-S sea un algoritmo MAP?

  Sí, cuando la distribución sobre H asigna más probabilidad a las hipótesis más específicas (en vez de 1 / | H |) y no hay ruido en la entrada.

- Podemos caracterizar los algoritmos aun cuando éstos no utilizan explícitamente probabilidades: basta encontrar las probabilidades P(D|h) y P(h) bajo las cuales los algoritmos dan hipótesis MAP.

- Esta es una alternativa al sesgo para caracterizar los supuestos bajo los cuales un algoritmo aprende.

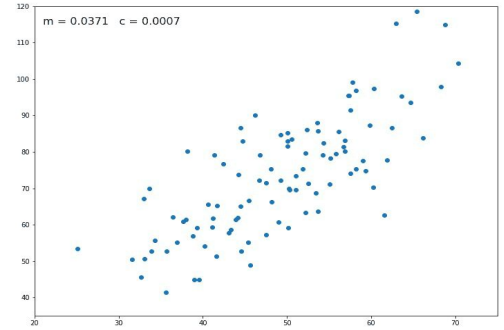
Siguientes clases

4/Set - Aprendizaje por casos

6/Set - Aprendizaje no supervisado

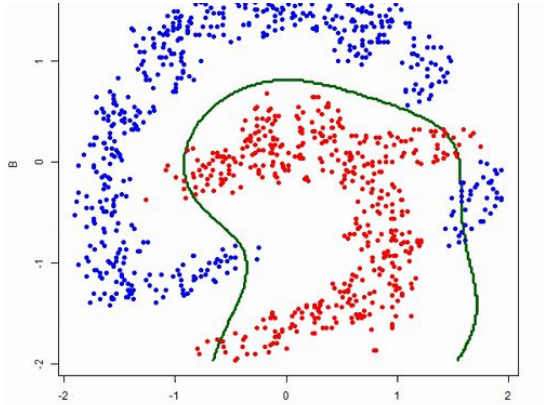
11 y 13/Set - Aprendizaje por refuerzos

Parciales



2/Oct - Regresión lineal

4/Oct - Regresión logística



Lecturas previas recomendadas para Regresiones:

- Repaso de álgebra lineal
- Linear Algebra Review and Reference
- Numpy: computación científica con Python