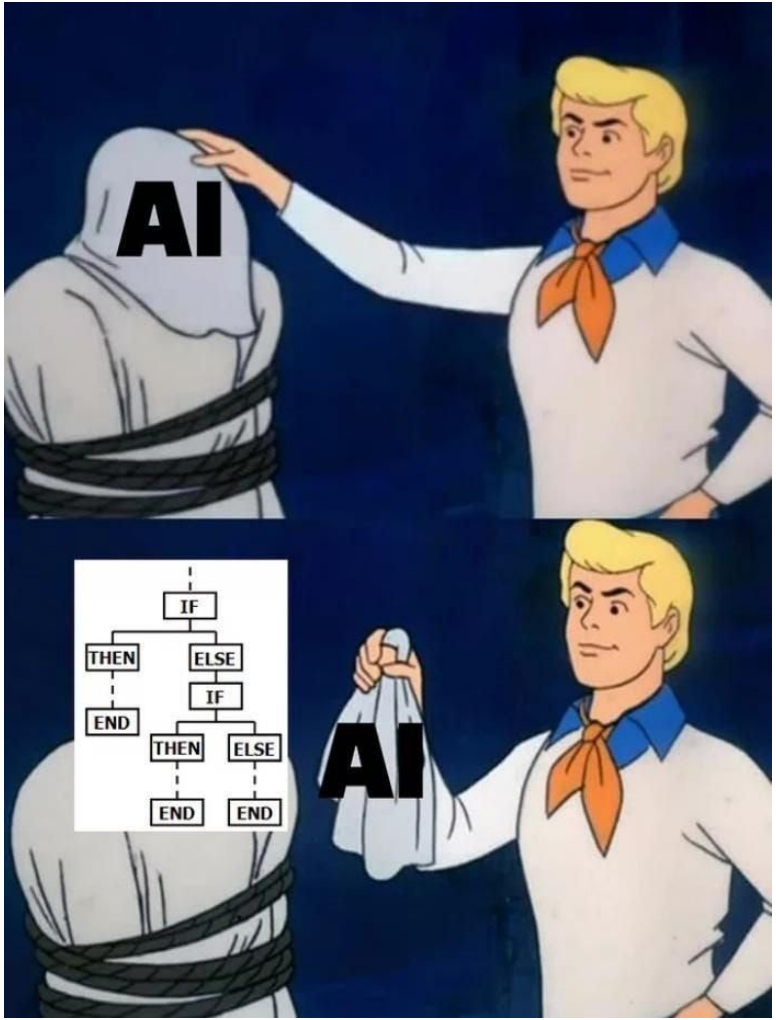
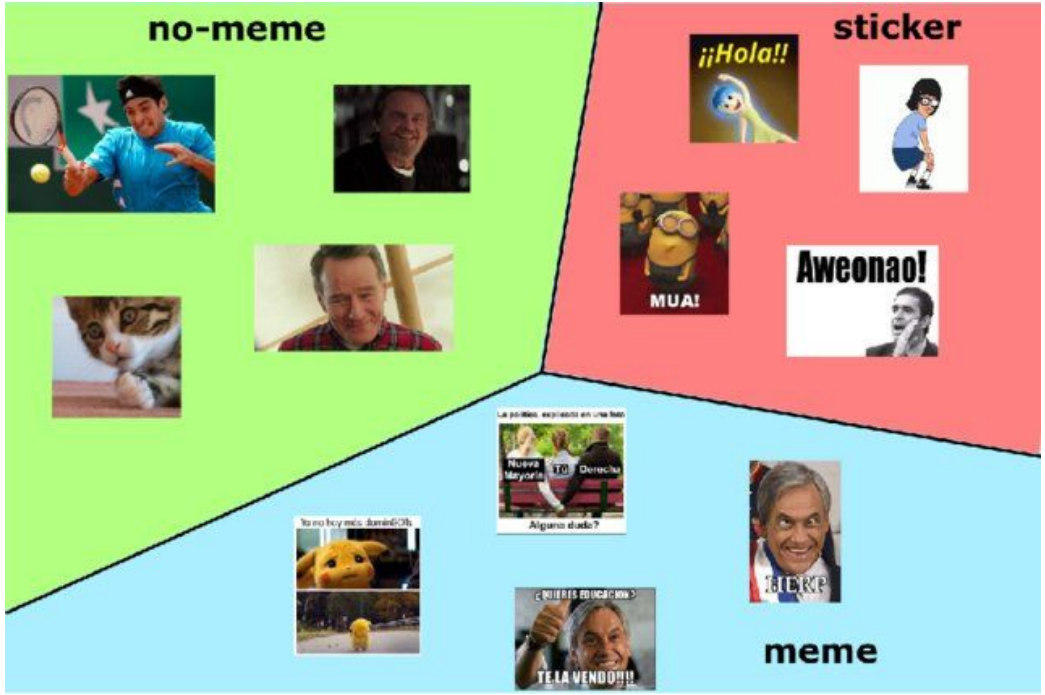
Introducción a la Ciencia de Datos

Maestría en Ciencias de la Computación

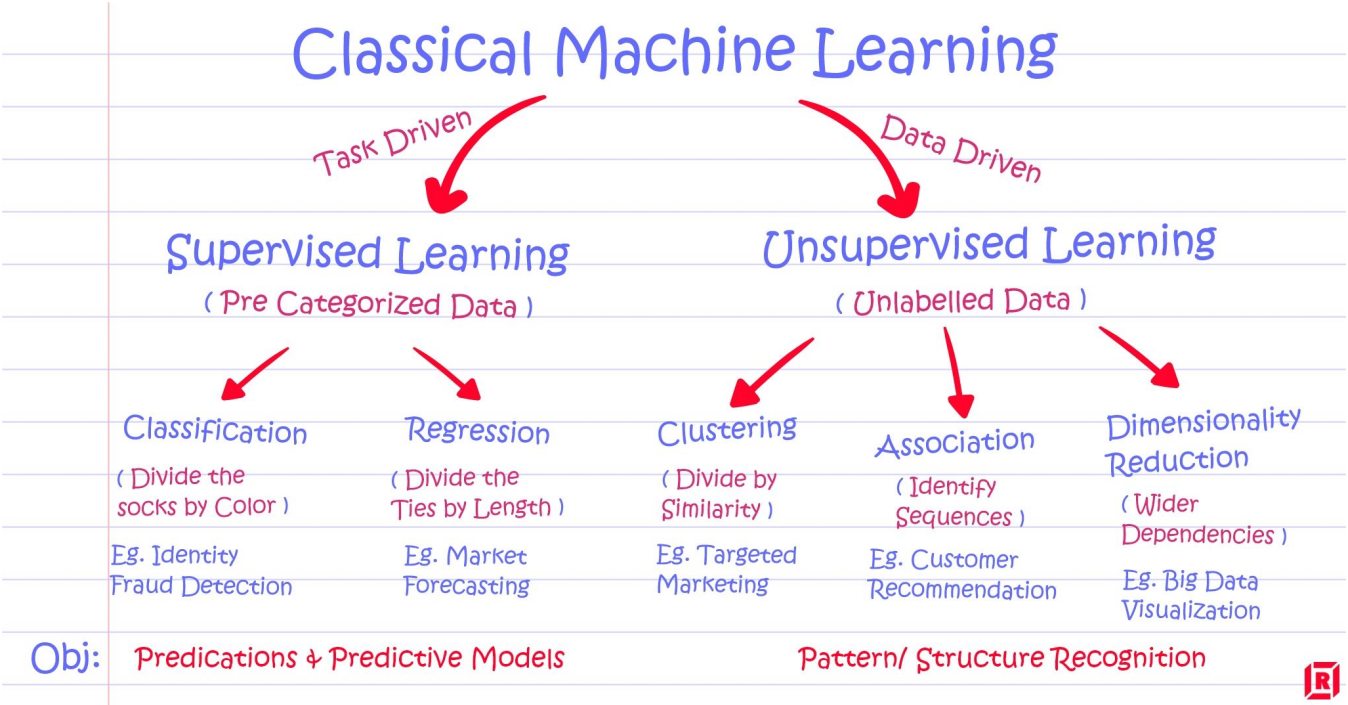
Dr. Irvin Hussein López Nava


AI
IF
THEN
ELSE
IF
END
THEN
ELSE
END
END
AI


no-meme
sticker
¡¡Hola!!
MUA!
Aweonao!
Nueva Mayoría
Derecha
Alguna duda?
¿QUIERES EDUCACION?
TE LA VENDO!!!!
meme

04

Clasificación


Classical Machine Learning
Task Driven
Data Driven
Supervised Learning
(Pre Categorized Data)
Unsupervised Learning
(Unlabelled Data)
Classification
(Divide the socks by Color)
Eg. Identity Fraud Detection
Regression
(Divide the Ties by Length)
Eg. Market Forecasting
Clustering
(Divide by Similarity)
Eg. Targeted Marketing
Association
(Identify Sequences)
Eg. Customer Recommendation
Dimensionality Reduction
(Wider Dependencies)
Eg. Big Data Visualization
Obj:
Predications & Predictive Models
Pattern/ Structure Recognition

4.1 Modelos de clasificación


¿Qué es
Clasificación?

Definiciones

ChatGPT

Es una tarea que implica asignar una etiqueta predefinida a un elemento de entrada en función de sus características. El objetivo es aprender a generalizar a partir de datos de entrenamiento para hacer predicciones precisas sobre nuevos datos no etiquetados.

https://chat.openai.com/

Wikipedia

Es el problema de identificar a cuál de un conjunto de categorías pertenece una observación; e.g., asignar un correo electrónico a la clase "spam" o "no spam", o asignar un diagnóstico a un paciente determinado según sus características observadas.

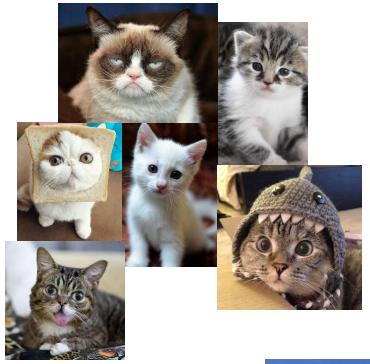
https://en.wikipedia.org/wiki/Statistical_classification

Gemini

Es un tipo de aprendizaje supervisado que se utiliza para asignar etiquetas a los datos. El objetivo es crear un modelo que pueda predecir la categoría de un nuevo dato, dado un conjunto de datos de entrenamiento con etiquetas conocidas.

https://gemini.google.com/

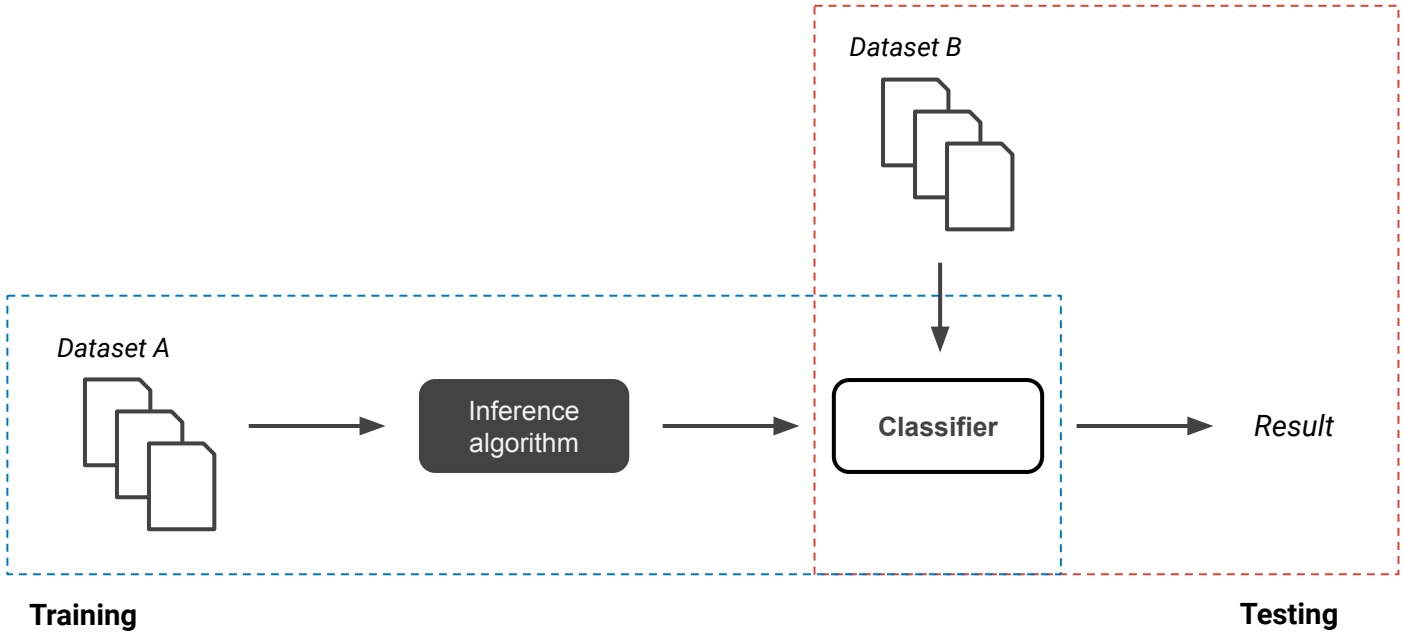
Un primer caso



Classification model

cat

Clasificador general



¿Qué es clasificación?

- En el **aprendizaje supervisado**, la **clasificación** es el proceso de **identificar** a cuál de un conjunto de clases pertenece una nueva **observación** (fase de prueba), a partir de un conjunto de **datos de entrenamiento** que contiene **observaciones** cuya clase se conoce (fase de entrenamiento).

Fase de entrenamiento

- El entrenamiento se lleva a cabo utilizando datos de entrenamiento $\mathcal{T} = \{(\mathcal{X}_i, y_i)_{i=1}^n\}$ con n pares de vectores de características $\mathcal{X}_i$ y sus correspondientes etiquetas y_i
- A partir de los datos de entrenamiento $\mathcal{T}$ se construye un modelo utilizando métodos de inferencia supervisada, $\mathcal{I}$, antes de utilizarlo en la **fase de prueba**.
- Los parámetros del modelo λ deben aprenderse para minimizar el error de clasificación en $\mathcal{T}$ si se ha utilizado un **algoritmo paramétrico** para construir modelos.
- En cambio, los **algoritmos no paramétricos** toman como parámetro los datos de entrenamiento etiquetados $\lambda = \mathcal{T}$ sin más entrenamiento.

Fase de prueba

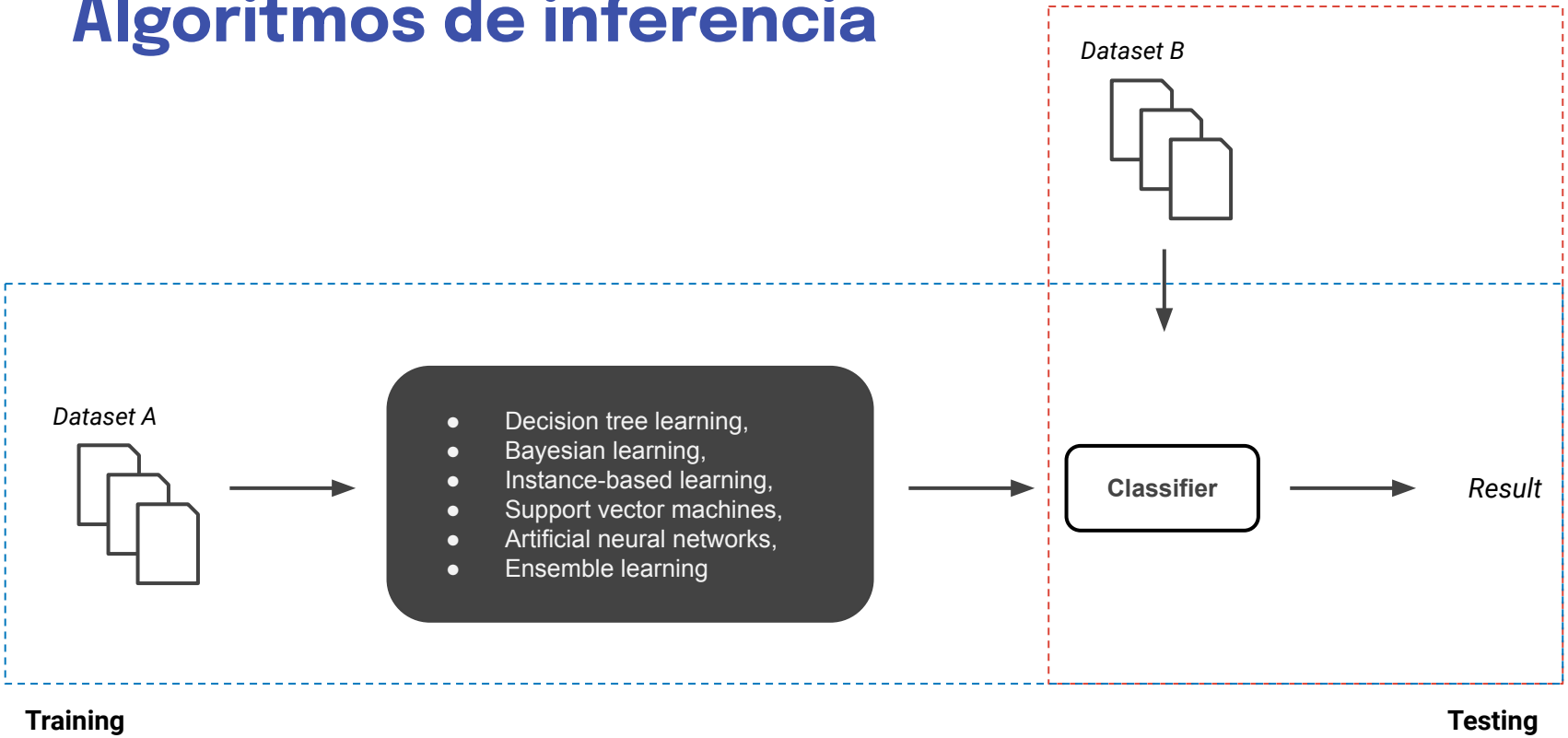
- Las pruebas se realizan utilizando un modelo entrenado con parámetros λ, que asigna cada nuevo vector de características $\mathcal{X}_i$ a un conjunto de etiquetas de clase $\mathcal{Y} = y^1, ..., y^c$ con sus correspondientes puntuaciones $\mathcal{P}_i = p_i^1, ..., p_i^c$.

$$p_i(y|\mathcal{X}_i, \lambda) = \mathcal{I}(\mathcal{X}_i, \lambda), \text{ for each } y \in \mathcal{Y}$$

- con el método de inferencia $\mathcal{I}$. A continuación, las puntuaciones $\mathcal{P}_i$ calculadas se utilizan para obtener la puntuación máxima y seleccionar la etiqueta de clase y_i correspondiente como salida de la clasificación.

$$y_i = \underset{y \in \mathcal{Y}, p \in \mathcal{P}_i}{\operatorname{argmax}} p_i(y|\mathcal{X}_i, \lambda)$$

Algoritmos de inferencia

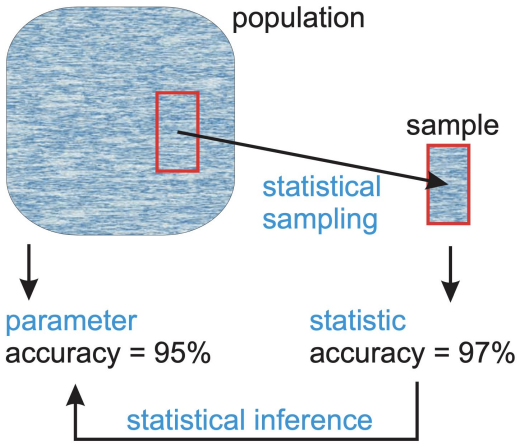


¿Hasta qué punto confiamos en el clasificador aprendido?

- Los clasificadores (tanto supervisados como no supervisados) se aprenden (entrenan) en un conjunto de **datos de entrenamiento finito**.
- Un clasificador, o modelo, aprendido debe probarse experimentalmente en una prueba diferente (datos de prueba).
- El rendimiento experimental sobre los **datos de prueba** es una aproximación al rendimiento en datos desconocidos – comprueba la capacidad de generalización del clasificador.
- Se necesita una función que evalúe experimentalmente el rendimiento del clasificador, e.g., su tasa de error, precisión, sensibilidad, especificidad.
    - Es necesario comparar los clasificadores experimentalmente.

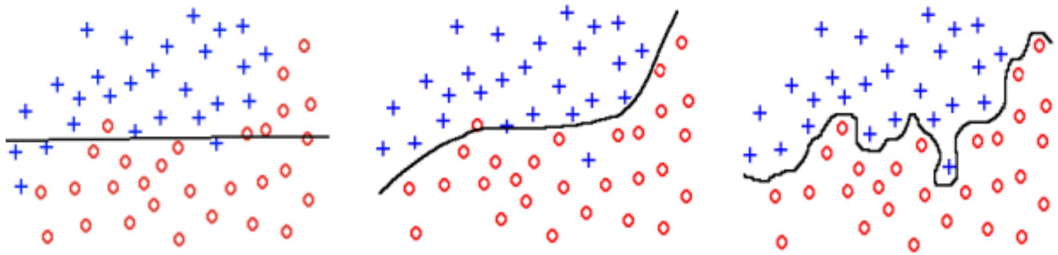
La evaluación es una prueba de hipótesis

- La evaluación debe tratarse como una **prueba de hipótesis** en estadística.
- El valor del parámetro poblacional debe deducirse estadísticamente a partir de las estadísticas de la muestra (i.e., un conjunto de entrenamiento).



Riesgo de sobreajuste

- **Aprender** de los datos de entrenamiento con demasiada precisión suele dar lugar a malos resultados de clasificación en nuevos datos.
- ¡El clasificador debe tener la capacidad de **generalizar**!



underfit fit overfit

Uso de los datos: conjuntos de entrenamiento y de prueba

Problema: sólo se dispone de datos finitos y deben utilizarse tanto para el entrenamiento como para las pruebas.

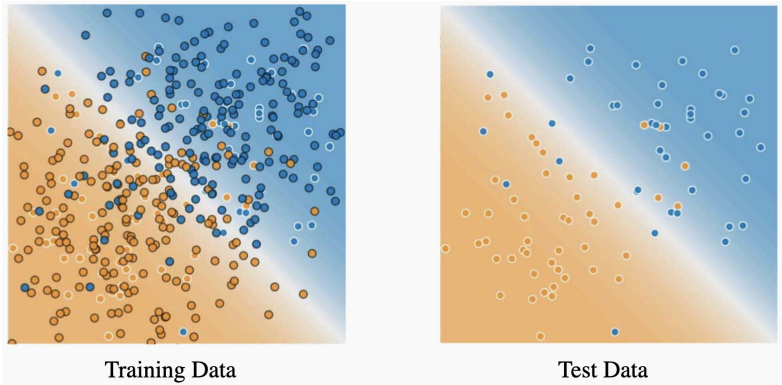
- Un mayor número de datos de entrenamiento mejora la generalización.
- Más datos de prueba dan una mejor estimación de la probabilidad de error de clasificación.
- **NUNCA** se debe evaluar el rendimiento de los clasificadores en función de los **datos de entrenamiento**: la conclusión tendría un sesgo optimista.

División de los datos

Partición (división) del conjunto finito de datos disponible en subconjuntos de entrenamiento/prueba:

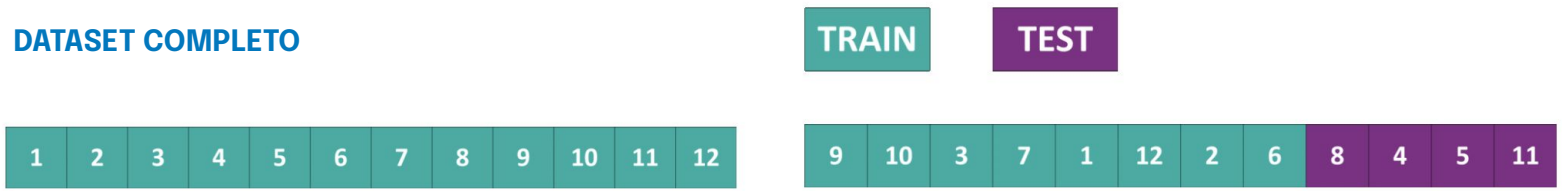
- Hold out.
- Cross validation.
- Bootstrap.

Una vez finalizada la evaluación, todos los datos disponibles pueden utilizarse para entrenar el clasificador final.



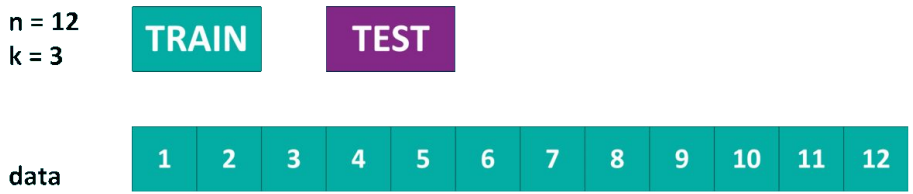
Hold out method

- Los datos se dividen **aleatoriamente** en dos conjuntos independientes.
- El **conjunto de entrenamiento** (e.g., 2/3 de los datos) para la construcción del modelo estadístico, i.e., el aprendizaje del clasificador.
  - Generalmente la proporción de los datos es mayor al conjunto de prueba.
- El **conjunto de prueba** (e.g., 1/3 de los datos) se utiliza para estimar la precisión del clasificador.



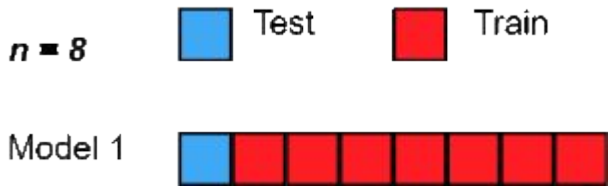
K-fold cross validation

- El **conjunto de entrenamiento** se divide **aleatoriamente** en *k* conjuntos disjuntos de igual tamaño en los que cada parte tiene aproximadamente la misma distribución de clases.
- El clasificador se evalúa *k* veces, cada vez con un conjunto diferente que se utiliza como **conjunto de prueba**.
- El rendimiento del clasificador es la media de estos *k* conjuntos.



Leave-one-out

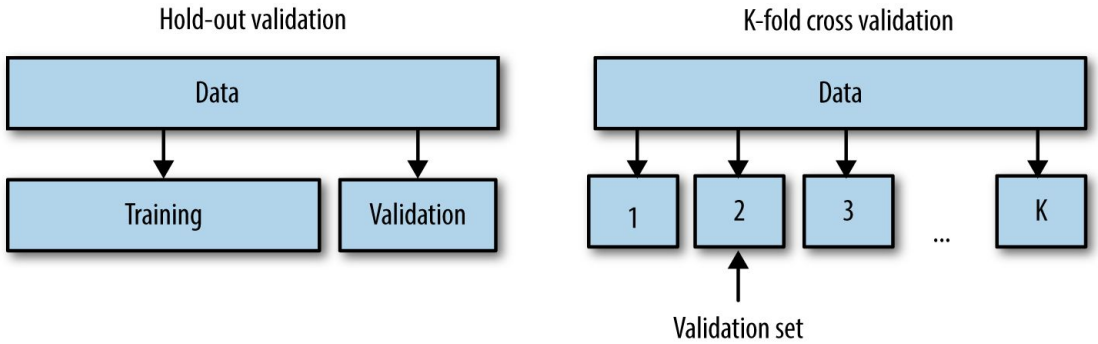
- Un caso especial de validación cruzada *k*–fold con *k* = *n*, donde *n* es el número total de muestras del conjunto de datos.
  - Se realizan *n* experimentos utilizando n – 1 muestras para el entrenamiento y la muestra restante para las pruebas.
- Es bastante costosa desde el punto de vista computacional.



Bootstrap method

- Utiliza el muestreo con sustitución para formar el **conjunto de entrenamiento**.
  Dado: el **conjunto de entrenamiento** T que consta de n entradas.
- Bootstrap genera m nuevos conjuntos T_i cada uno de tamaño $n' < n$ muestreando T uniformemente con reemplazo.
  - Como consecuencia, algunas entradas pueden repetirse en T_i.
  - En un caso especial (632 boosting) cuando $n' = n$, para n grandes, se espera que T_i tenga $1 - 1 \approx 63{,}2\%$ de muestras únicas. El resto son duplicados.
- Los m modelos estadísticos (e.g., clasificadores, regresores) se aprenden utilizando las m muestras bootstrap anteriores.
  - Por último, los modelos se combinan, e.g., promediando los resultados (regresión) o votando (clasificación).

Resumen de los métodos



Métricas

Retomando el *accuracy*

- La **exactitud** es el porcentaje de clasificaciones correctas.
- La **tasa de error** es el porcentaje de clasificaciones incorrectas.
- *Accuracy = 1 – Error rate*.

Problemas con estas métricas:

- Asume costes iguales para la clasificación errónea.
- Supone una distribución de clases relativamente uniforme (e.g., 0.5% de pacientes con una determinada enfermedad).

Se pueden derivar otras métricas de la matriz de confusión.

Matriz de confusión (caso binario)

- Una matriz de confusión, o matriz de error, es un diseño de tabla específico que permite visualizar el rendimiento de un modelo de clasificación.

		Predicted		
		Positive	Negative	
Actual	Positive	True positive (TP)	False negative (FN)	Sensitivity
	Negative	False positive (FP)	True negative (FN)	Specificity
		Precision		F-measure

Matriz de confusión (tipos de errores)

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN (error type I)
	Negative	FP (error type II)	TN

- El primer tipo de error es **el rechazo** de una hipótesis nula verdadera. El segundo es **el NO rechazo** de una hipótesis nula falsa.
- Como ejemplo, un error de tipo I corresponde a condenar a un acusado inocente; mientras que un tipo II corresponde a absolver a un criminal.

- Accuracy (ACC):

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

- Sensitivity, recall, o true positive rate (TPR):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- Specificity o true negative rate (TNR):

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

- Precision o positive predictive value (PPV):

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- F1 score es la media armónica entre precision and sensitivity:

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Costes desiguales de las decisiones

- **Diagnóstico médico:** El coste de una indicación falsa de cáncer de mama en el cribado de la población es menor que el coste de pasar por alto una enfermedad verdadera.
- **Defensa contra misiles aéreos:** El coste de no detectar un ataque real es mucho mayor que el de una falsa alarma.

Problema de la distribución de clases desconocida

En muchas circunstancias, se desconoce la distribución de clases, e.g., un filtro de spam de correo electrónico. Los modelos estadísticos deben aprenderse de antemano.

- Diagnóstico médico: 95 % sano, 5 % enfermedad.
- Comercio electrónico: el 99 % no compra, el 1 % compra.
- Seguridad: el 99.999 % de los ciudadanos no son terroristas.

Una situación similar ocurre con los clasificadores multiclase. La clase mayoritaria puede acertar el 99 %, pero es inútil.

Dealing with the imbalance: sampling

Construir un conjunto de entrenamiento equilibrado para entrenar el clasificador.

- Seleccione aleatoriamente el número deseado de casos de clase minoritaria.
- Añada el mismo número de casos de clase mayoritaria seleccionados aleatoriamente.

Construya un conjunto de prueba equilibrado (diferente del conjunto de entrenamiento) para probar el clasificador.

Dealing with the imbalance: augmentation

Construir un conjunto de entrenamiento equilibrado para entrenar el clasificador.

- Seleccione el número deseado de instancias de clase mayoritaria.
- Añada el mismo número de casos sintéticos de clases minoritarias.

Construya un conjunto de prueba equilibrado (**sin incluir datos sintéticos**) para probar el clasificador.

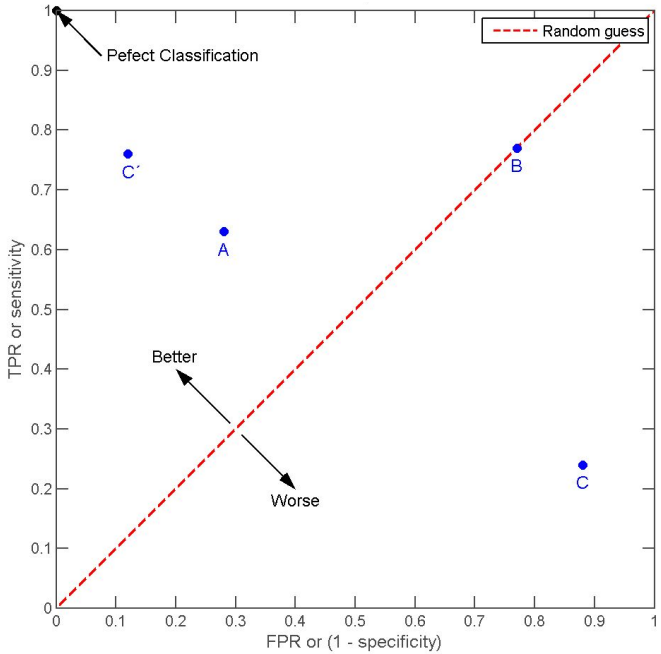
¿Es suficiente una métrica para evaluar el rendimiento de un clasificador?

- Las características escalares, como la exactitud, no proporcionan suficiente información.
- Dos números –la tasa de verdaderos positivos y la tasa de falsos positivos– son mucho más informativos que un solo número.
- Estos dos números se visualizan mejor mediante una curva, e.g., mediante una *Receiver Operating Characteristic* (ROC), que informa sobre:
  - Rendimiento para todos los posibles costes de clasificación errónea.
  - Rendimiento para todas las relaciones de clase posibles.
- ¿En qué condiciones el clasificador c1 supera al clasificador c2?

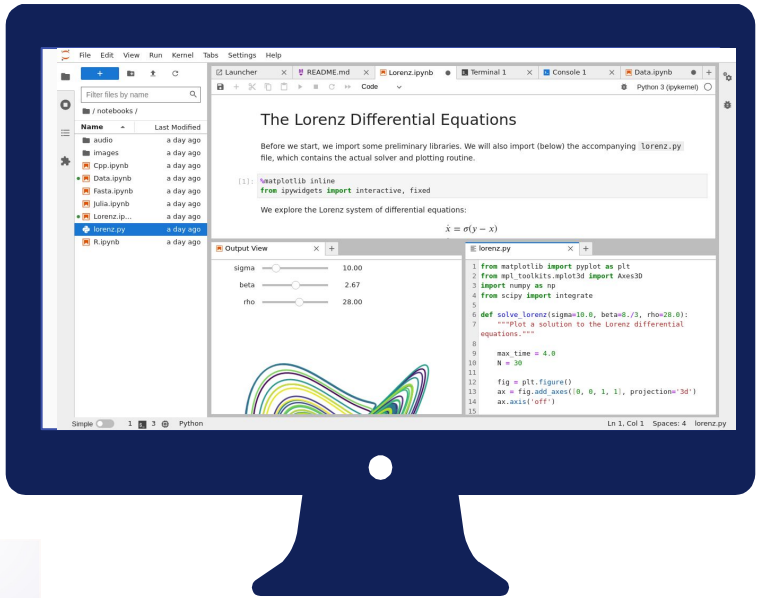
ROC – Receiver Operating Characteristic

- Gráfico que muestra el TPR (aciertos) frente al FPR (falsas alarmas).
- Caracteriza el grado de solapamiento de las clases para una única característica.
- La decisión se basa en un único umbral Θ (llamado también punto operativo).
- Diferentes curvas ROC corresponden a diferentes clasificadores.
- La curva única es el resultado de cambiar el umbral Θ.

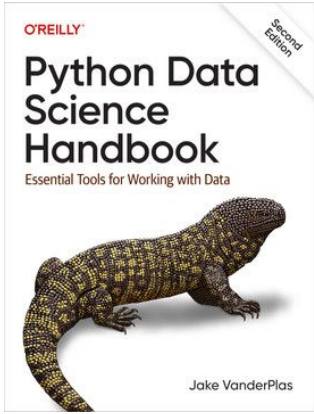
ROC – Receiver Operating Characteristic



(Go to live notebook)



Extra Libro



GitHub

- 05.00-Machine-Learning.ipynb
- 05.01-What-Is-Machine-Learning.ipynb
- 05.02-Introducing-Scikit-Learn.ipynb
- **05.03-Hyperparameters-and-Model-Validation.ipynb**
- 05.04-Feature-Engineering.ipynb
- 05.06-Linear-Regression.ipynb

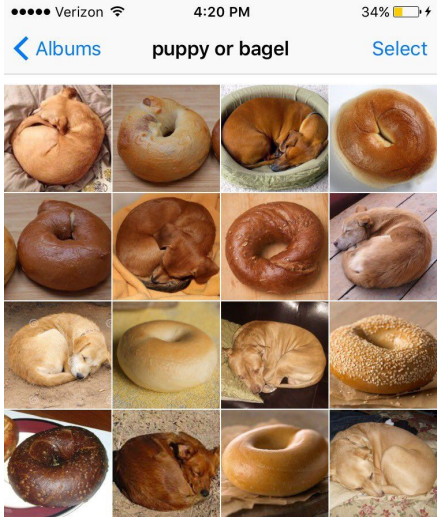
¿El reto?

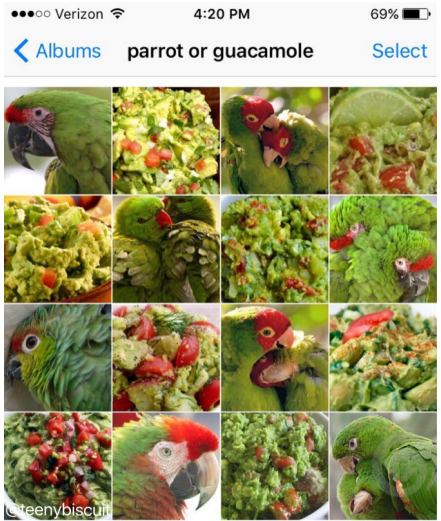

Seguir
karen zack
@teenybiscuit
nice lady | creative at @wiedenkennedy
Traducir la biografía
los angeles
karenzack.com
Se unió en junio de 2009
845 Siguiendo
6.171 Seguidores
Ninguna de las cuentas que sigues sigue a este usuario


Verizon
4:20 PM
76%
Albums
chihuahua or muffin
Select
@teenybiscuit


Verizon 10:50 PM 4%
Back labradoodle or fried chicken Select


Verizon LTE 3:02 PM 42%
Albums sheepdog or mop Select

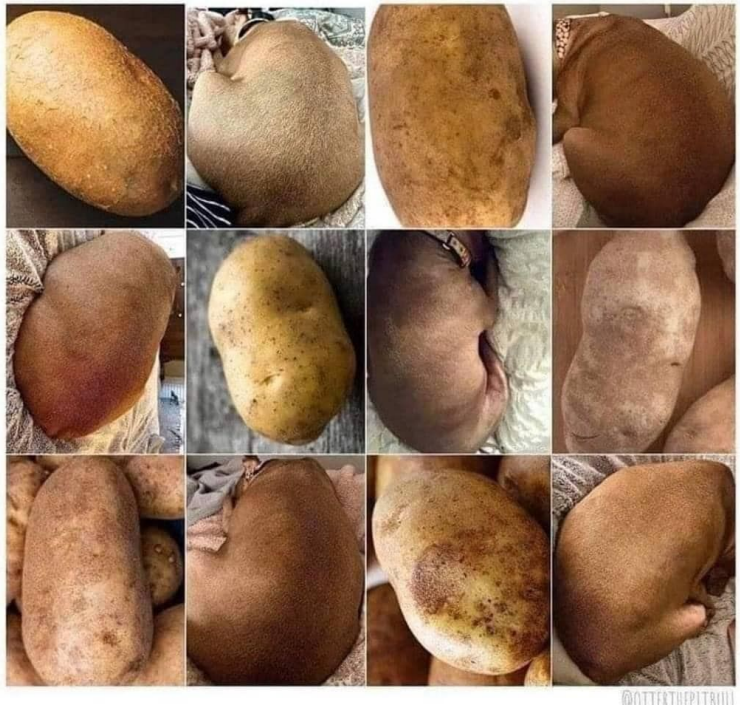

Verizon 4:20 PM 34%
Albums puppy or bagel Select


Verizon 4:20 PM 69%
Albums parrot or guacamole Select
@teenybiscuit


Verizon 4:20 PM 69%
Albums shiba or marshmallow Select
@teenybiscuit


Verizon 4:20 PM 69%
Albums kitten or ice cream Select
@teenybiscuit

POTATO OR PIT BULL



Gracias!

¿Alguna pregunta?

hussein@cicese.mx
https://sites.google.com/view/husseinlopeznava

SCAN ME

CREDITS: This presentation was based on a template by **Slidesgo**, and includes icons by **Flaticon**.