Aprendizaje Bayesiano

→ capítulo 6 libro

[La ausencia de evidencia
no es evidencia de ausencia

datos, resultados → [A.A] → reglas

datos nuevos → [modelo] → Prediciones

- Observar el entorno y aprender de él

CARACTERÍSTICAS → capacidad de adaptarse y capacidad de procesar datos contradictorios

Alg. Bayesianos → decisiones difusas (se manejan con prob.)

Inferencia estadística es:

¿cómo encontrar a partir de algo que parece caos un cierto orden?

- David Hume → idea: los milagros son imposibles
  ↳ 'es imposible romper las leyes de la naturaleza'
- Richard Price →

conclusión:

Razonamiento Bayesiano:

(1) Ningún evento puede tener probabilidad $\emptyset$ o 1
(no es correcto)

· La intuición como inferencia estadística
⇒ es muy difícil decir que un evento no va a pasar nunca

Repaso Prob.

- Eventos independientes $(A, B) \Rightarrow P(A \cap B) = P(A) \cdot P(B)$
- Prob. condicional : $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

- Prob total ( ) :

ejemplo — Covid — prevalencia de covid en pobl. = 0,001

$P(\text{Prueba} = + \mid \text{Covid} = +) = 0{,}85$

$P(\text{Covid} = +) = 0{,}001$

ejemplo — Aprender de la experiencia

#	Tiempo	temp	Humedad	Viento	Juega
.	Soleado				
:	Nublado				
.	Lluvioso				
:					

Quiero predecir (inferir) si la persona va a jugar al golf un día < s, f, a, f > → $P(\langle s,f,a,f\rangle \mid si)$

→ saco prob. a partir de los datos que tengo

ej: $P(\text{jugar} = si) = 9/14$ → para el conj. de datos que tengo

$P(\text{tiempo} = f \mid \text{jugar} = si) = 3/9$

$P(\text{humedad} = \quad \mid \quad) =$

⋮

usando la form. prob. cond →

$P(\text{soleado} \mid si) =$ . . .

$=$

$P(\text{soleado} \mid no) =$

$P(\text{soleado} \mid si) < P(\text{soleado} \mid no)$

⟹ la mejor respuesta es no

- se asume que las caract. del clima son independientes

en este caso h es si juego o no y D la combinación del clima

Fórmula de Bayes

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

D = evidencia
h = hipótesis

teo de Bayes:

Sea H un espacio de hipótesis y D un conj. de datos de entrenamiento

- $P(D|h) \sim$ prob
- $P(h|D) \sim$

es la que nos interesa conocer

$$P(h|D) = \frac{}{P(D|h) \times P(h) + P(D|\neg h)}.$$

Hipótesis MAP y ML

Buscamos h que maximiza $P(h|D)$

- MAP $\longrightarrow P(h|D) =$

$$h_{opt} = \max_{h \in H} \{ P(D|h) \cdot P(h) \}$$

si asumo hipótesis equipr. $\longrightarrow h_{opt} = \max_{h \in H} \{ P(D|h) \}$

CARACTERÍSTICAS

-
- Soporta datos con ruido
- Las implementaciones tienen un costo computacional alto

Fortalezas:

- Actualizar nu → cada vez que hay datos nuevos, podemos re-entrenar y volver a calcular las prob. dando que sea más preciso
- Obser
-

Dificultades:

- Conocer muchas probabilidades ← conocer valores iniciales
- Hipótesis y distribuciones
- Costo computacional (argmax P( ).P( )) → se pueden tomar prob. de otros lados
  - ↳ depende de cant. de hipótesis y cant. de posibles valores de cada variable
  - ↳ gran limitante

este es el que se usa

→ CLASIFICADOR SENCILLO (Naive Bayes) · ⊕ en el web

simplificación: • independencia de atributos
•

⋮

Sea < > .

$$\Rightarrow V_{NB} = \max_{v_i \in V} \prod P(\mid).P( )$$

características:

+

+

→ La prob. de un atributo nunca es $\emptyset$ o 1 !!!

Acumulación de experiencias y certezas

¿cómo estimar la prob. cuando tenemos pocos datos?

→ esto genera prob. en la mult. de prob.

sol: TÉCNICA de M-ESTIMADOR

$$= \frac{c + m \cdot p}{n + m}$$

p =
m = tamaño de la muestra

→ se agregan valores 'ficticios' adicionales

hay que determinar m

suavizado de Laplace

atributo	n datos	frecuencia	(m=1) - estimador	m=2	m=5

→

Resumen

- Prob. Cond $\longrightarrow P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B)}$
- Fórmula de Bayes $\longrightarrow P(h|D) = \dfrac{P(D|h) \times \overbrace{P(h)}^{\text{probabilidad a priori}}}{P(D)}$
- Hipótesis MAP $\rightarrow H_{map} = \underset{h \in H}{\operatorname{argmax}}\ P(D|H) \cdot P(h)$
  - ↳ max. probabilidad a posteriori
- Naive Bayes $P(\langle a_1, \ldots, a_n \rangle | H) = P(a_1|h) \times \ldots \times P(a_n|h)$
  $\Rightarrow \operatorname{argmax} p(h) \cdot \prod P(a_i|h)$
- m-estimador
  - ↳ técnica para evitar que las prob. den $\varnothing$
  - $\Rightarrow$ aumento espacio muestral
    $f = 1/n \rightarrow f = \dfrac{1+}{n+m}$

TAREA 2 - Predicción de Palabras

- usar alg. Naive Bayes
  simplificaciones [

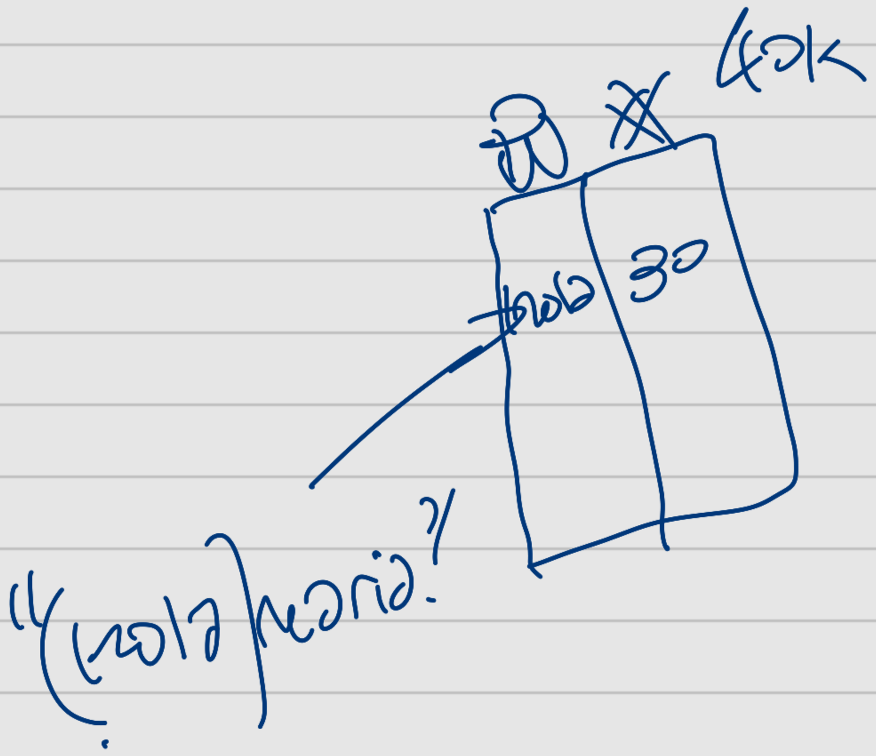
- para trabajo: cuaderno dado (conjunto)

⟹ modificar func. bayesiana

Solo quedarse con los
mensajes
locales no nombre y fecha

n-gramas

→ evitar loops
→ no más de 10 min



① Preprocesar los datos
  - palabras en lista
  - y minúsculas
  - ordenar

→ pdok lo pasa

② Crear con palabras del diccionario
③ Estructura de datos

ej: mi hijo perdió la ______
↓
para cada palabra del diccionario calculo:

ej: mochila

N-últimos palabras
↑
hiperparámetro

$P(mochila) \times P(la \mid mochila) \times P(hijo \mid mochila) \times P(perdió \mid mochila) \times \ldots$

→ tomo la palabra con mayor probabilidad

(8) elegir criterio que para cuando una
prob es $\emptyset$
→ justificarlo

(*) PM [h][d]
↳ prob. h dado once d

P dewocodigo
```
D = [ ' ' ,
horizonte = 4
h-MAP = ''
P_map = Ø
for h in P :
    prob = P[h]
    for d in D[-horizonte:] :
        prob = prob x PD[h].get(d, P-nool)

    if prob > P-MAP :
        h_map , P-map = h , prob

print (h-map)
```
← palabra recomendada

ej: cant. total palabras

valor default

⇒ entrenar cuando termino una frase
palabras nuevas agregar
↳ (puede ser cada n-frases)
ej: 10

(*) Poner foco en el entrenamiento

(*) Verificar que aprendió

(*) Asumimos que en el entrenamiento ingresan palabras

(*) ¡Usar log!

Predictor de texto ← uso del alg.

Clasificador de Texto

→ uso del alg.

· clasificar textos en categorías

→ muchas palabras ⇒ prob. chicos

⇒ usar prob log.

∘ considerando el orden de palabras

⋮

⇒ Asumiendo ind. de palabras

↳ Reducimos tamaño del modelo

calcular $V_{NB} = \underset{v_j}{\operatorname{argmax}}$

⋮

problema ⟶

sol. → aplicar $\underline{\underline{\log}}$

prop:

$$\log(P(E) \cdot P(F)) = \log(P(E)) + \log(P(F))$$

⇒ achic rango de valores

$$-\infty \leq \log(P(E)) \leq 0$$

-prob +prob

⇒ si $a < b$

⇒ $\log(a) < \log(b)$

$$\log\left(P(v_j) \prod P(a_i \mid v_j)\right)$$

$$\log(P v_j) + \sum \log(P a_i v_j)$$

Otras simplificaciones:

- limitar vocabulario
- eliminar palabras vacías

¿ ?

¿ ?

decisiones que hay que tomar en las implementaciones (justificando)

Clasificadores de texto

→ modelo sencillo

→ se calcula rápido

→

Clasificadores Bayesianos óptimos