

Tópicos de inteligencia artificial: de métodos clásicos a modelos generativos

Gabriela Ramírez de la Rosa

<https://gabyrr.github.io/>

TecNM - Oaxaca
12 al 16 de agosto de 2024

Brevemente, sobre mi

- Ingeniera en Computación, UTM
- Maestra en Ciencias Computacionales, INAOE
- Doctora en Ciencias Naturales e Ingeniería, UAM

Intereses

Principal área de investigación es la intersección del Procesamiento del Lenguaje Natural y Psicología

Investigar, proponer y desarrollar sistemas de IA para ayudar en el diagnóstico de enfermedades y condiciones mentales a partir del análisis automático del lenguaje

Trabajo principalmente en el desarrollo de nuevas representaciones interpretables basada en teorías cognitivas

Temario general

1. Aprendizaje supervisado, conceptos básicos
2. Métodos clásicos de aprendizaje supervisado
3. Métodos generativos: fundamentos y ejemplos

Hoy

De 9 a 12 hrs

Break (por reunión) de 12 a 13:30

De 13:30 a 15:00

- Inteligencia artificial y Aprendizaje automático
 - Conceptos básicos
 - Datos (práctica)
- Métodos clásicos de aprendizaje
 - KNN, NB, GNB, Lineales
- Break -- reunión (12:00-13:30)
- Ejercicio (13:30-15:00)

Inteligencia artificial

Campo de investigación en ciencias computacionales que estudia, diseña e implementa modelos para que las computadoras “se comporten” o “piensen” como humanos

	Como humanos	Racionalidad
PENSAR	Pensar como humanos Modelos cognitivos	Pensar racionalmente Leyes del pensamiento
ACTUAR	Actuar como humanos Test de turing	Actuar racionalmente Agente racional
<i>Objetivo</i>	<i>Ser fiel al desempeño que tendría un humano. “Imitar”</i>	<i>Ser fiel a un desempeño ideal. Hace lo correcto de acuerdo a lo que conoce</i>

Inteligencia artificial

	Como humanos	Racionalidad
PENSAR	Ciencias cognitivas Multi-disciplinas: psicología, neurología, etc.	Lógica
ACTUAR	Procesamiento del Lenguaje Natural Representación del conocimiento Aprendizaje automático (ML) Visión Robótica	Agente racional
Objetivo	<i>Ser fiel al desempeño que tendría un humano. “Imitar”</i>	<i>Ser fiel a un desempeño ideal. Hace lo correcto de acuerdo a lo que conoce</i>

¿Por qué enfocarnos en ML?

- Motores de búsqueda (e.g. Google)
- Sistemas de recomendación (e.g. Netflix)
- Traducción automática (e.g. Google Translate)
- Comprensión del habla (e.g. Siri, Alexa)
- Jugadores autónomos (e.g. AlphaGo)
- Autos autónomos
- Medicina personalizada
- Uso en otras ciencias: Genética, astronomía, química, neurología, física, etc...

Machine Learning (Aprendizaje Automático)

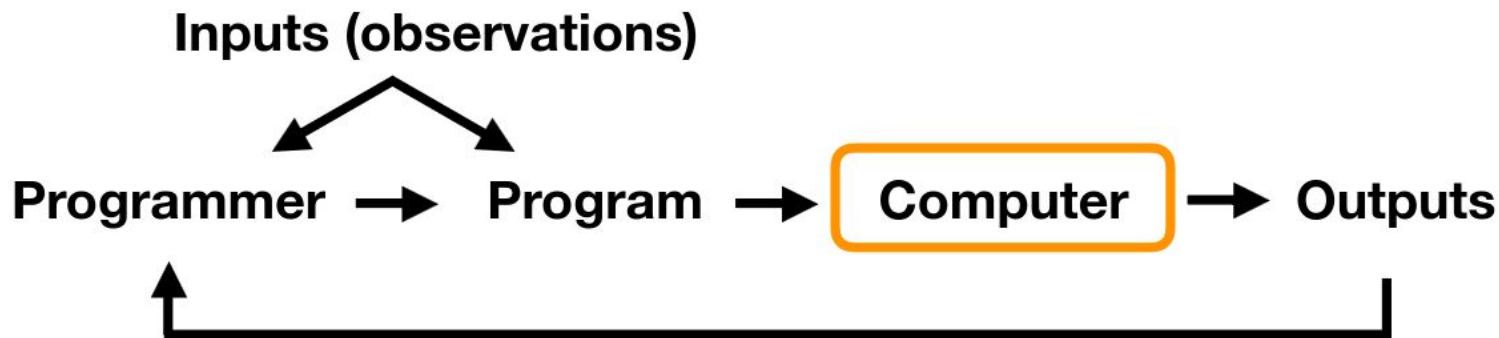
Subcampo de la Inteligencia Artificial (IA) que se enfoca en el desarrollo de programas de computadora que:

- Adapten su conocimiento automáticamente a partir de datos
- Sean capaces de aprender sin ser programados explícitamente



<https://xkcd.com/1838/>

El método de programación tradicional

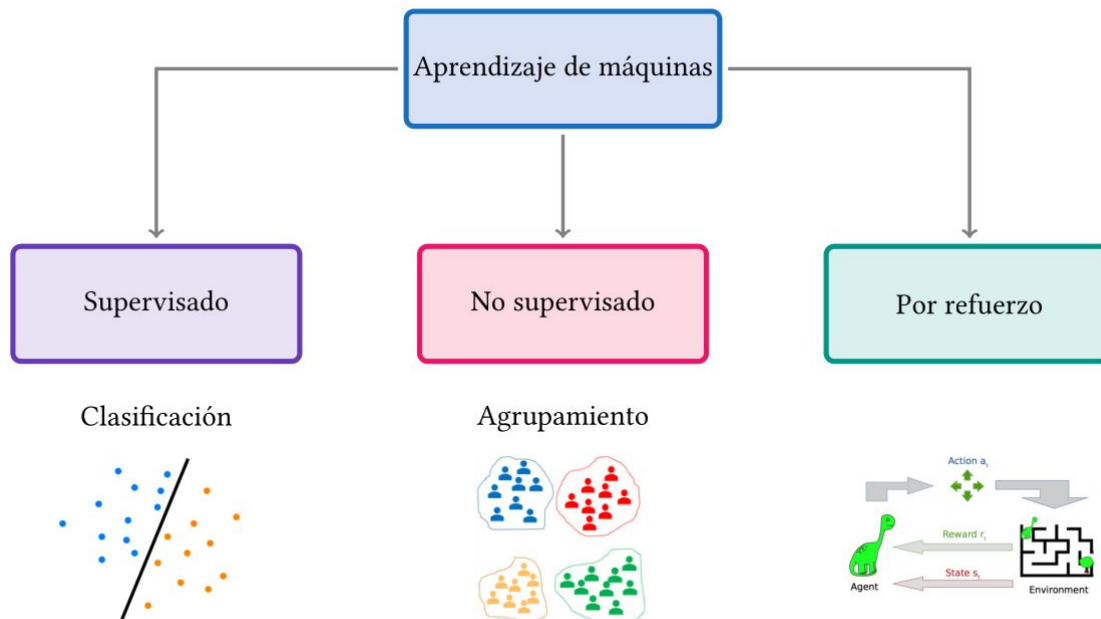


El método de programación tradicional

Sistemas que sean capaces de aprender sin ser programados explícitamente...



Diferentes tipos de aprendizaje automático



Diferentes tipos de aprendizaje automático

Supervised Learning

- > Labeled data
- > Direct feedback
- > Predict outcome/future

Unsupervised Learning

- > No labels
- > No feedback
- > Find hidden structure in data

Reinforcement Learning

- > Decision process
- > Reward system
- > Learn series of actions

(Raschka et al., chap.1)

Aprendizaje supervisado


Es una función h que mapea datos de entrada \mathcal{X} a un etiqueta (clase) y de un conjunto fijo de posibles clases.

$$\mathcal{X} \Rightarrow y$$

Donde \mathcal{X} es un conjunto de instancias x expresadas en forma de vector

La función aprendida puede responder consultas del tipo:

$$h(x) = y$$

$h(\text{) = \text{gato}$

Aprendizaje supervisado - definición

El aprendizaje supervisado se define como (Russell and Norvig, 2009):

Dado un conjunto de entrenamiento de N pares de ejemplos, con la forma (x_1, y_1) , (x_2, y_2) , \dots , (x_N, y_N) , donde cada y_j es generada por una función desconocida $f(x_j) = y_j$, se construye una función h que aproxime la función real f .

La función h se evalúa con un conjunto diferente de pares entrada–salida; si la función h predice de forma correcta el valor de y para los nuevos ejemplos, entonces h generaliza bien a f .

Aprendizaje supervisado - esquema general

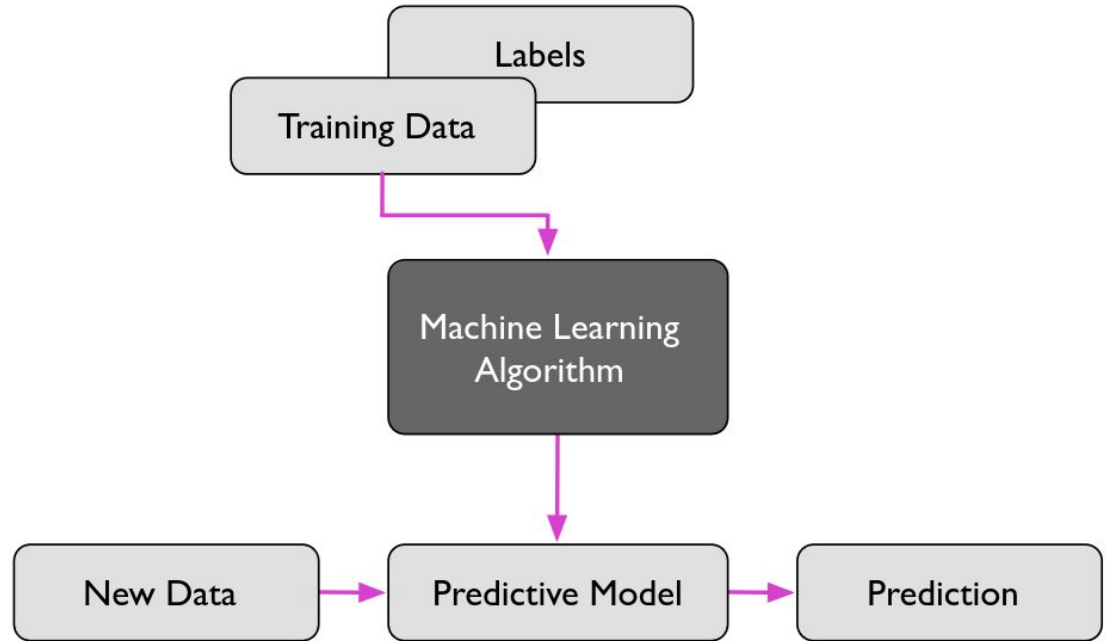
Aprende un modelo a partir de datos etiquetados, para hacer predicciones

- Se conoce a priori la salida correcta (etiqueta)

$$\mathcal{X} \Rightarrow y$$

Hay dos tipos dependiendo de la forma de y

- **Clasificación** - predice una clase de un conjunto de clases predefinidas
- **Regresión** - predice un valor



(Raschka et al., chap.1)

Clasificación - ejemplo

Clasificar tipos de Iris



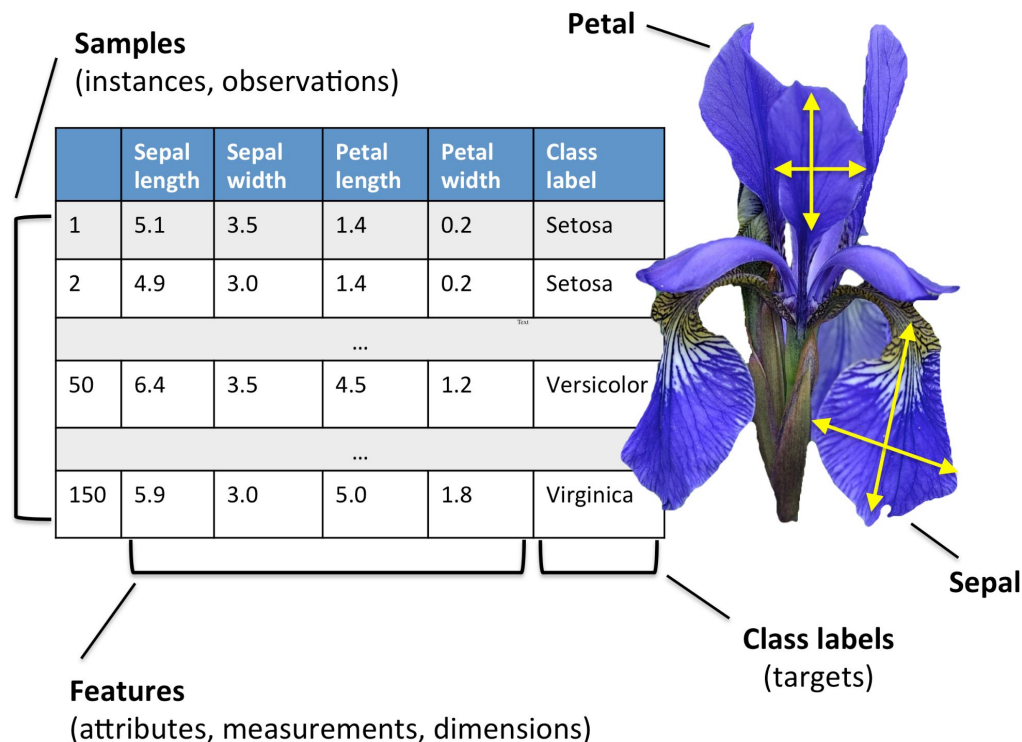
¿Cómo lo harían?

Clasificación - ejemplo

Representación de los ejemplos




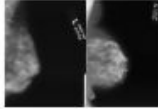



- tomar una foto usar los valores de los píxeles como entrada (--> deep learning)
- definir un número de características de entrada (variables)

Los ejemplos son puntos en un espacio de alta dimensionalidad



(Raschka et al., chap.1)

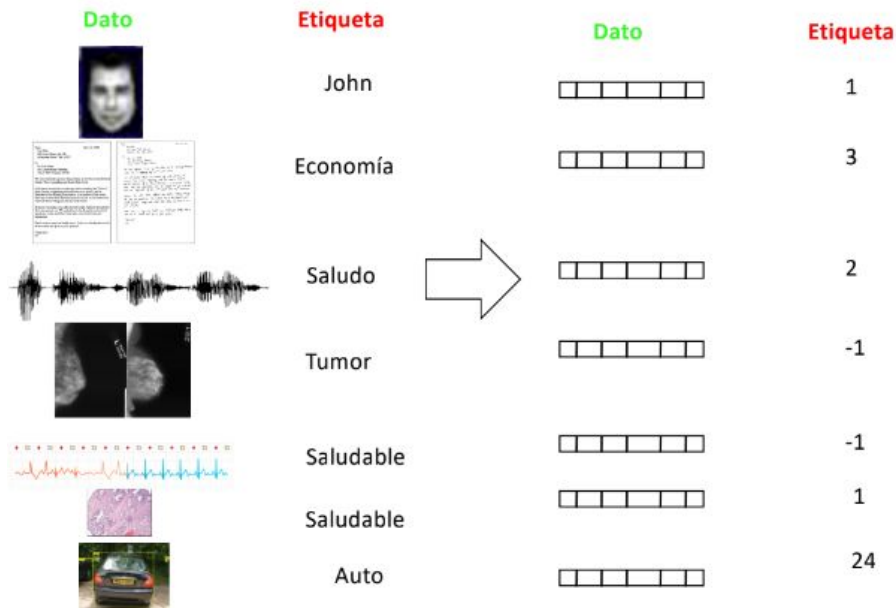
Datos etiquetados

Dato	Etiqueta
	John
	Economía
	Saludo
	Tumor
	Saludable
	Saludable
	Auto

Usualmente los datos se representan como vectores numéricos y las etiquetas con números entre 1 y K

Datos etiquetados

El paso que nos atañe ahora es cómo pasar el DATO a un **vector numérico** (que la computadora *entienda*) y las etiquetas a números.



Hugo Jair Escalante Main/Aprendizaje Computacional II 2020.
<https://ccc.inaoep.mx/~hugojaír/pmwiki/index.php?n=Main.AprendizajeComputacionalII2020>.

Datos etiquetados

Los datos se representan por un conjunto de atributos o mediciones

- Instancia: Fruta
 - Atributos: color, alto, ancho, masa
 - Etiqueta: manzana
-
- Instancia: Imágenes en color
 - Atributos: contenido de color en RGB, textura, intensidad, etc
 - Etiqueta: mujer con sombrero



Datos etiquetados

Los datos se representan por un conjunto de atributos o mediciones

- Instancia: Documento p.e. nota periodística
- Atributos: ??
- Etiqueta: Deportes

Tras la convincente **actuación sellada con victoria (0-2) en su debut frente a Estados Unidos**, Colombia afronta el partido de este martes contra **Paraguay**, con la intención de firmar el pase a cuartos de final frente a un rival con hambre y urgencias tras su empate frente a Costa Rica (0-0).

El Rose Bowl de Pasadena (California) acogerá el primer duelo de Copa América entre ambas selecciones desde la goleada (5-0) que los paraguayos infligieron a los cafeteros en 2007. Además, el choque supondrá el partido número cincuenta para José Pékerman al frente de Colombia.

Se trata de un envite relevante para ambas escuadras. **Una victoria pondría a Colombia en cuartos de final** y una derrota supondría la temprana eliminación de Paraguay.

El empate, por su parte, aún permitiría a los del argentino **Ramón Díaz mantener el sueño de poder clasificarse**, aunque centrarían todas sus opciones en el último encuentro del grupo contra Estados Unidos.

Para este partido, Colombia cuenta con la más que probable **baja de su capitán, James Rodríguez**, aún no confirmada oficialmente por Pékerman.

El jugador del Real Madrid recibió un fuerte golpe en el hombro izquierdo durante el duelo ante Estados Unidos, que le obligó a ser sustituido.

"Esperaremos a ver cómo evoluciona", dijo este domingo Néstor Lorenzo, entrenador asistente de la selección, en conferencia de prensa.

Dos enfoques en machine learning

J. Wang et al. / Journal of Manufacturing Systems 48 (2018) 144–156

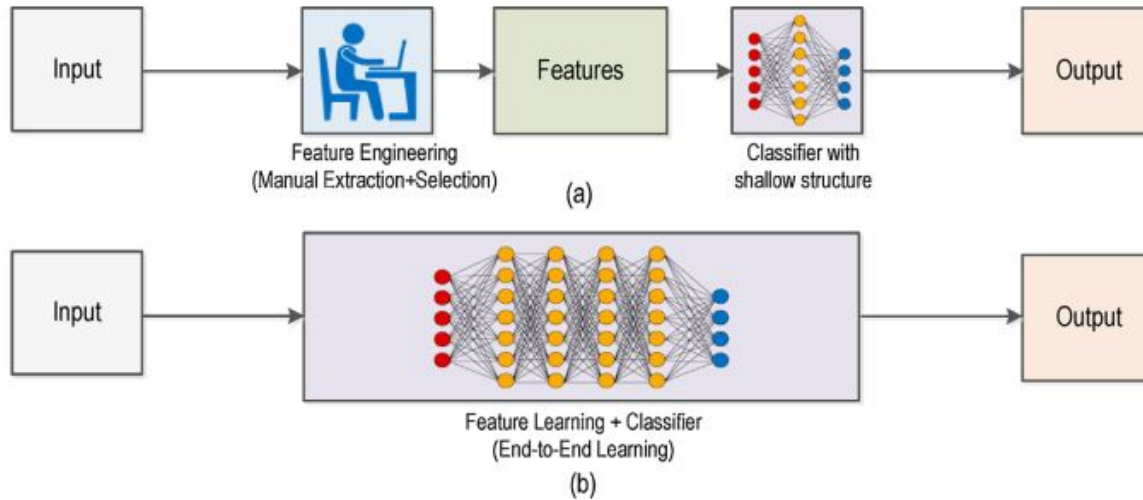
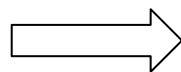


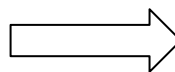
Fig. 2. Comparison between two techniques: a) traditional machine learning, b) deep learning.

Representación

Datos de
entrada



Ingeniería
de
atributos



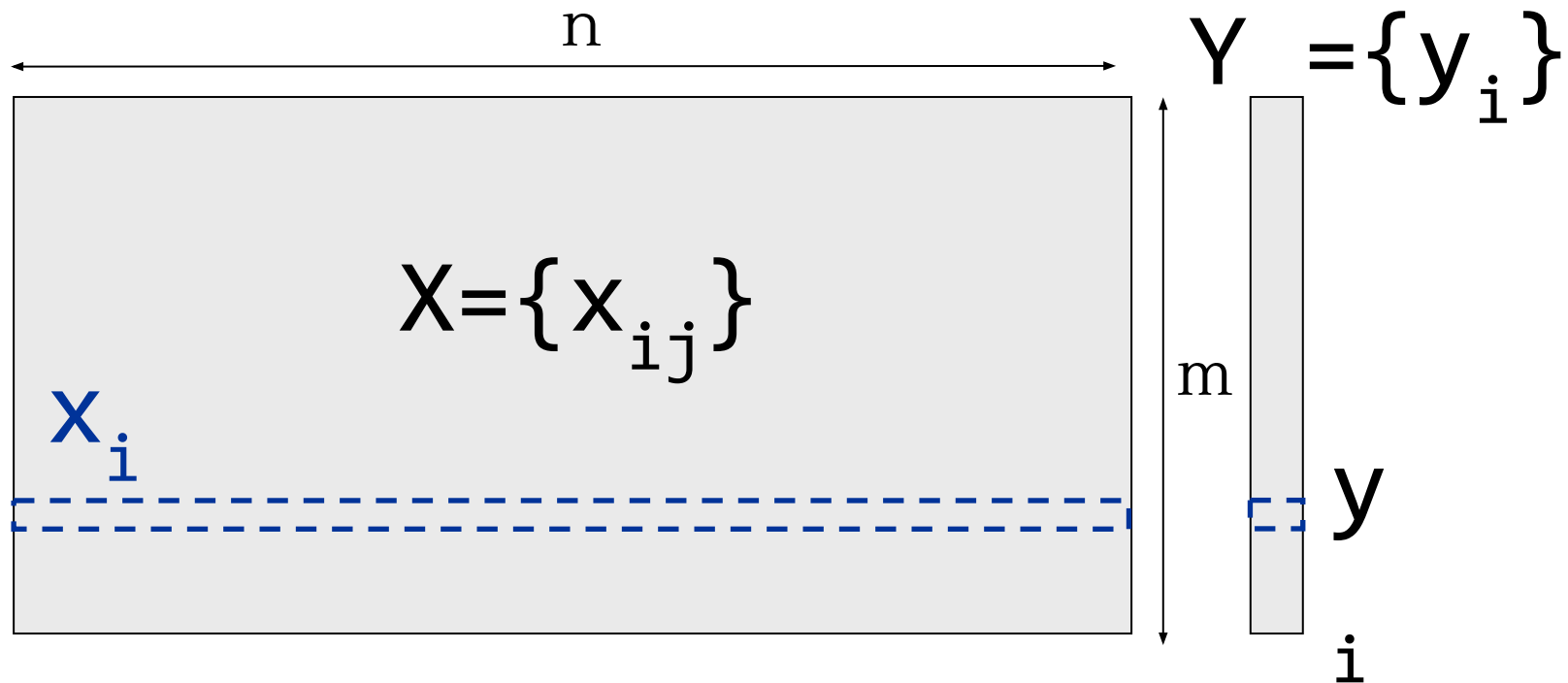
	t_1	t_1	...	t_n
d_1				
d_2				
:		w_{ij}		
d_m				



color: categórico,
altura: continuo (cm),
anchura: continuo (cm),
masa: continuo (g)

$x_1 = \langle \text{rojo}, 5.0, 5.5, 70 \rangle$

Convención

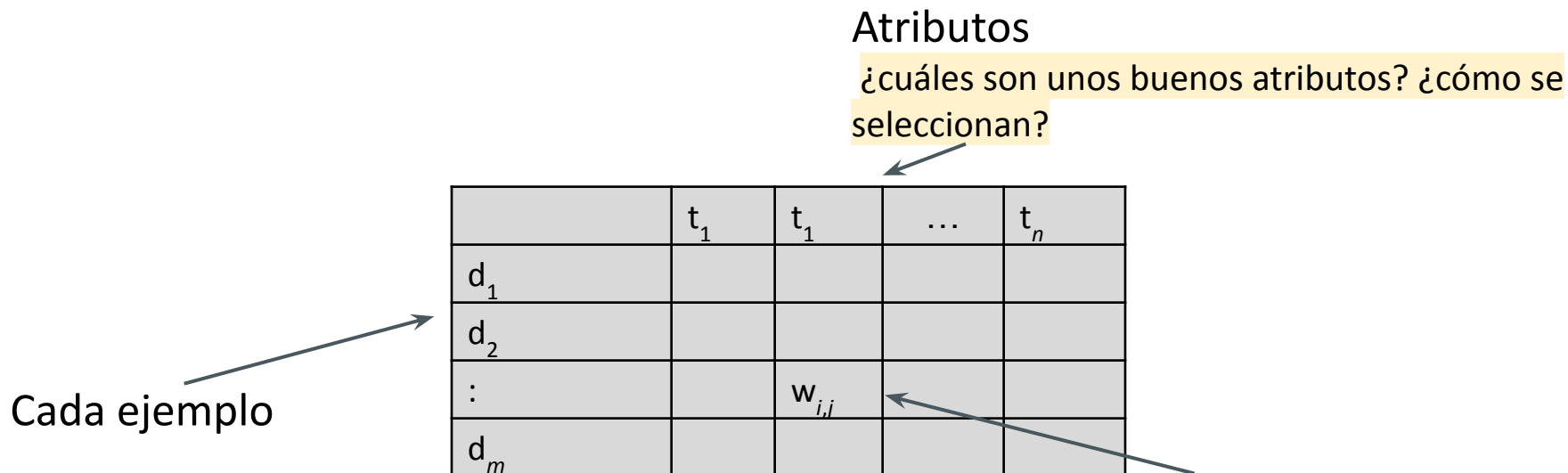


Representación vectorial

Atributos

¿cuáles son unos buenos atributos? ¿cómo se seleccionan?

Cada ejemplo



The diagram shows a feature matrix with m rows and n columns. The columns are labeled t_1, t_1, \dots, t_n at the top. The rows are labeled $d_1, d_2, :, d_m$ on the left. An arrow points from the text 'Cada ejemplo' to the first column. Another arrow points from the text 'Atributos' to the first row. A third arrow points from the text 'Peso del atributo j en un ejemplo i ' to the cell containing $w_{i,j}$ in the row labeled $:$ and the second column labeled t_1 .

	t_1	t_1	\dots	t_n
d_1				
d_2				
$:$		$w_{i,j}$		
d_m				

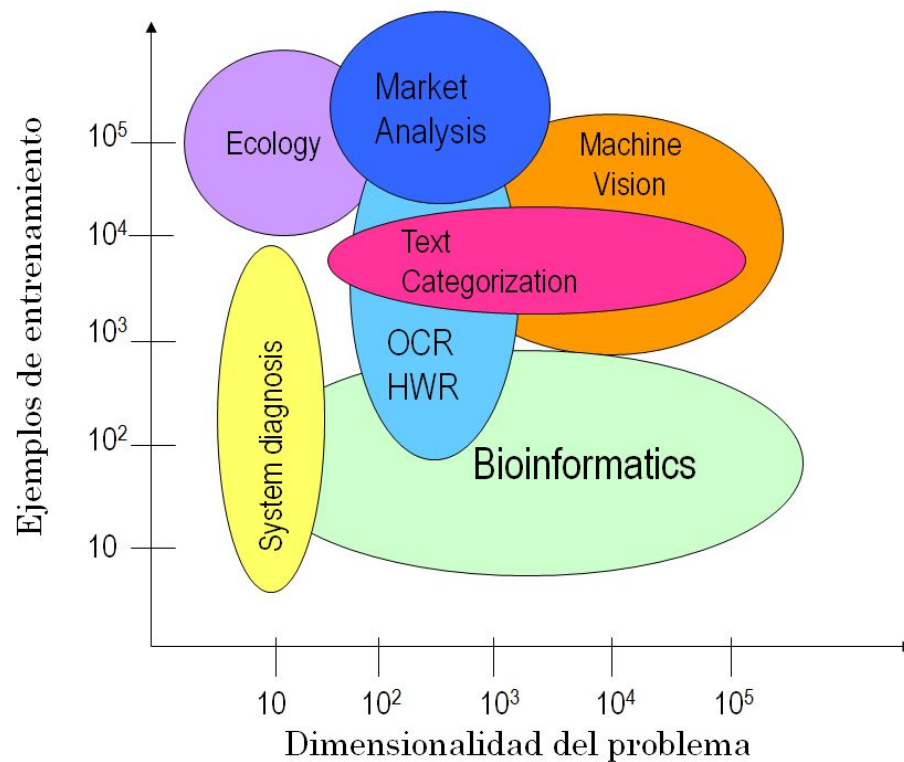
Peso del atributo j en un ejemplo i
¿Cómo determinar la importancia de un atributo?

Tipo de atributos

Dependiendo de la naturaleza de los ejemplos:

- Continuos (altura, temperatura, masa)
- Ordinales (“de acuerdo”, “ni de acuerdo ni en desacuerdo”, “en desacuerdo”)
- Categóricos (e.g. color, género, tema)

Dimensionalidad de la matriz

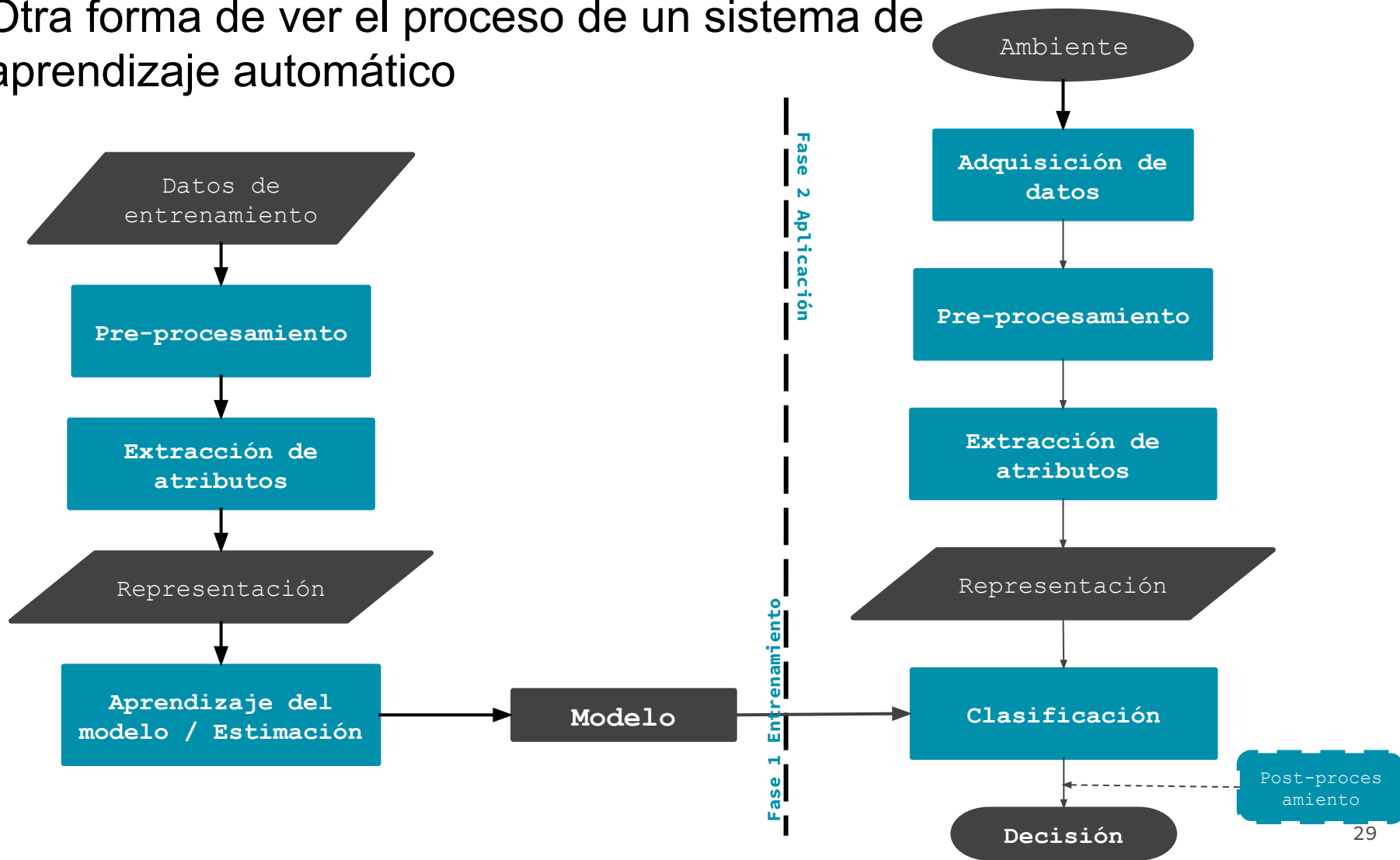


Ejercicio 1 en codalab

<https://github.com/gabyrr/CursoTIA>

- TIA:Explorando_datasets.ipynb

Otra forma de ver el proceso de un sistema de aprendizaje automático



Aprendizaje del modelo / Estimación: ¿Qué algoritmo de aprendizaje utilizar?

- Clasificación basada en prototipos
- K-vecinos más cercanos
- Regresión logística
- Discriminadores lineales
- Clasificador Bayesiano simple (Naïve Bayes)
- Redes neuronales artificiales
- Árboles de decisión
- Máquinas de vectores de soporte (Support vector machines)
- Ensamblados

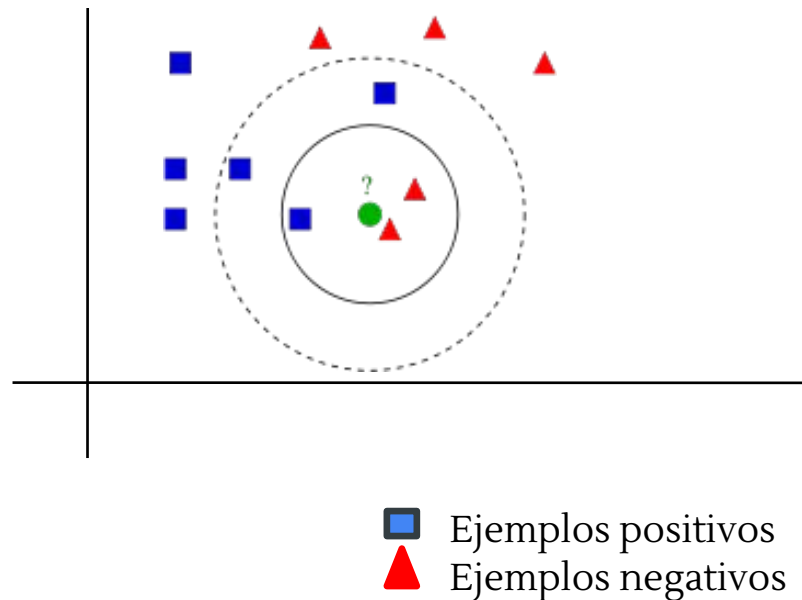
Métodos basados en similitud

Se basan en medidas de similitud entre patrones de entrenamiento y de prueba.

- Generalmente no hay etapa de “aprendizaje”

Ejemplo: **clasificador KNN**

- Asignar la clase dominante en los k-vecinos más cercanos a la instancia de prueba



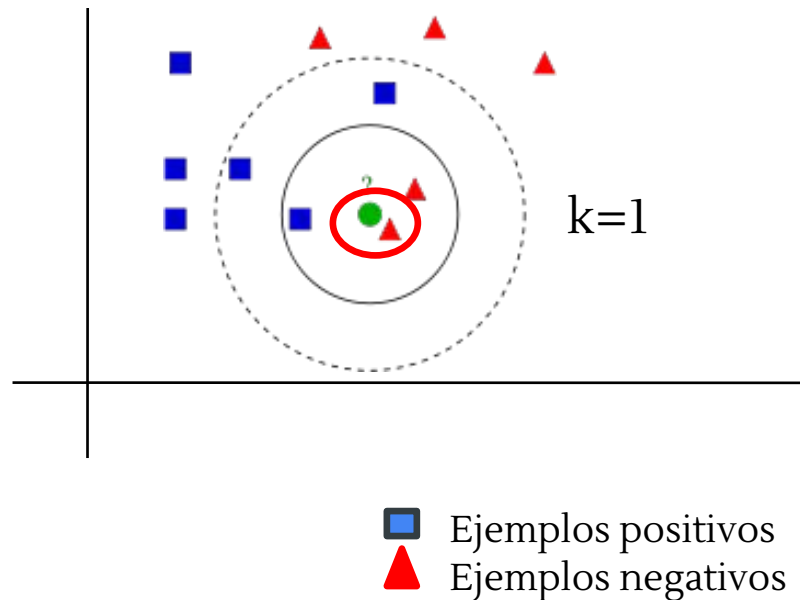
Métodos basados en similitud

Se basan en medidas de similitud entre patrones de entrenamiento y de prueba.

- Generalmente no hay etapa de “aprendizaje”

Ejemplo: **clasificador KNN**

- Asignar la clase dominante en los k -vecinos más cercanos a la instancia de prueba



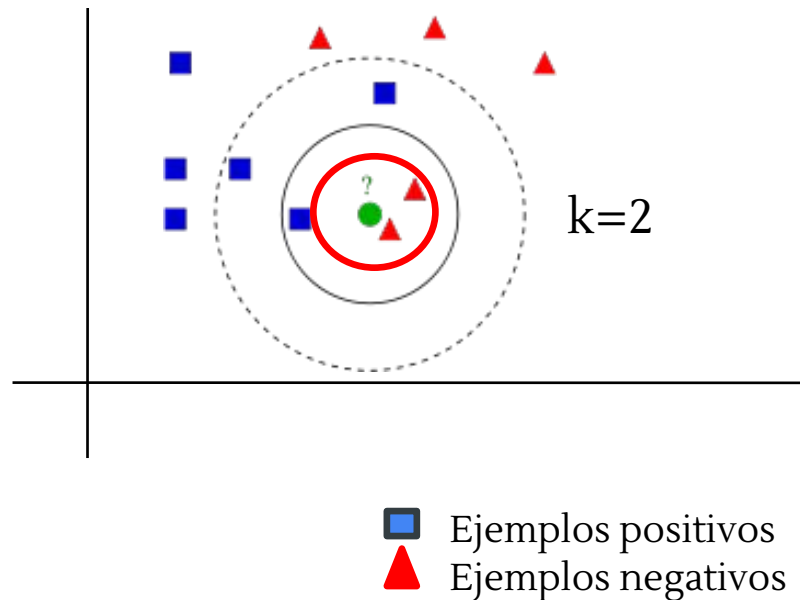
Métodos basados en similitud

Se basan en medidas de similitud entre patrones de entrenamiento y de prueba.

- Generalmente no hay etapa de “aprendizaje”

Ejemplo: **clasificador KNN**

- Asignar la clase dominante en los k -vecinos más cercanos a la instancia de prueba



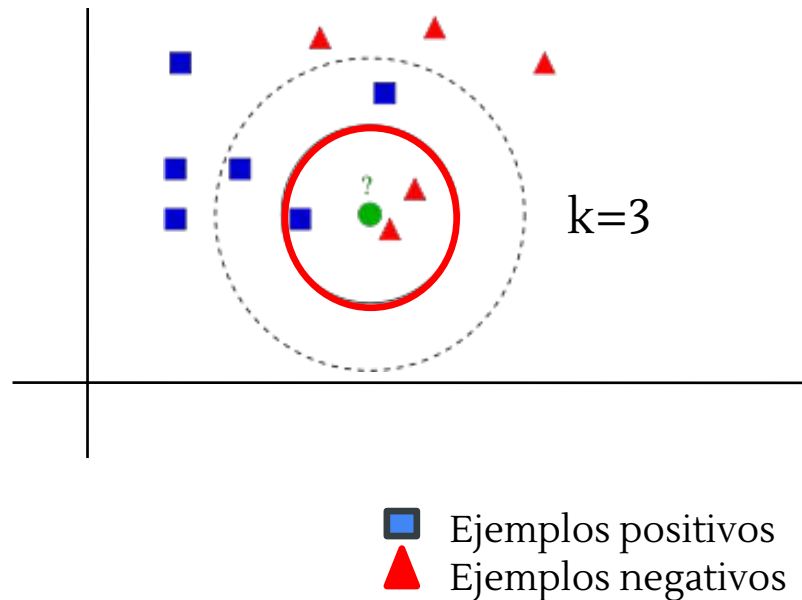
Métodos basados en similitud

Se basan en medidas de similitud entre patrones de entrenamiento y de prueba.

- Generalmente no hay etapa de “aprendizaje”

Ejemplo: **clasificador KNN**

- Asignar la clase dominante en los k -vecinos más cercanos a la instancia de prueba



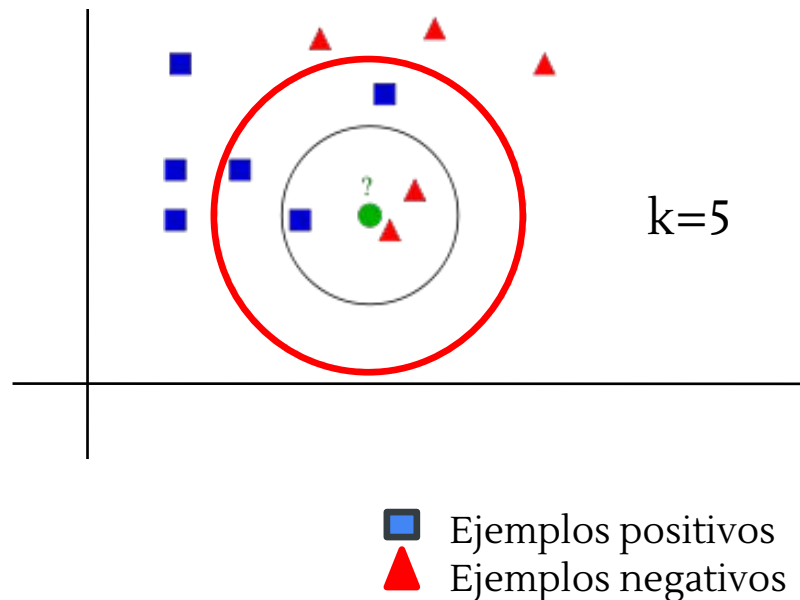
Métodos basados en similitud

Se basan en medidas de similitud entre patrones de entrenamiento y de prueba.

- Generalmente no hay etapa de “aprendizaje”

Ejemplo: **clasificador KNN**

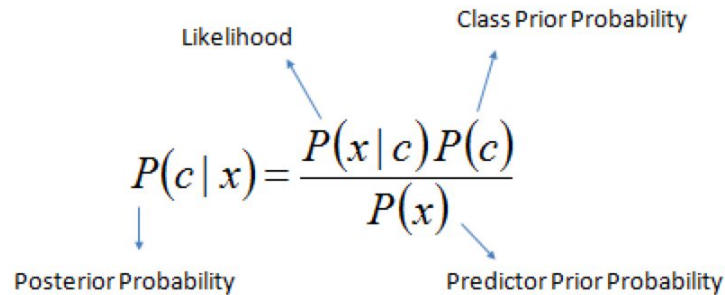
- Asignar la clase dominante en los k -vecinos más cercanos a la instancia de prueba



Métodos probabilísticos

Bayes's rule

Una regla para actualizar la probabilidad de una hipótesis c dado un dato x .

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$


The diagram shows the Bayes' rule formula with four labels and arrows pointing to the corresponding terms: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$P(c|x)$ - probabilidad posterior de la clase c dado el dato x .

$P(c)$ - probabilidad a priori de la clase c : qué creía antes de ver el dato x

$P(x|c)$ - probabilidad del dato x dada la clase c (calculada a partir de los datos de entrenamiento)

$P(x)$ - probabilidad previa de los datos (verosimilitud marginal): la probabilidad de los datos x bajo cualquier circunstancia (sin importar la clase)

Métodos probabilísticos

Ejemplo: prueba de COVID

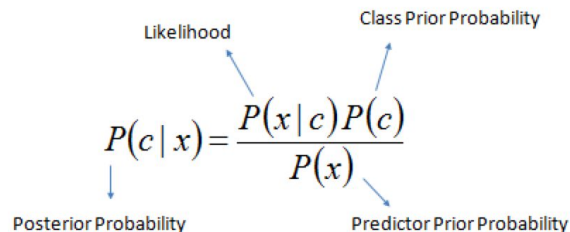
Cuál es la probabilidad de tener COVID-19 si el 96% de las pruebas son positivas. Asumiendo que la tasa de falsos positivos es del 4%

- $P(C) = 0.096$ (760M casos, 7900M personas)
- $P(TP) = P(\text{pos}|C) = 0.96$
- $P(FP) = P(\text{pos}|\text{not}C) = 0.04$

Si el test es positivo, entonces $P(C)=0.718$

$$P(C|\text{pos}) = P(\text{pos}|C) * P(C) / P(\text{pos})$$

$$P(\text{pos}) = P(\text{pos}|C) * P(C) + P(\text{pos}|\text{not}C) * (1 - P(C))$$



The diagram shows the formula for Posterior Probability: $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$. Arrows point from labels to parts of the formula: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

¿Cuál es la probabilidad de tener COVID dada una segunda prueba positiva?

Naive Bayes

Predice la probabilidad que un punto pertenezca a cierta clase, usando el teorema de Bayes

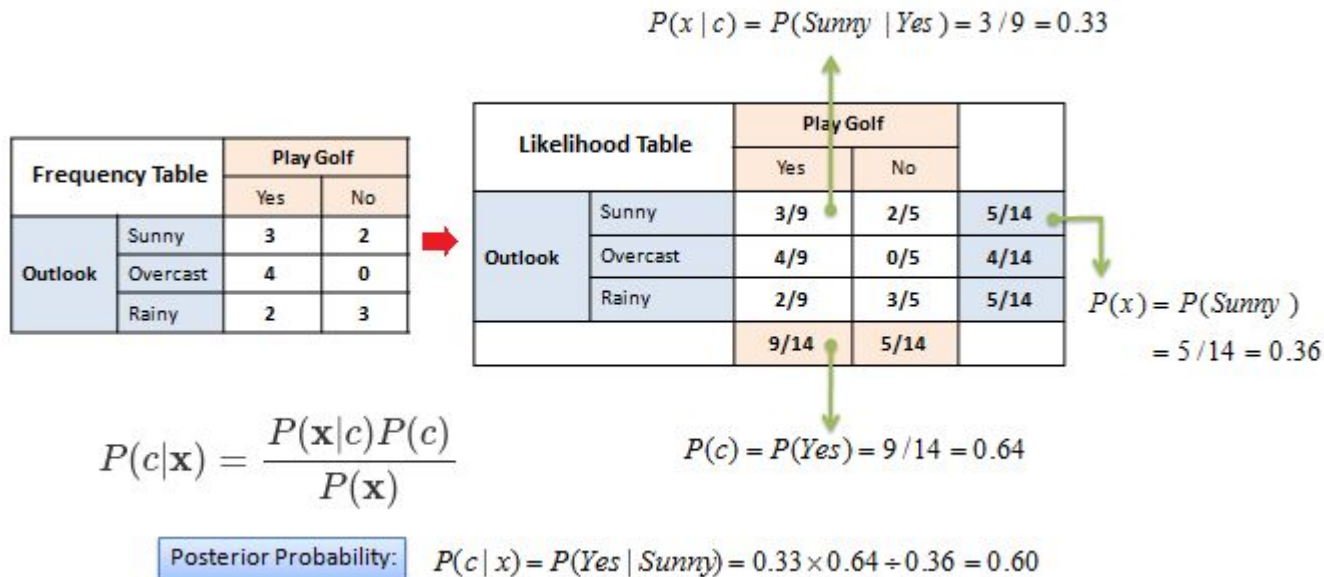
$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})}$$

Problema: \mathbf{x} es un vector!!

- Calcular $P(\mathbf{x}|c)$ es muy complejo,
- Se asume que todos los atributos son condicionalmente independientes, de modo que:
- $P(\mathbf{x}|c) = P(x_1|c) * P(x_2|c) * \dots * P(x_n|c)$
- Muy rápido pues solo necesita calcular estadísticas de cada atributo

Naive Bayes - ejemplo con datos categóricos

¿Cuál es la probabilidad de que tu amigo juegue golf si está soleado?



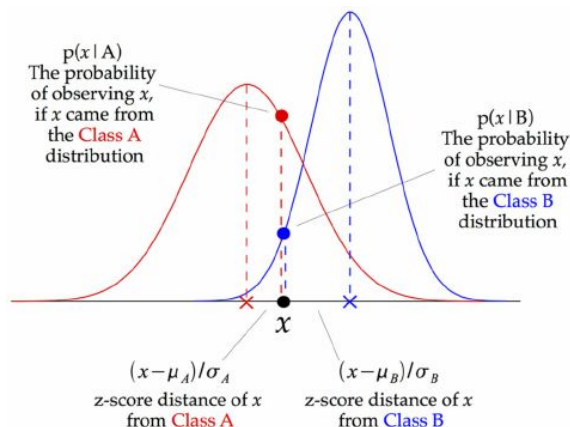
¿Cuál es la probabilidad de que no vaya a jugar golf si está soleado?

Naive Bayes - ejemplo con datos numéricos

¿Cuál es la probabilidad de que tu amigo juegue golf si está soleado?

- Primero, necesitamos ajustar la distribución de los datos (p.e. Gaussiana)
- **GaussianNB**: calcula la media y la desviación estándar de los valores de los atributos por clase

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

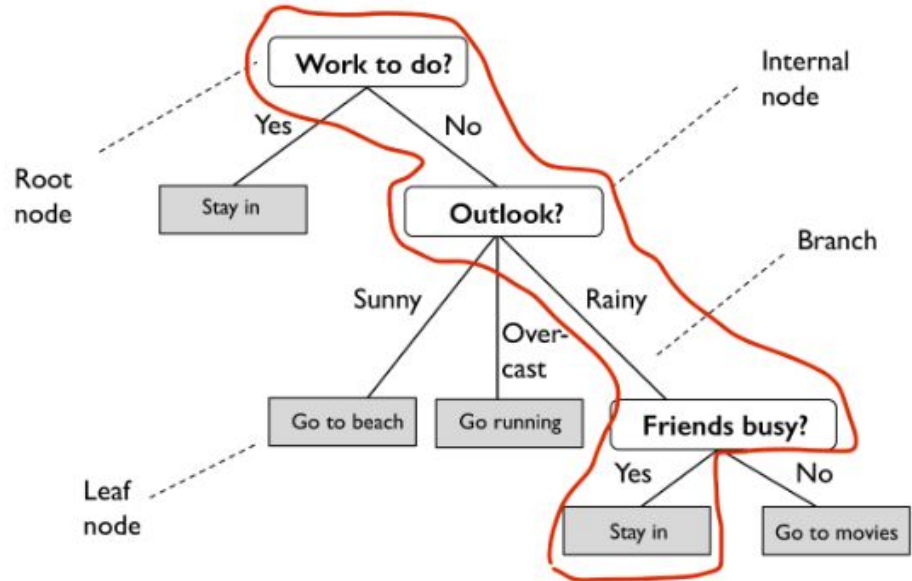


- Ahora ya se pueden hacer predicciones usando el teorema de Bayes:

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})}$$

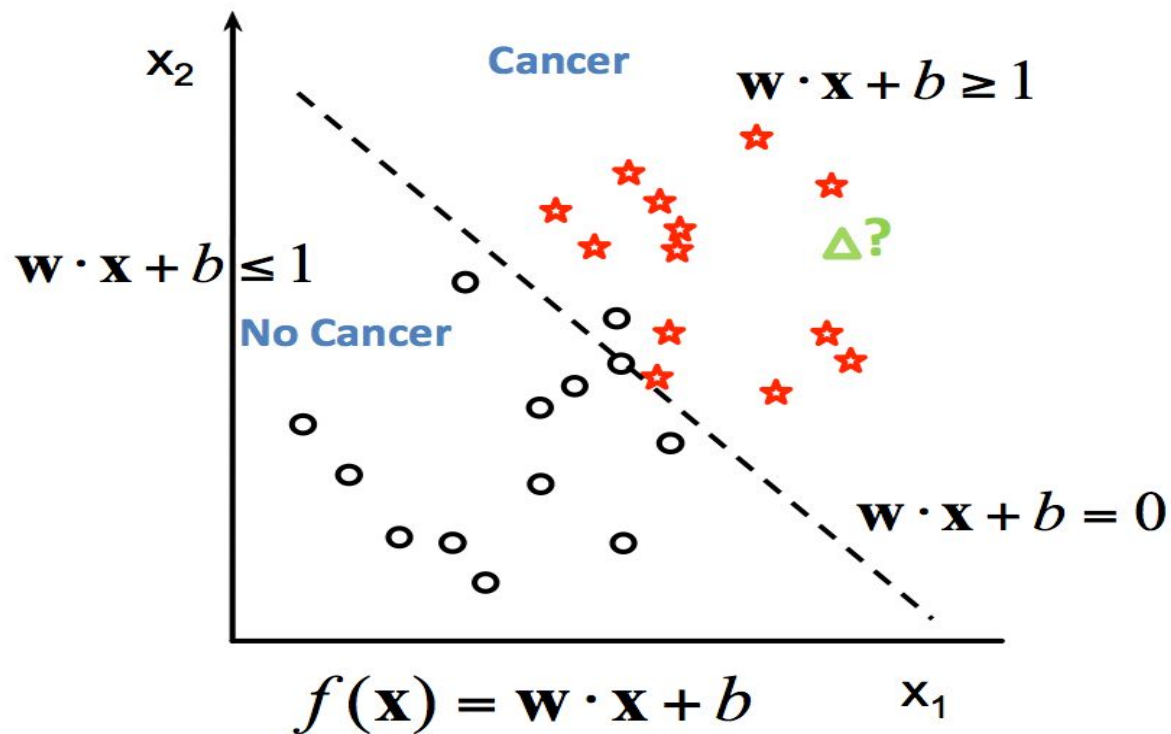
Árboles de decisión

- El conocimiento extraído es fácilmente interpretable por humanos
- Llegar a una conclusión hay que seguir reglas:
 - p.e. Si no tengo trabajo que hacer Y está lluvioso Y mis amigos están ocupados, entonces me quedo en casa.



Modelos lineales

- Generan una función lineal para discriminar entre elementos de dos clases.
- Aprende w dado un conjunto de datos X , dado a una función de pérdida
-



Modelos lineales para clasificación

- El objetivo es encontrar un hiperplano que separe los ejemplos de cada clase
- Para clasificación binaria, el objetivo es ajustar la función:

$$y = w_1 * x_1 + w_2 * x_2 + \dots + w_p * x_p + w_0 > 0$$

- Cuando $y < 0$, se predice la clase -1; en otro caso se predice la clase +1

Modelos lineales para clasificación

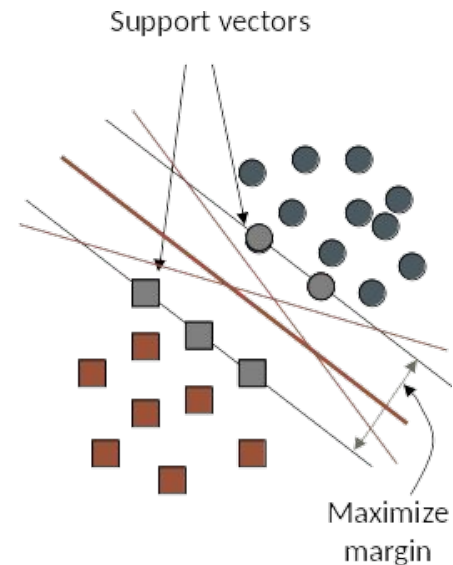
Existen varios algoritmos para clasificación lineal. Técnicas más comunes y ejemplos de algoritmos:

- Convierte las clases objetivo {neg, pos} en {0,1} y los trata como un problema de regresión
 - Logistic regression (log loss)
 - Ridge classification (mínimos cuadrados + L2 loss)
- Encuentran el hiperplano que maximiza el margen entre clases
 - Linear SVM (Hinge loss)
- Redes neuronales sin funciones de activación
 - Perceptrón (perceptron loss)

SVM (Modelos de vectores de soporte)

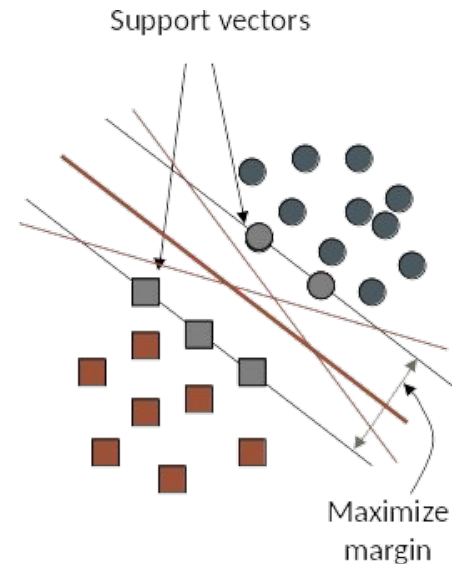
Un clasificador binario SVM se puede ver como un hiperplano en el espacio de atributos que separa los puntos que representan instancias positivas de las negativas.

- SVM selecciona el hiperplano que maximiza el margen alrededor de las instancias (ejemplos) representativos



SVM (Modelos de vectores de soporte)

- Es flexible al elegir la mejor función de similitud (kernels)
- Funciona bien con bases de datos grandes, pues al final solo importan los vectores de soporte (en la frontera de las clases)
- Maneja bien espacios de muy alta dimensión (muchos atributos)



Ejercicio 2 en codalab

<https://github.com/gabyrr/CursoTIA>

- TIA_algoritmos_aprendizaje.ipynb

Evaluación

La Evaluación nos permitirá
contestar a:

¿Cómo elegir el mejor modelo?

Aprendizaje del
modelo / Estimación



Modelo

Selección del modelo

¿Cómo saber si estamos aprendiendo lo correcto?

¿Estamos sobreajustando (*overfitting*) el modelo?

El modelo debe seleccionarse cuidadosamente para ajustarse a la aplicación

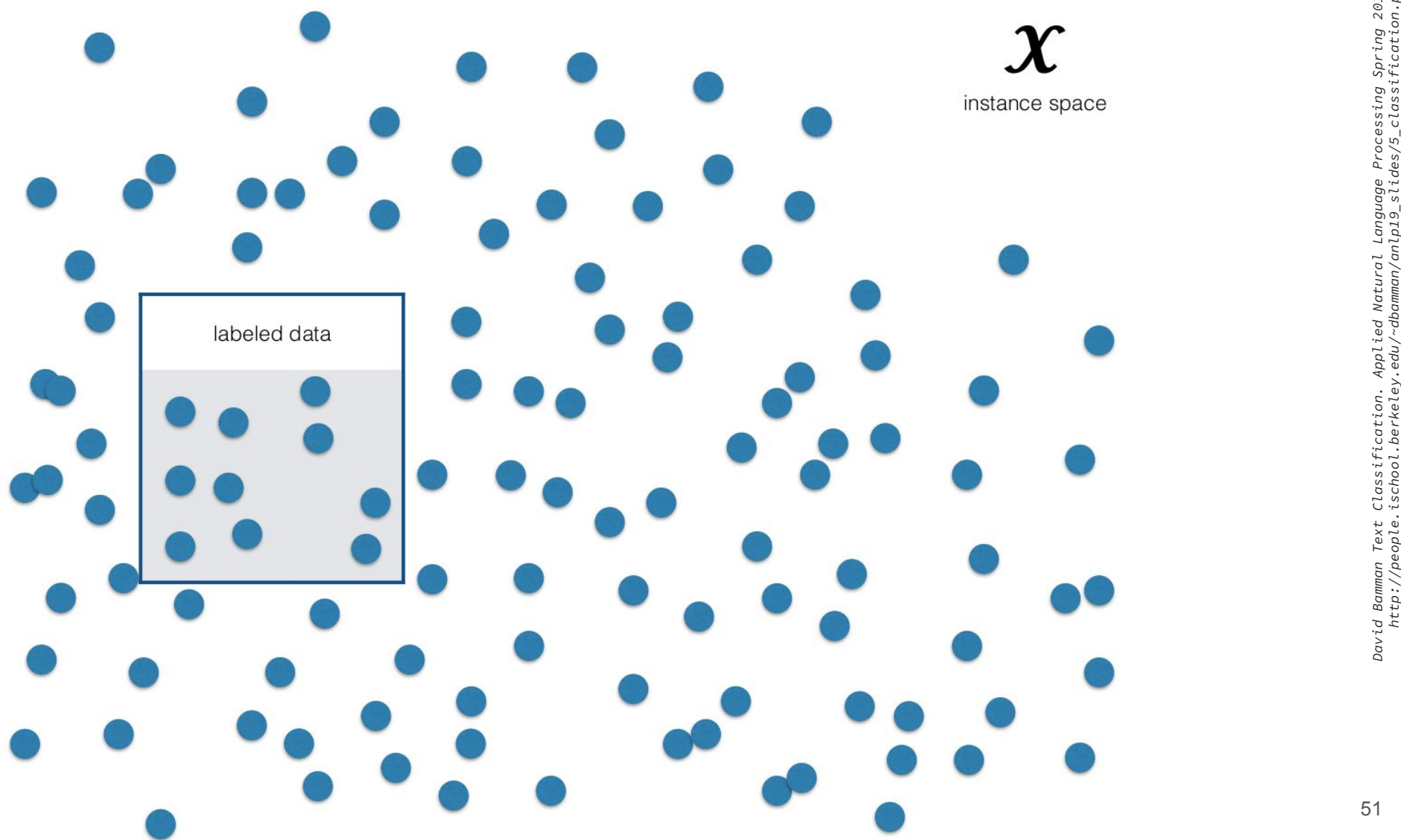
Debe existir una forma de selección entre modelos (y configuraciones de hiperparámetros)

Los datos etiquetados son escasos. Del universo de posibles datos, una pequeña parte está etiquetada.

- Entrenar el modelo (**train**),
- Optimizar el modelo (**dev** -*opcional*-), y
- Evaluar el desempeño del modelo (es muy importante que este conjunto se guarde -no se vea nunca- mientras construimos y adaptamos el modelo) (**test**).

Dos enfoques:

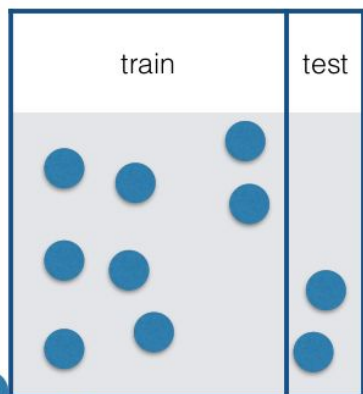
- División fija (**split**)
- Cross Validation

 X

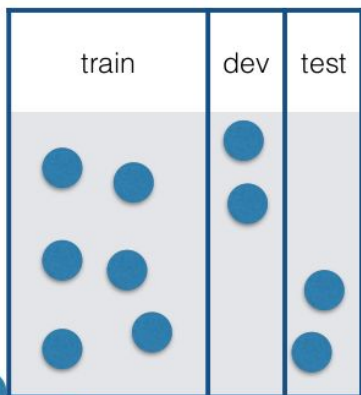
instance space

x

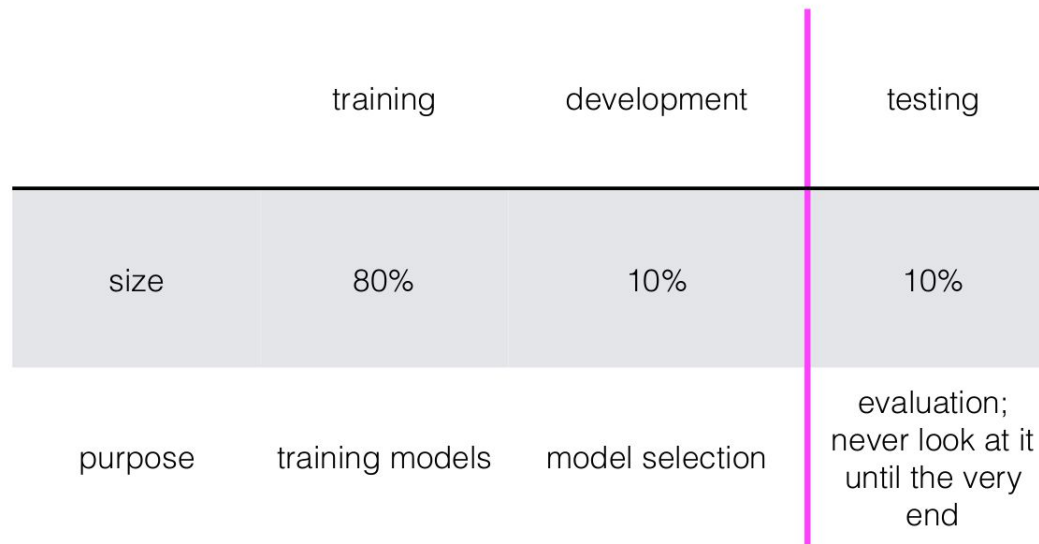
instance space



instance space



Experiment design



54
David Bamman Text Classification. Applied Natural Language Processing Spring 2019.
http://people.ischool.berkeley.edu/~dbamman/anlp19_slides/5_classification.pdf

Primer enfoque de evaluación: SPLIT

Primer enfoque de evaluación: SPLIT

Especialmente **útil** cuando se cuenta con una cantidad **abundante de datos**.

Usualmente se usa este enfoque cuando se quieren hacer **comparaciones directas** con diferentes sistemas (la partición de TEST se usa para todos los sistemas a evaluar)

Se usa la partición de **TRAIN** para aprender el modelo.

Cuando los algoritmos tienen parámetros a optimizar, se usa la partición **DEV** para hacer estos ajustes al modelo.

La partición **TEST** se sólo al final, para medir el desempeño general del modelo.

Segundo enfoque: n-fold cross validation

- Útil cuando hay pocos datos etiquetados.
- Bajo este enfoque usamos todo el conjunto de datos pero de forma sistemática, siempre habrá un conjunto que no se use para entrenar el modelo.
- En este ejemplo, se hace una evaluación: **5-folds cross validation**:
 - Se repite el procedimiento 5 veces
 - En cada vez se divide el conjunto en 5 partes iguales
 - Se usan 4 para **TRAIN** y se deja 1 parte para **TEST**



Métricas tradicionales de evaluación

- Medida-F
- Accuracy (exactitud)
- Precisión
- Recuerdo

Todas estas medidas se obtienen de la matriz de confusión

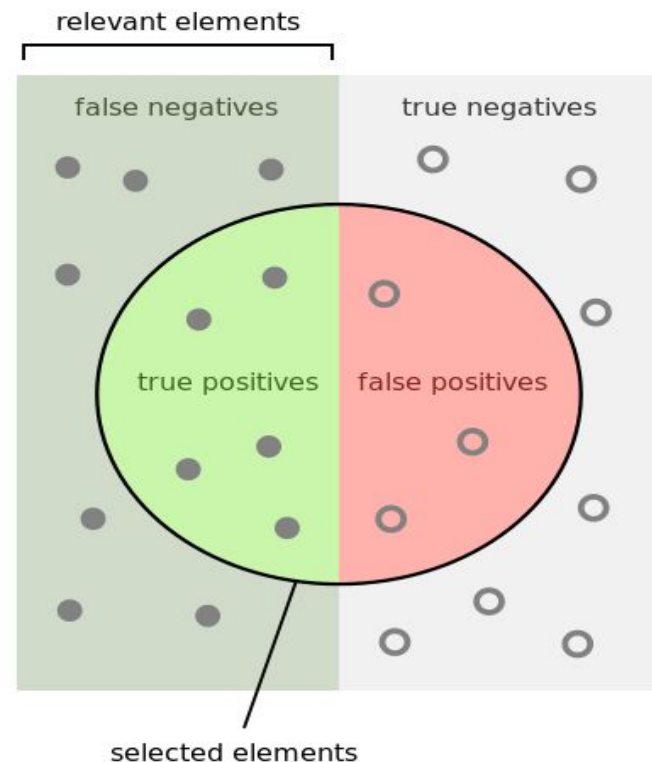
Matriz de confusión

Para un problema de clasificación binaria:

- Positivo (+)
- Negativo (-)

El modelo clasifica como:

- True Positives (TP) : positivo los datos que están etiquetados como positivo
- True Negatives (TN) : negativo los datos que están etiquetados como negativos
- False Positives (FP) : positivos los datos que están etiquetados como negativos
- False Negatives (FN) : negativos los datos que están etiquetados como positivos



		Class	
		+	-
Classified	+	TP	FP
	-	FN	TN

Matriz de confusión

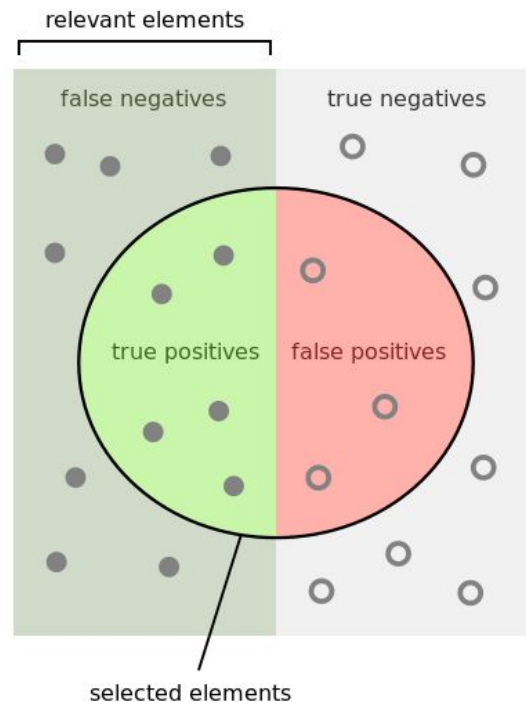
- Exactitud (Acc), indica el porcentaje de elementos clasificados correctamente

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}$$

- Precisión: $P = \frac{TP}{TP + FP}$

- Recuerdo: $R = \frac{TP}{TP + FN}$

- F-measure: $F_measure = 2 \cdot \frac{P \cdot R}{P + R}$



How many selected items are relevant?

$$Precision = \frac{\text{Green}}{\text{Green} + \text{Red}}$$

How many relevant items are selected?

$$Recall = \frac{\text{Green}}{\text{Green}}$$

Matriz de confusión: ejemplo

Piense en un sistema que entrenó para detectar **cáncer** en pacientes

- La distribución de sus datos de evaluación son: 100 pacientes tienen cáncer y 300 no tienen cáncer

	Cáncer	No_Cáncer
Cáncer	100	0
No_cáncer	0	300

- Si empleamos la medida de Exactitud para medir el desempeño, tenemos **ACC= 1.0**

Matriz de confusión: ejemplo

¿Qué pasa en el siguiente escenario?

	Cáncer	No_Cáncer
Cáncer	20	0
No_cáncer	80	300

$ACC = 0.8$

- Note que en este problema nos interesa clasificar de bien los casos positivos (los pacientes con cáncer)
- Este clasificador sólo identifica 20 de 100 casos

¿Qué pasa en el siguiente escenario?

$$Acc = 0.80$$

$$P = ?$$

$$R = ?$$

$$F = ?$$

	Cáncer	No_Cáncer
Cáncer	20	0
No_cáncer	80	300

$$P_{(+)} = ?$$

$$R_{(+)} = ?$$

$$F_{(+)} = ?$$

$$P_{(-)} = ?$$

$$R_{(-)} = ?$$

$$F_{(-)} = ?$$

¿Qué pasa en el siguiente escenario?

	Cáncer	No_Cáncer
Cáncer	20	0
No_cáncer	80	300

$$Acc = 0.80$$

$$P = 0.89$$

$$R = 0.60$$

$$F = \mathbf{0.60}$$

¿Qué medida conviene más?

$$P_{(+)} = 1.0$$

$$R_{(+)} = 0.20$$

$$F_{(+)} = \mathbf{0.33}$$

$$P_{(-)} = 0.78$$

$$R_{(-)} = 1$$

$$F_{(-)} = \mathbf{0.87}$$

Ejercicio:

Para la siguiente matriz de confusión, calcule las medidas de P, R, F globales así como las medidas para cada clase.

	Cáncer	No_Cáncer
Cáncer	60	70
No_cáncer	40	230

Últimos puntos

- Si los datos están muy desbalanceados conviene emplear métricas como la medida F
- La matriz de confusión es un buen medio para visualizar el comportamiento del clasificador
- Sólo cuando tiene datos balanceados puede confiar (hasta cierto punto) en la medida ACC

Al hacer particiones, tomar en cuenta:

El dataset debe dividirse en conjuntos de **entrenamiento** y **prueba**

- Optimizar los parámetros del modelo en el conjunto de **entrenamiento**, evaluar el desempeño en el conjunto de **prueba**

Evite fugas de dato:

- Nunca optimizar la configuración de hiper-parámetros en el conjunto de **prueba**
- Nunca se debe elegir el tipo de preprocesamiento basado en el conjunto de **prueba**

Para optimizar parámetros y preprocesamiento, se deja apartado un pedazo del conjunto de **entrenamiento** como conjunto de **validación**

- Mantener oculto el conjunto de prueba durante todo el proceso de entrenamiento

Aprendizaje = Representación + evaluación + optimización

Todos los algoritmos de aprendizaje consisten de 3 componentes:

1. **Representación**.- Un modelo debe ser representado en un lenguaje formal que la computadora pueda manejar
 - a. Define los conceptos que puede aprender, el espacio de hipótesis
2. **Evaluación**.- Una manera interna de elegir una hipótesis sobre otra
 - a. La función objetivo. función de medición, función de pérdida
3. **Optimización**. - Una forma eficiente de buscar en el espacio de hipótesis
 - a. Empieza por una hipótesis simple, la relaja si no se ajusta a los datos
 - b. Empieza con un conjunto inicial de parámetros, y gradualmente los refina

Ejercicio 2 en codalab

Parte 2

<https://github.com/gabyrr/CursoTIA>

- TIA_algoritmos_aprendizaje.ipynb

DATOS en la vida real

Importancia de los datos



Más que ver la calidad de los datos, en las siguientes diapositivas veremos los retos que conlleva trabajar con datos en la vida real.

Datos crudos - raw data-

- Los datos vienen en múltiples formas y tamaños
 - Archivos CSV (valores separados por comas)
 - PDFs
 - SQL dumps
 - .txt
 - .png
 - etc
- Los archivos vienen con diferentes formatos
 - Con comillas dobles en lugar de simples
 - Con codificación diferente
 - Cadenas vacías o espacios en lugar de NULL
 - Renglones con títulos extras
 - etc
- Los datos están “sucios”
 - Duplicados
 - Anomalías no deseadas

Domando a los datos

(Data wrangling o data munging)

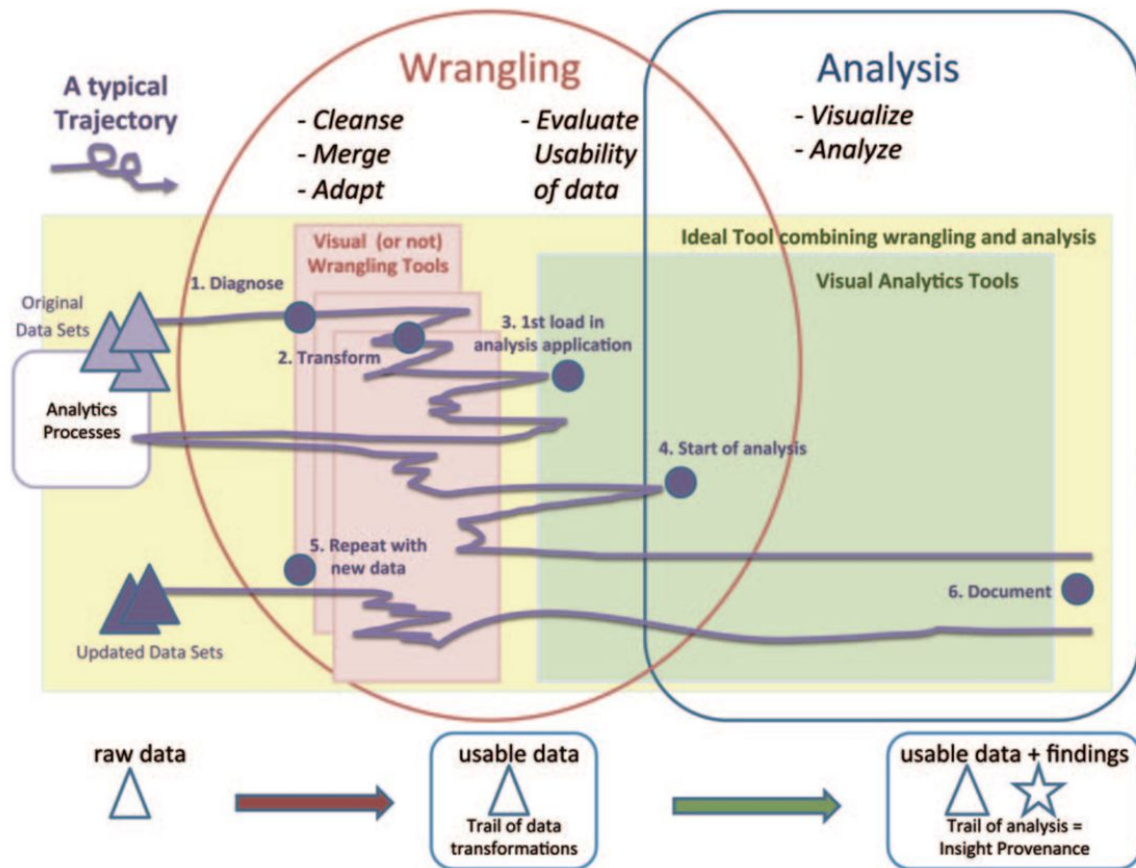
¿Qué es *data wrangling*?

Objetivo: extraer y estandarizar los datos crudos (combinando múltiples recursos y limpiando anomalías)

Estrategia: Combinar automatización con visualizaciones interactivas (python puede ayudar aquí) para ayudar en la limpieza

- Proceso iterativo

Un proceso a la medida



Tipos de problemas de datos

- Datos faltantes
 - No todos los usuarios tienen asignada una ubicación
- Datos incorrectos
- Inconsistencias en la representación del mismo conjunto de datos
 - un color se puede representar con su código RGB o CMYK
- Cerca del 75% de los problemas en los datos requiere intervención humana (e.g. expertos, crowdsourcing, etc)
- Hay que saber balancear entre limpiar los datos y sobre-sanitizar los datos
 - Queremos poder usarlos de forma automática
 - No queremos eliminar elementos inherentes al fenómeno que queremos estudiar

Diagnosticando problemas con los datos

La visualización y estadísticas básicas pueden apoyar en identificar los problemas en los datos

- Diferentes representaciones resaltan diferentes tipos de problemas
 - Outliers (elementos raros) pueden ser identificados en ciertas gráficas
 - Datos faltantes pueden casar espacios o valores ceros en algunas gráficas
- Este enfoque puede resultar complicado conforme el conjunto de datos se vuelve grande o muy grande



Práctica

- Hacer el tutorial de Rachael Tatman en Kaggle (posiblemente necesiten una cuenta en Kaggle)

<https://www.kaggle.com/rtatman/data-cleaning-challenge-handling-missing-values/notebook>