

Aprendizaje Supervisado

Facundo Carrillo
fcarrillo@udesa.edu.ar

Taller I
2022-09-10

¿Cuándo?

- Sábado 10-09-2022:
 - Métricas
 - Validación cruzada
 - Naive Bayes
 - Gaussian Naive Bayes
- Sábado 24-09-2022: Árboles de decisión
- Sábado 01-10-2022: Vecinos más cercanos + Scikit-learn como framework + Presentación TP

Organización:

- Clase expositiva cortas (esperemos <1hs)
- Taller con ejercicios

Comunicación (esperemos mucha):

- En Clase!
- Mail

Hoy: Taller I

Temas:

- Métricas
- Validación cruzada
- Naive Bayes
- Gaussian Naive Baye

¿Dónde estamos parados?

- ¿Qué es un clasificador?
- ¿Para qué sirve?
- ¿Qué diferencia hay con un regresor?

Bibliografía:

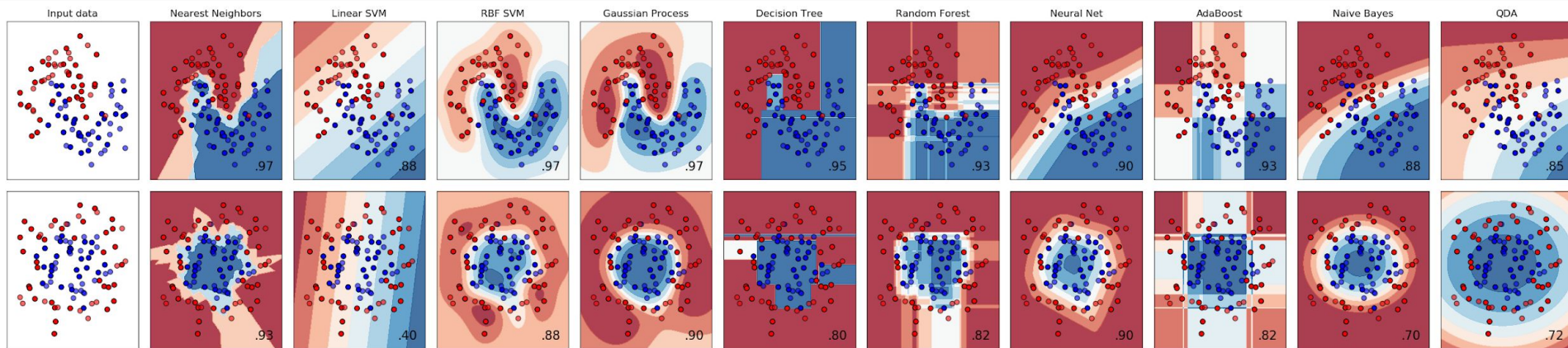
- Mitchell Machine Learning: [website libro](#) , [pdf](#)
- Scikit-learn: <https://scikit-learn.org/>

- Modelos para resolver el problema de clasificación...
- Aprendizaje Supervisado
- Identificar cual o cuales categorías (finitas) son las correctas para una instancia a **partir de muestras ya conocidas**
- Las categorías:
 - No necesariamente tienen orden (total o parcial)
 - Ej: **spam** vs **no-spam**
 - No necesariamente son 2 (es decir puede ser multiclase).
 - Ej: predecir a qué **idioma** pertenece un texto
 - No necesariamente una sola categoría/etiqueta (multi label).
 - Ej: la descripción de un nuevo item en un marketplace podría indicarnos que el producto es de categoría **electrodoméstico** pero también de categoría **cocina**

Clasificadorosssssss

- Asunciones que hacemos
 - Árboles cortos vs árboles largos
 - K-vecinos: el más cerca, el más símil
 - Naive Bayes: independientes

Más en el Capítulo 2 Mitchell *Concept Learning*



Definiciones, objetos y flujo de trabajo

Instancias:

Elementos con los que queremos trabajar: un texto, un vector de números, una imagen, etc.

Instancias de entrenamiento:

Las instancias acompañadas por un valor de la categoría a predecir. Ej:

- Texto de un email acompañado por la etiqueta spam
- Una imagen acompañado por la etiqueta hay_un_perro

Flujo:

1. Entreno un modelo usando datos de entrenamiento
2. Predigo nuevos datos

Artesanal:

¿Qué modelo? ¿Cómo entreno? ¿Estoy seguro de que generaliza mi modelo? ¿Cuáles son las limitaciones que tiene? ¿Cuán bien anda?

Clasificadores: ¿Cuán bien funciona mi modelo?

¿Cómo sé si mi modelo funciona bien?
¿Qué significa que funcione bien mi modelo?
¿Generalizará bien para datos no vistos?
...

Muchas preguntas con muchas estrategias diferentes para contestarlas!

Matriz de confusión

Tabla de doble entrada (definamos SPAM como la clase positiva)

		Clase predicha	
		NO-SPAM	SPAM
Clase real	NO-SPAM	TN	FP
	SPAM	FN	TP

Ejemplos: Supongamos que tenemos 50 spam y 50 no-spam y 3 clasificadores distintos

		N	S
Clase real	N	40	10
	S	0	50
		Clasificador 1	

		N	S
Clase real	N	50	0
	S	10	40
		Clasificador 2	

		N	S
Clase real	N	45	5
	S	5	45
		Clasificador 3	

Matriz de confusión

		N	S
Clase real	N	40	10
	S	0	50

Clasificador 1

		N	S
Clase real	N	50	0
	S	10	40

Clasificador 2

		N	S
Clase real	N	45	5
	S	5	45

Clasificador 3

¿Cuál anda mejor?

Clasificador de **spam**:

- ¿Son igual de importantes los errores? FN vs FP
- Si un **spam** me llega al **inbox** ¿cuánto pierdo?
- Si un **mail genuino** me llega a **spam** ¿cuánto pierdo?

Solución de compromiso de información parcial

Pensemos en variables categóricas binarias

1. Accuracy: $(TP+TN) / TOTAL$

En el ejemplo: $(\# \text{ Spam bien tagueado} + \# \text{ no-spam bien tagueado}) / \# \text{ mails}$

2. Precision: $(TP) / (TP+FP)$

En el ejemplo: $(\# \text{spam_predicho_spam}) / (\# \text{spam_predicho_spam} + \# \text{no-spam_predicho_spam})$

3. recall $(TP) / (TP+FN)$

En el ejemplo: $(\# \text{spam_predicho_spam}) / (\# \text{spam_predicho_spam} + \# \text{spam_predicho_no-spam})$

4. F1-score: $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

5. Fb-score: $[(1 + \beta^2) * (\text{precision} * \text{recall})] / [(b^2 * \text{precision}) + \text{recall}]$

Más métricas en [Confusion matrix](#)

sensitivity, recall, hit rate, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

precision or positive predictive value (PPV)

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN} = 1 - FOR$$

miss rate or false negative rate (FNR)

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$$

fall-out or false positive rate (FPR)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

false discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

false omission rate (FOR)

$$FOR = \frac{FN}{FN + TN} = 1 - NPV$$

Prevalence Threshold (PT)

$$PT = \frac{\sqrt{TPR(-TNR+1)} + TNR - 1}{(TPR + TNR - 1)}$$

Threat score (TS) or critical success index (CSI)

$$TS = \frac{TP}{TP + FN + FP}$$

accuracy (ACC)

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

balanced accuracy (BA)

$$BA = \frac{TPR + TNR}{2}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

Matthews correlation coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Fowlkes-Mallows index (FM)

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} = \sqrt{PPV \cdot TPR}$$

informedness or bookmaker informedness (BM)

$$BM = TPR + TNR - 1$$

markedness (MK) or deltaP

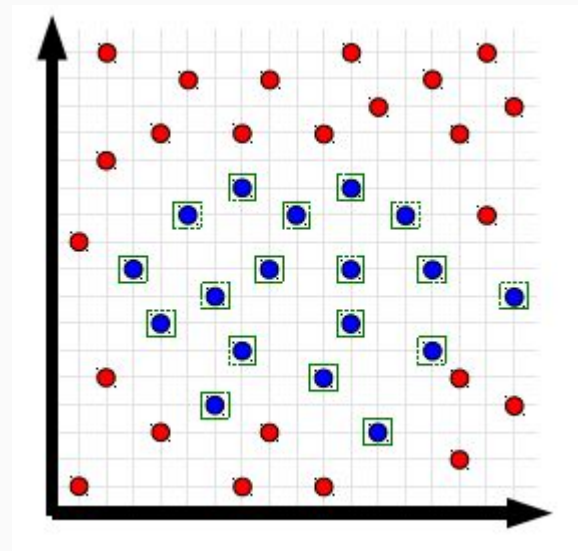
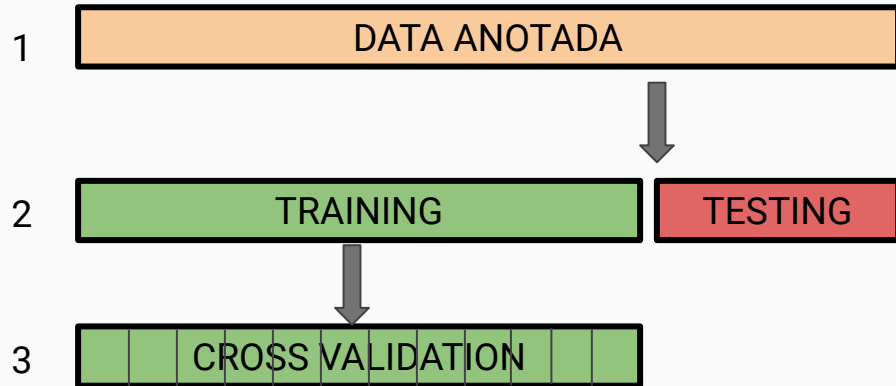
$$MK = PPV + NPV - 1$$

Métricas: Clase positiva SPAM

Clase real	N		S	
	N	S	N	S
	40	10	50	0
	0	50	10	40
accuracy (TP+TN) / TOTAL	0.9 (40+50 / 100)		0.9 (50+40 / 100)	
precision (TP) / (TP+FP)	0.8333 (50 / (50+10))		1 (40 / (40+0))	
recall (TP) / (TP+FN)	1 (50/(50+0))		0.8 (40/(40+10))	
F1 2 * (pre * rec) / (prec + rec)	0.9090		0.8888	

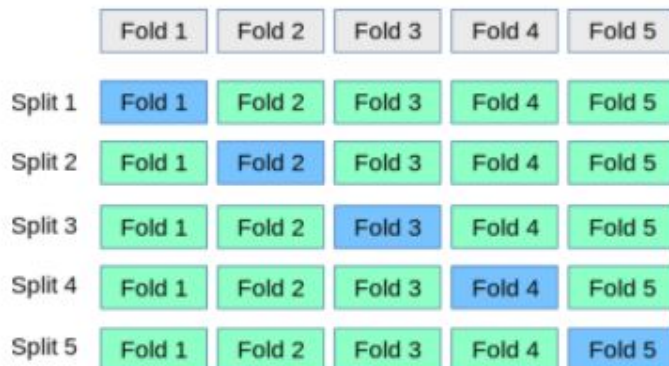
Overfitting

- Generalización
- ¿Entrenamos y testeamos en el mismo conjunto de datos?
- ¿Cómo mitigamos esto?
 - Cross validation, stratified cross validation, nested validation
 - Training - Validation - Test



3

CROSS VALIDATION



Distintos parámetros de los modelos:

- Tamaño del árbol
- Cuantos árbol
- Cuantas capas ocultas
- etc

Distintos parámetros en la etapa de feature extraction (si existe):

- Nuevos pre-procesamientos
- Nuevos filtros
- etc

3

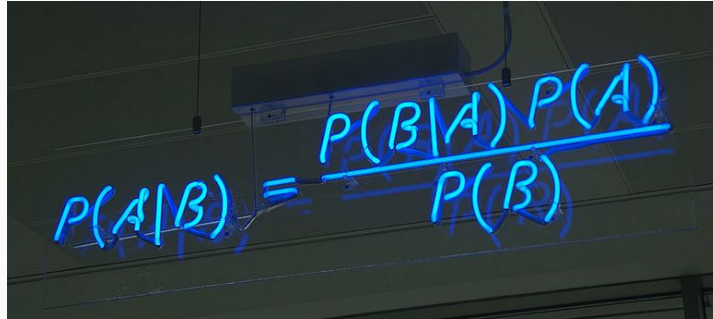
CROSS VALIDATION

Habiendo terminado la etapa de cross-validation: debemos tener **el mejor modelo encontrado** (es decir los mejores parámetros (o meta-parámetros), mejor pre-procesamiento, feature extraction, etc.

Entrenamos el modelo elegido anteriormente con todo el dataset de *Training* y luego predecimos en el set de *Test* y reportamos las diferentes métricas

No vale cambiar el modelo una vez que ya decidimos que era el mejor (para que cuando lo evaluamos en *test* estemos reproduciendo lo mejor posible el escenario de predecir sobre datos no vistos)

¿Cómo determinamos: #folds en el cross-validation, % de dataset para test, etc...? Con la experiencia, hay algunas buenas prácticas dependiendo de cada subdominio



A photograph of a whiteboard with the Naive Bayes formula written in blue marker. The formula is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The whiteboard is dark, and the marker is bright blue. The formula is written in a slightly slanted, handwritten style. The background of the slide is a solid blue color.

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

Naive Bayes

- 1) ¿Qué queremos? (donde V son el conjunto de clases posibles)

$$\operatorname{argmax}_{v_j \in V} P(v_j | a_1, a_2 \dots a_n)$$

- 2) Lo re-escribimos:

$$\operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)}$$
$$\operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j)$$

- 3) Asumimos independencia (por eso Naive!)

$$: \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

¿Cómo calculamos la probabilidad de que el atributo i -ésimo tenga un cierto valor dado que suponemos que proviene de la categoría v_j ?

$$P(a_i | v_j)$$

Caso discreto:

Si el atributo/feature i -ésimo es categórico/finito podemos medir la frecuencia.

Es decir, me fijo las instancias que tengan la clase v_j , me fijo cuantas en el i -ésimo atributo tienen el valor a_i y listo, con esto tengo la frecuencia.

Caso continuo:

Si el atributo/feature i -ésimo es numérico, podemos suponer alguna distribución (por ejemplo una normal) y fitearla para estimar la probabilidad para luego estimar la probabilidad de a_i

Esto se conoce como Gaussian Naive Bayes

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

Naive Bayes: Ejemplo

Dataset de entrenamiento:

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	Calor	Alta	Débil	No
Sol	Calor	Alta	Fuerte	No
Lluvia	Frío	Normal	Fuerte	No
Sol	Templado	Alta	Débil	No
Lluvia	Templado	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Nublado	Frío	Normal	Fuerte	Sí
Sol	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Sol	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Debil>

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	Calor	Alta	Débil	No
Sol	Calor	Alta	Fuerte	No
Lluvia	Frío	Normal	Fuerte	No
Sol	Templado	Alta	Débil	No
Lluvia	Templado	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Nublado	Frío	Normal	Fuerte	Sí
Sol	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Sol	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Debil>

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	Calor	Alta	Débil	No
Sol	Calor	Alta	Fuerte	No
Lluvia	Frío	Normal	Fuerte	No
Sol	Templado	Alta	Débil	No
Lluvia	Templado	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Nublado	Frío	Normal	Fuerte	Sí
Sol	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Sol	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Voy armando calculando cada probabilidad:

$P(\text{Cielo} = \text{Sol} \mid \text{No})$

Cielo	Tenis
Sol	No
Sol	No
Lluvia	No
Sol	No
Lluvia	No

$$\frac{3}{5} = 0.6$$

$P(\text{Cielo} = \text{Sol} \mid \text{Si})$

Cielo	Tenis
Nublado	Sí
Lluvia	Sí
Lluvia	Sí
Nublado	Sí
Sol	Sí
Lluvia	Sí
Sol	Sí
Nublado	Sí
Nublado	Sí

$$\frac{2}{8} = 0.25$$

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Debil>

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	Calor	Alta	Débil	No
Sol	Calor	Alta	Fuerte	No
Lluvia	Frío	Normal	Fuerte	No
Sol	Templado	Alta	Débil	No
Lluvia	Templado	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Nublado	Frío	Normal	Fuerte	Sí
Sol	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Sol	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Voy armando calculando cada probabilidad:

$P(\text{Temperatura} = \text{Calor} \mid \text{No})$

Temperatura	Tenis
Calor	No
Calor	No
Frío	No
Templado	No
Templado	No

$$2/5 = 0.4$$

$P(\text{Temperatura} = \text{Calor} \mid \text{Si})$

Temperatura	Tenis
Calor	Sí
Templado	Sí
Frío	Sí
Frío	Sí
Frío	Sí
Templado	Sí
Templado	Sí
Templado	Sí
Calor	Sí

$$2/8 = 0.25$$

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Debil>

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	Calor	Alta	Débil	No
Sol	Calor	Alta	Fuerte	No
Lluvia	Frío	Normal	Fuerte	No
Sol	Templado	Alta	Débil	No
Lluvia	Templado	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Nublado	Frío	Normal	Fuerte	Sí
Sol	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Sol	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Voy armando calculando cada probabilidad:

$P(\text{Humedad} = \text{Alta} \mid \text{No})$

Humedad	Tenis
Alta	No
Alta	No
Normal	No
Alta	No
Alta	No

$$4/5 = 0.8$$

$P(\text{Humedad} = \text{Alta} \mid \text{Sí})$

Humedad	Tenis
Alta	Sí
Alta	Sí
Normal	Sí
Normal	Sí
Normal	Sí
Normal	Sí
Normal	Sí
Alta	Sí
Normal	Sí

$$3/8 = 0.375$$

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Débil>

Cielo	Temperatura	Humedad	Viento	Tenis
Sol	Calor	Alta	Débil	No
Sol	Calor	Alta	Fuerte	No
Lluvia	Frío	Normal	Fuerte	No
Sol	Templado	Alta	Débil	No
Lluvia	Templado	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Nublado	Frío	Normal	Fuerte	Sí
Sol	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Sol	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Voy armando calculando cada probabilidad:

$P(\text{Viento} = \text{Débil} \mid \text{No})$ $P(\text{Viento} = \text{Débil} \mid \text{Sí})$

Humedad	Tenis
Débil	No
Fuerte	No
Fuerte	No
Débil	No
Fuerte	No

$$2/5 = 0.4$$

Humedad	Tenis
Débil	Sí
Débil	Sí
Débil	Sí
Fuerte	Sí
Débil	Sí
Débil	Sí
Fuerte	Sí
Fuerte	Sí
Débil	Sí

$$6/8 = 0.75$$

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Debil>

Caso v=No

$$P(\text{Cielo} = \text{Sol} \mid \text{No}) * P(\text{Temperatura} = \text{Calor} \mid \text{No}) * P(\text{Humedad} = \text{Alta} \mid \text{No}) * P(\text{Viento} = \text{Débil} \mid \text{No}) = \\ 0.6 * 0.4 * 0.8 * 0.4 = \quad \quad \quad \mathbf{0.0768 = P(\text{features} \mid \text{No})}$$

Caso v=Sí

$$P(\text{Cielo} = \text{Sol} \mid \text{Sí}) * P(\text{Temperatura} = \text{Calor} \mid \text{Sí}) * P(\text{Humedad} = \text{Alta} \mid \text{Sí}) * P(\text{Viento} = \text{Débil} \mid \text{Sí}) = \\ 0.25 * 0.25 * 0.375 * 0.75 = \quad \quad \quad \mathbf{0.017578125 = P(\text{features} \mid \text{Sí})}$$

$$\underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i \mid v_j)$$

Prior:

$$\begin{array}{lll} P(\text{No}) & = 6/14 & = 0.4285 \\ P(\text{Sí}) & = 9/14 & = 0.6428 \end{array}$$

Naive Bayes: Ejemplo

Quiero clasificar: < Cielo = Sol, Temperatura = Calor, Humedad=Alta, Viento= Debil>

Caso v=No

$$P(\text{No}) * P(\text{features} | \text{No}) = \mathbf{0.03290}$$

Caso v=Sí

$$P(\text{Sí}) * P(\text{features} | \text{Sí}) = \mathbf{0.0112}$$

$$\underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

¿Entonces cuál es el máximo v ?

v=No

A programar!

1. Metricas
2. Cross validation
3. Naive Bayes

Link Colab:

https://drive.google.com/file/d/17SMIKhFc_Jkgh5Chlf81UK1AG7iVxcC1/view?usp=sharing

Reglas:

1. Individual
2. Pueden charlar, discutir y aprender entre ustedes
3. No tienen que entregar nada
4. No hagan copy paste de internet porque no tiene mucho sentido (son todas funciones ya super programadas, las hacemos para aprender!)