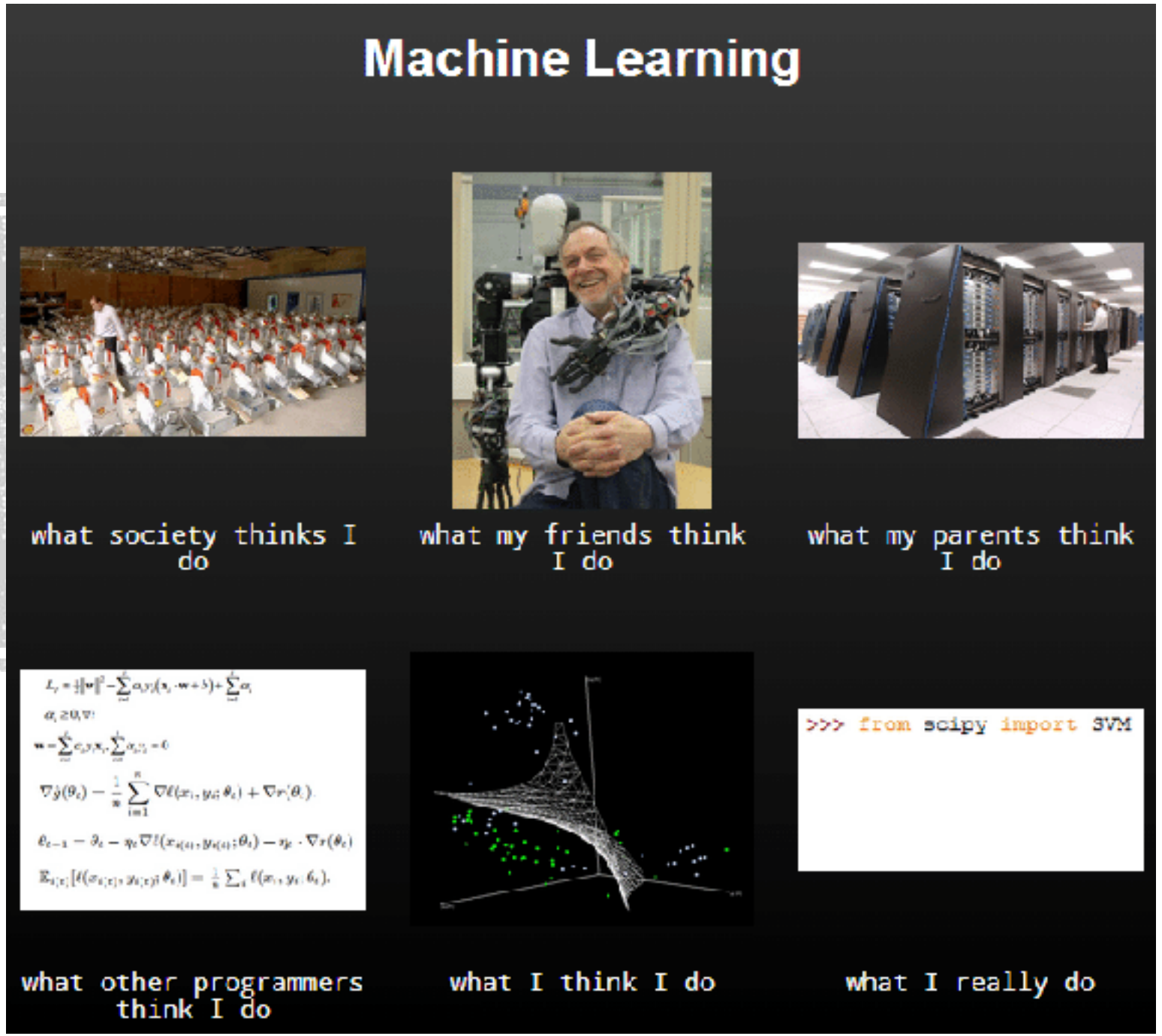
APRENDIZAJE AUTOMÁTICO

Anibal Sosa, PhD

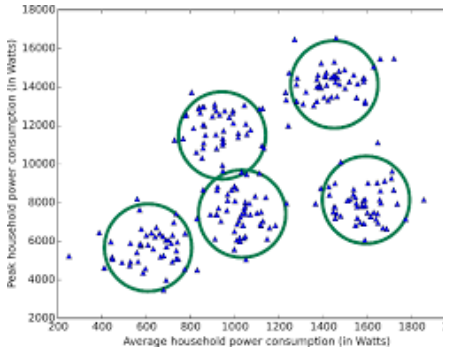


UNIVERSIDAD ICESI

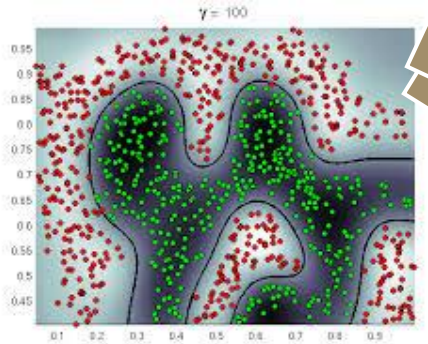
OBJETIVOS DE APRENDIZAJE

- Identificar problemas que se puedan resolver a partir de modelos de aprendizaje supervisado, ya sea de clasificación o de regresión.
- Aplicar modelos de K-NN y de regresión logística a conjuntos de datos, para responder a preguntas de negocio involucrando modelos de aprendizaje supervisado utilizando el lenguaje python.
- Comparar modelos de aprendizaje supervisado con respecto a métricas de ajuste, utilizando diferentes protocolos de evaluación

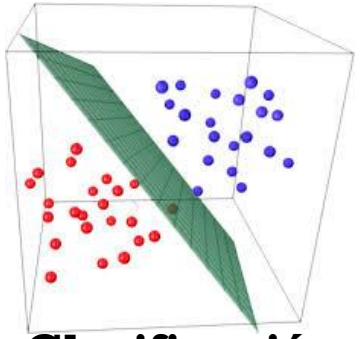
Aprendizaje automático



Aprendizaje no supervisado

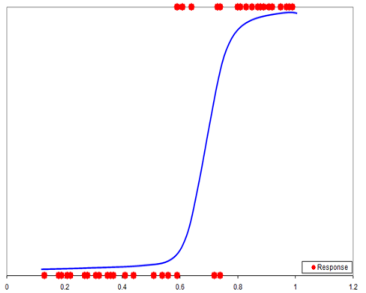


Aprendizaje supervisado

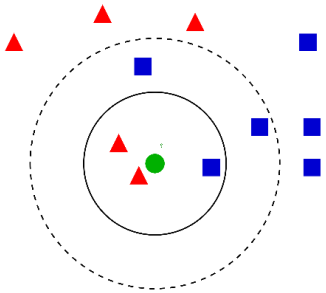


Clasificación

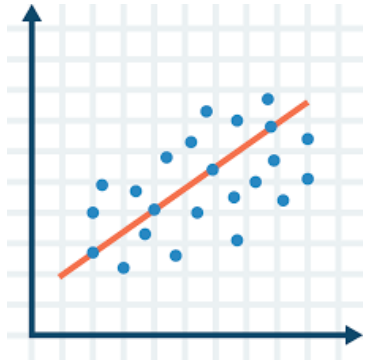
Métricas de Evaluación de la clasificación



Regresión logística



KNN



Regresión

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

DATOS:
Materia prima

PREGUNTA:

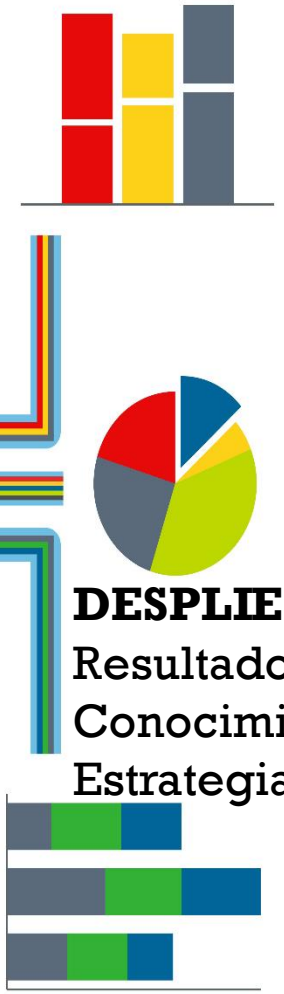
¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

ANALISIS EXPLORATORIO:
Limpieza de datos
Preparación / transformación
Escogencia de variables

MODELO:
Implementación del modelo de analítica

EVALUACION:
Calidad del modelo
Ajuste del modelo

DESPLIEGUE:
Resultados
Conocimiento
Estrategia



UNIVERSIDAD ICESI

APRENDIZAJE AUTOMÁTICO (APRENDIZAJE DE MAQUINA VS APRENDIZAJE ESTADISTICO)

- El aprendizaje de máquina tiene que ver más con los resultados. En este sentido su valor será caracterizado únicamente por su rendimiento.
- En el modelado estadístico se trata más de encontrar relaciones entre variables y la importancia de esas relaciones, al tiempo que se puede buscar una predicción.

APRENDIZAJE SUPERVISADO

¿Cómo le puedo enseñar a un niño que es una pelota?

Set de entrenamiento

¿Qué patrones distinguen las pelotas de los demás juguetes?

¿Es esta una pelota?

UNIVERSIDAD ICESI

APRENDIZAJE NO SUPERVISADO

Organiza tus juguetes

¿Qué estructura hay en los datos?

UNIVERSIDAD ICESI

APRENDIZAJE AUTOMÁTICO

Aprendizaje supervisado

- Aprender a partir de un "experto"
- Datos de entrenamiento **etiquetados** con una clase o valor:

$$(x1, x2, \ldots, xn, y)$$

Predictores, variables de entrada (independientes)

Respuesta, variable de salida (dependiente)

- **Meta**: predecir una clase o valor

Aprendizaje no supervisado

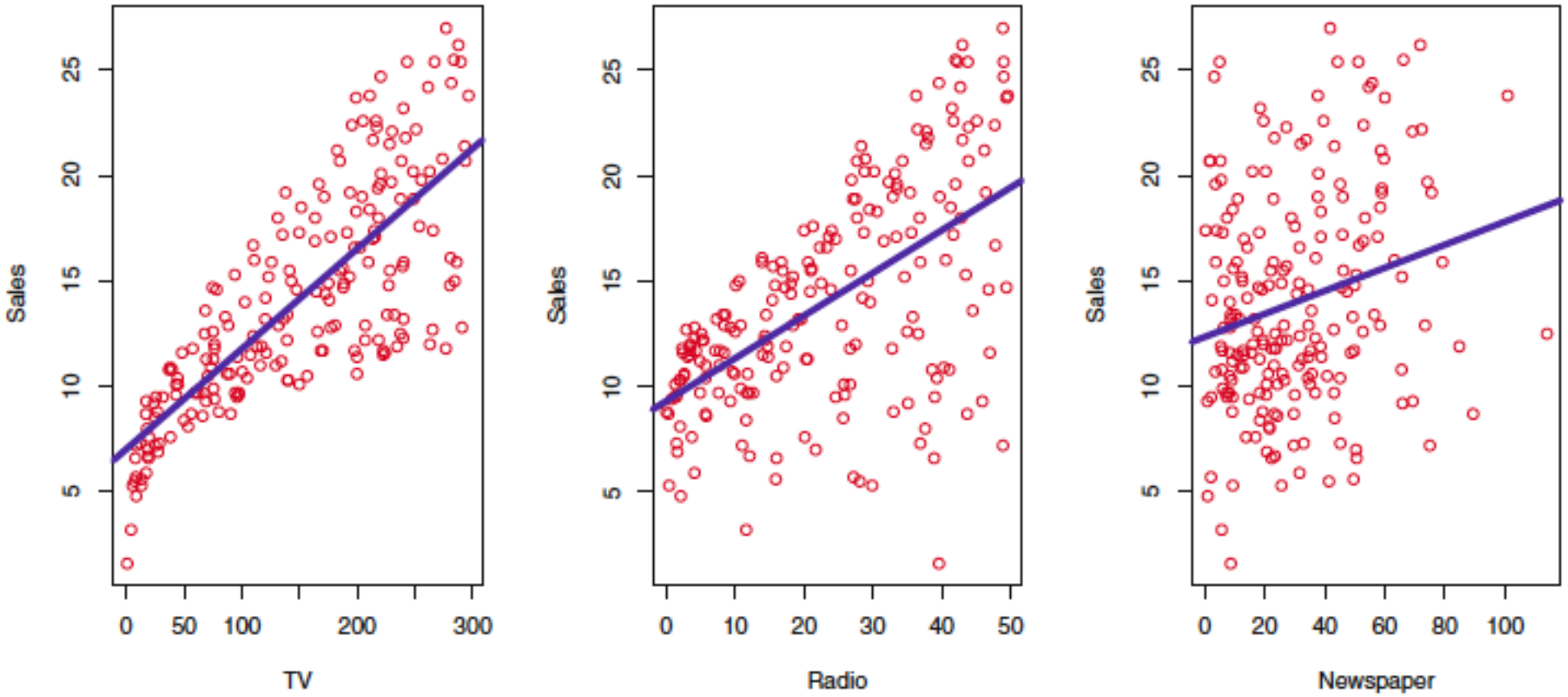
- Sin conocimiento de una clase o valor objetivo
- Los datos **no** están **etiquetados**

$$(x1, x2, \ldots, xn)$$

- **Meta**: descubrir patrones, estructura, factores no observados o una representación mas simple

APRENDIZAJE AUTOMÁTICO

Presupuestos de Publicidad



UNI ICESI

APRENDIZAJE AUTOMÁTICO

Aprendizaje supervisado

Edad	Ingresos	Tiene carro?
24	1'200.000	NO
23	4'500.000	SI
45	1'250.000	SI
32	1'100.000	NO

Datos **etiquetados**: "Respuestas correctas" disponibles

Factores/atributos/variables independientes, predictores, explicativos

Dependiente, objetivo, respuesta, salida

34	3'500.000	?

¿Cuál es el valor predicho para una instancia dada?

Aprendizaje no supervisado

Edad	Ingresos
24	1'200.000
23	4'500.000
45	1'250.000
32	1'100.000

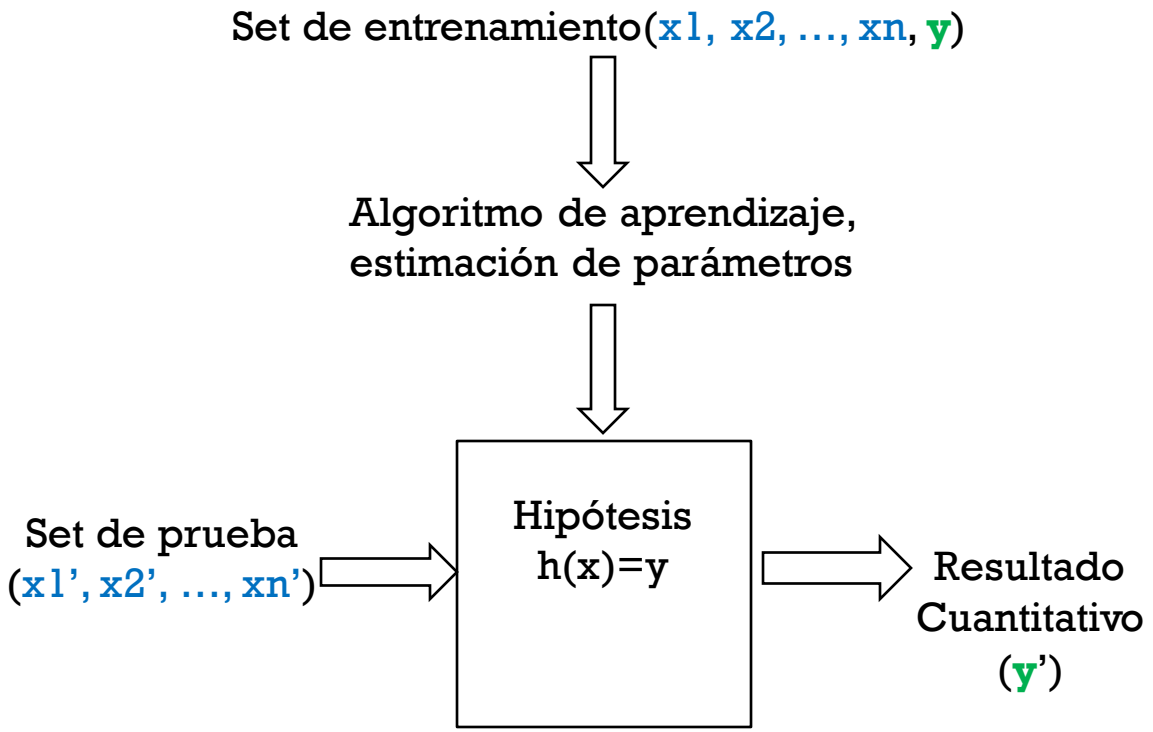
Factores/atributos/variables

¿Se puede encontrar alguna estructura en los datos?

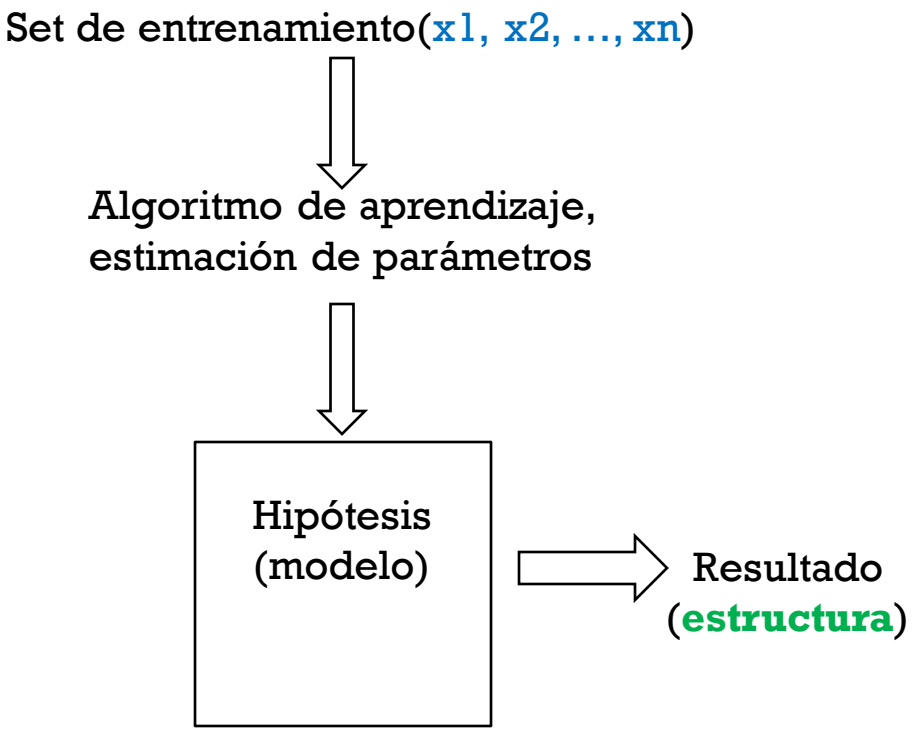
UNIVERSIDAD ICESI

APRENDIZAJE AUTOMÁTICO

Aprendizaje supervisado

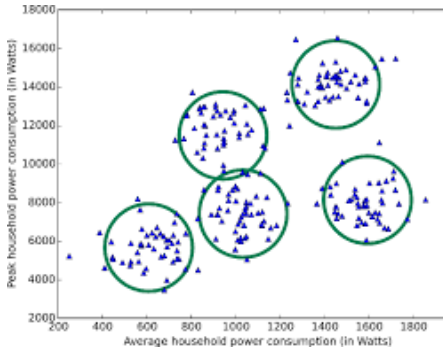


Aprendizaje no supervisado

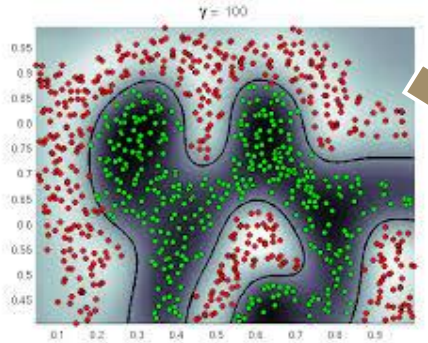


UNIVERSIDAD ICESI

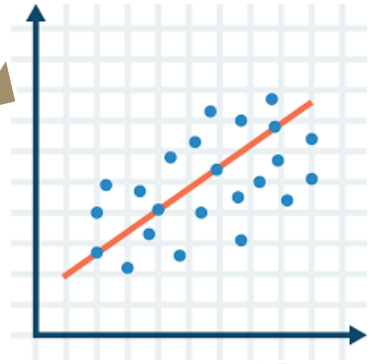
Aprendizaje automático



Aprendizaje no supervisado



Aprendizaje supervisado

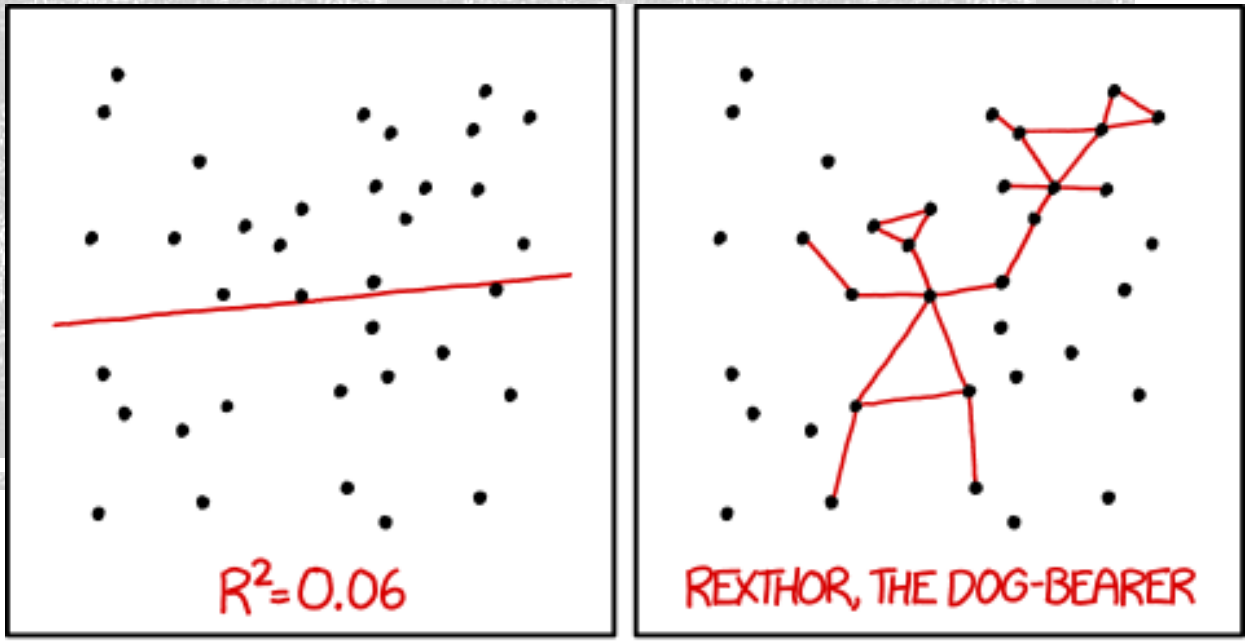


Regresión

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

REGRESIÓN

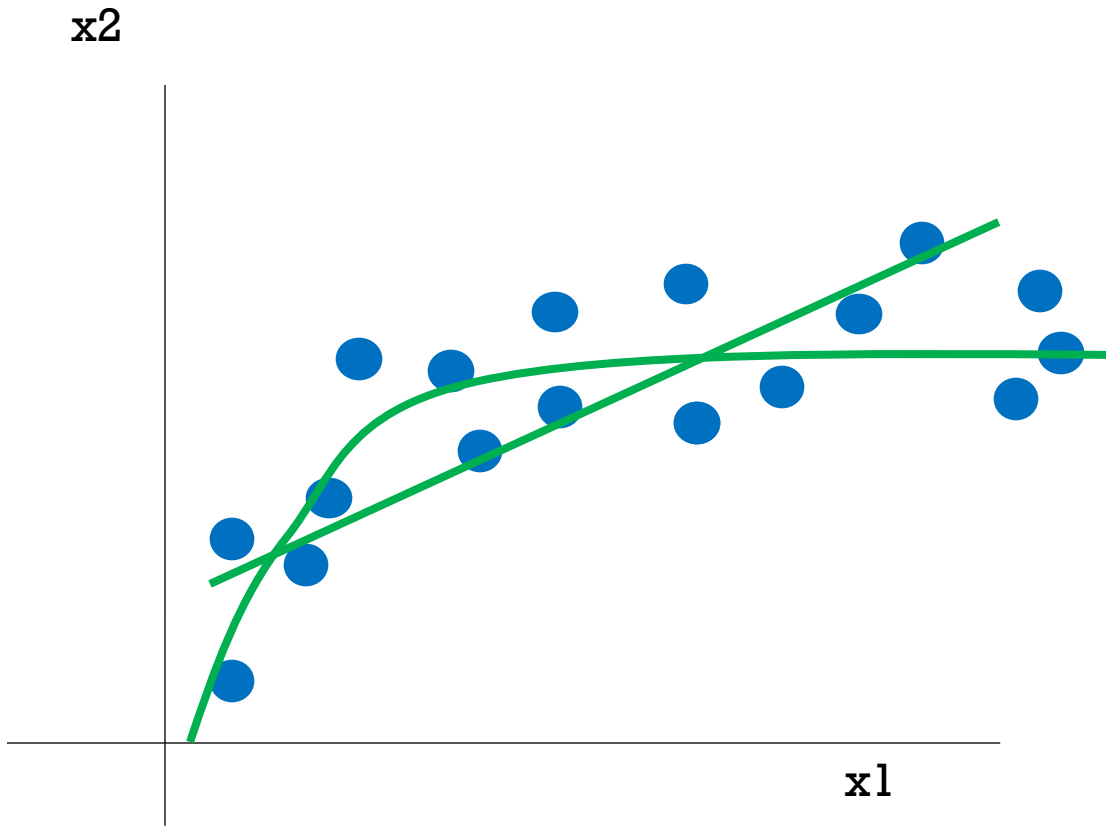


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

UNIVERSIDAD ICESI

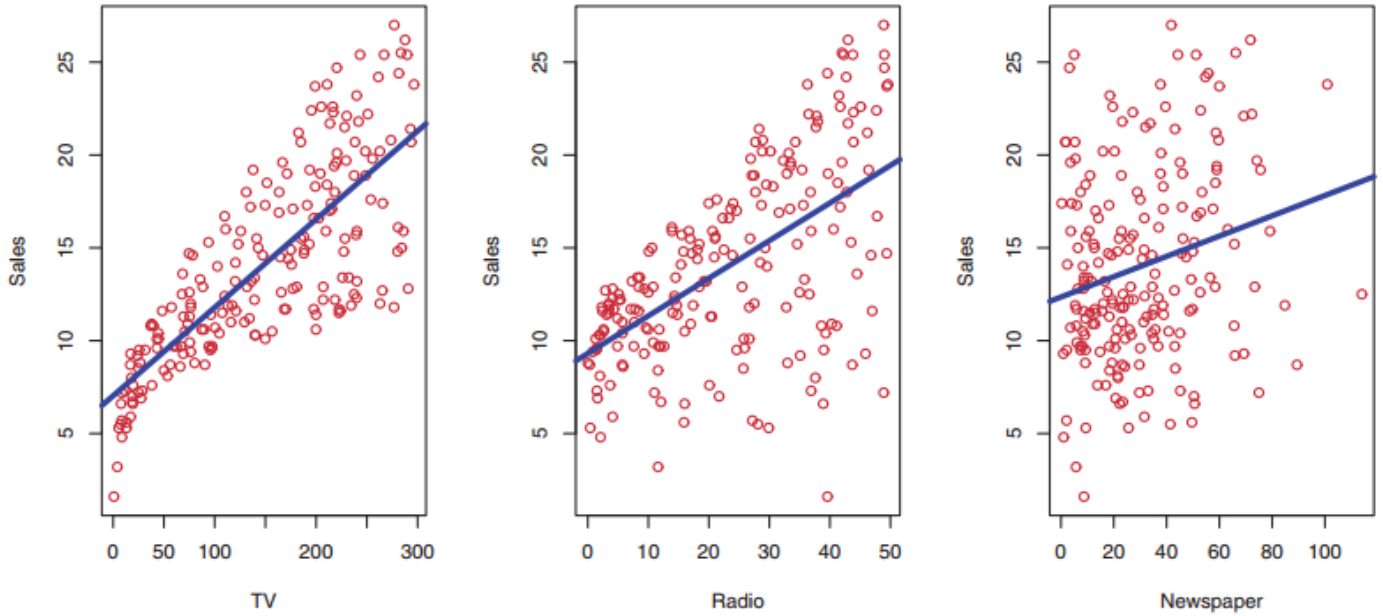
REGRESIÓN

- Encontrar modelos, f, que permitan predecir valores continuos:
  - KNN
  - Regresión lineal
  - Regresión polinómica
  - Árboles de regresión
  - …
- Valores **numéricos** de la variable o función objetivo
- **Baseline**: medida de evaluación dada por un modelo que predice una medida de tendencia central (e.g. el promedio)



REGRESIÓN

- **Predicción**:
  - Procesos de caja negra
  - Estimar el valor objetivo Y dado los valores de los predictores X
- **Inferencia**:
  - ¿Cuáles son los predictores asociados con la respuesta?
  - ¿Cuál es la relación entre la variable respuesta y cada uno de los predictores?
  - ¿Se puede considerar esa relación lineal o se trata de una relación más compleja?

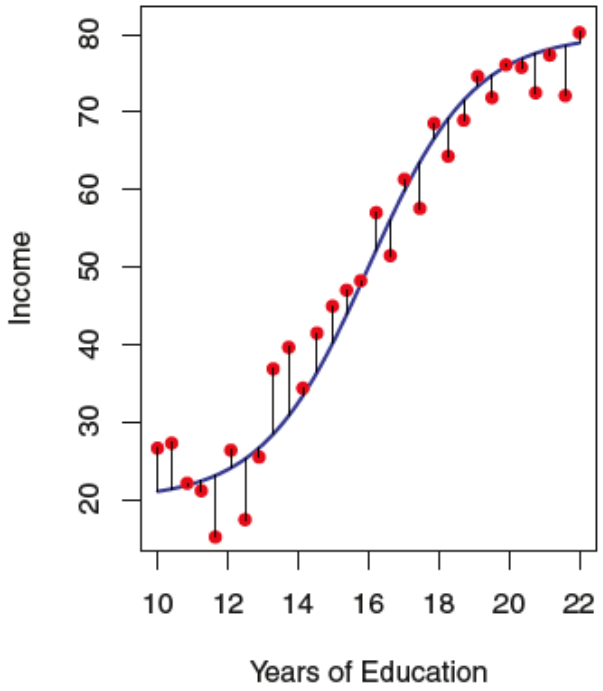


$$Ventas = f(TV, Radio, Periódicos)$$

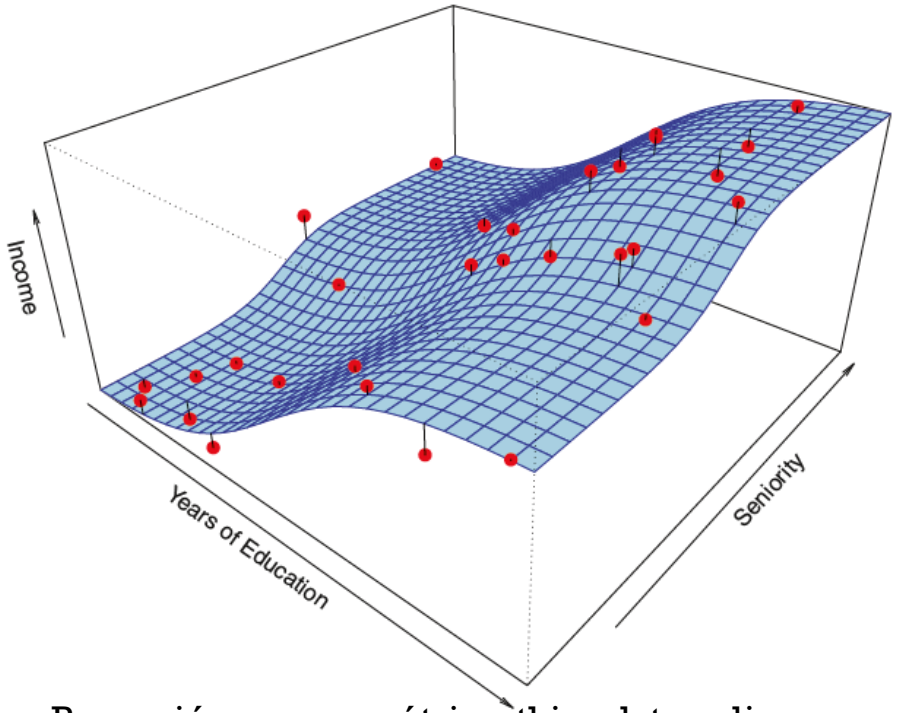
ISLR, 2013

RESIDUOS

Residuos: diferencia entre los valores reales y los valores predichos
RSS (Residual Sum of Squares): suma de los residuos al cuadrado



Regresión polinómica



Regresión no paramétrica: thin-plate spline

ISLR, 2013

UNIVERSIDAD ICESI

MÉTRICAS DE REGRESIÓN

Coeficiente de correlación (Pearson: [-1;1]): indica la fuerza de la relación lineal entre los predictores y la variable objetivo, que puede ser positiva o negativa

- | rho | = 0 no hay correlación
- | rho| = 0.10 correlación muy débil
- | rho | = 0.25 correlación débil
- | rho | = 0.50 correlación media
- | rho | = 0.75 correlación fuerte
- | rho | = 0.90 correlación muy fuerte
- | rho | = 1 correlación perfecta

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Coeficiente de determinación (R^2): indica el porcentaje de la varianza que pudo ser explicada por los predictores a partir de la relación lineal

UNIVERSIDAD ICESI

MÉTRICAS DE REGRESIÓN

- MAE (mean absolute error):

$$\frac{1}{m}\sum_{1}^{m}|h_\theta(x_i)-y_i|$$

- MSE (mean square error):

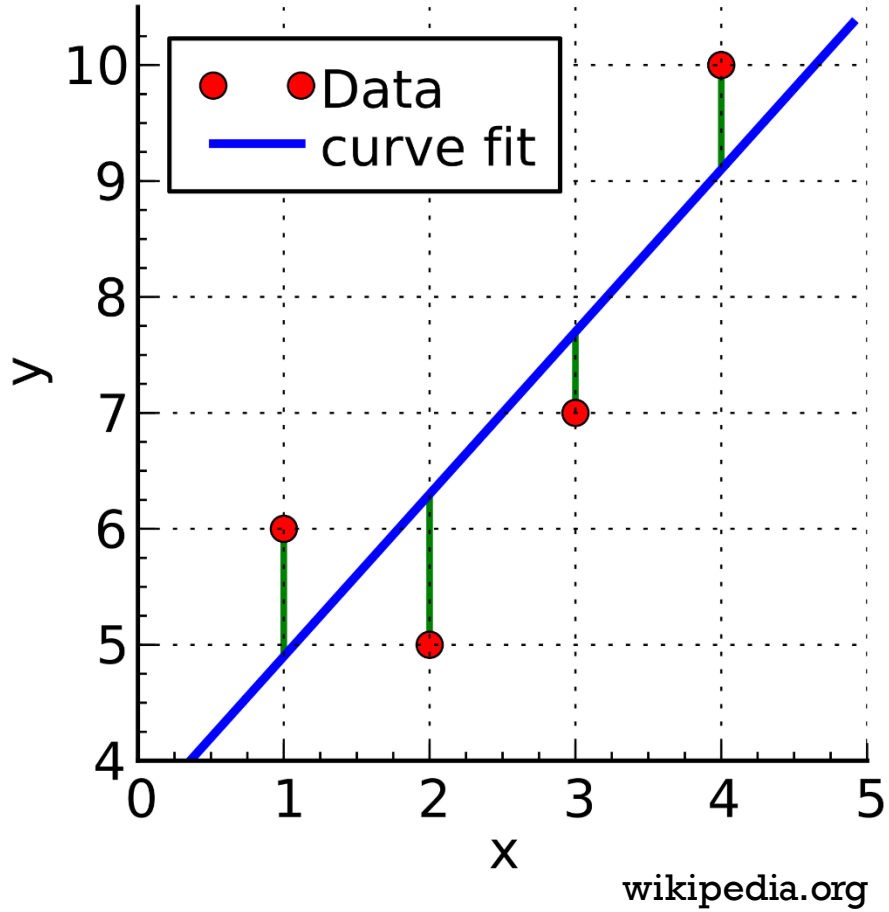
$$\frac{1}{m}\sum_{1}^{m}(h_\theta(x_i)-y_i)^2$$

- RMSE (root mean square error):

$$\sqrt{\frac{1}{m}\sum_{1}^{m}(h_\theta(x_i)-y_i)^2}$$

- R^2 (coeficiente de determinación):

$$1-\frac{\sum_1^m(h_\theta(x_i)-y_i)^2}{\sum_1^m(y_i-\bar{y})^2}$$

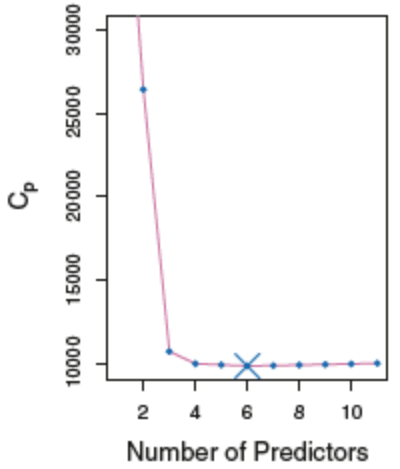


UNIVERSIDAD ICESI

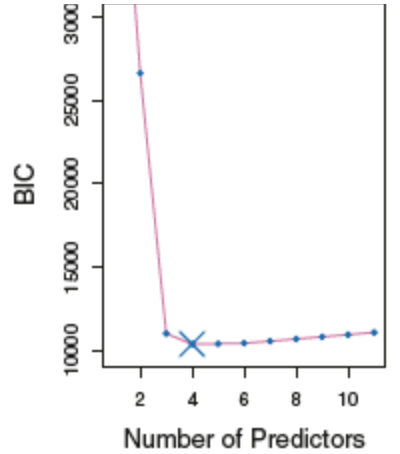
MÉTRICAS DE REGRESIÓN

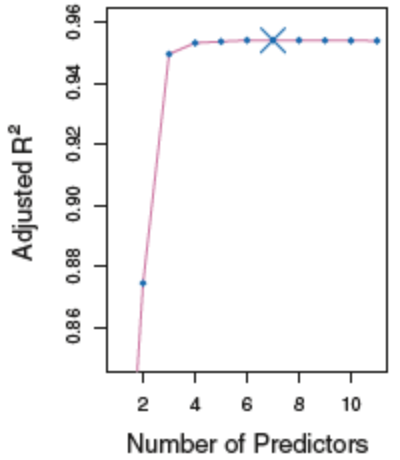
- Se puede demostrar que el error de las métricas basadas en el RSS de un modelo, es decreciente con respecto al número de variables predictivas ***d*** que considera.
- No se puede considerar MSE, RMSE, R^2 como métricas de evaluación para comparar modelos con un número de predictores diferentes → se deben ajustar con una penalización de la cardinalidad

  - Mallow's Cp (minimizar) $$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$
  - AIC (Akaike Information Criteria) (minimizar) $$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$
  - BIC (Bayesian Information Criteria) (minimizar) $$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$$
  - R^2 ajustado (maximizar) $$R^2_{adj} = 1 - \frac{\sum_1^m (h_\theta(x_i) - y_i)^2/(n-d-1)}{\sum_1^m (y_i - \bar{y})^2/(n-1)}$$



ISLR, 2013

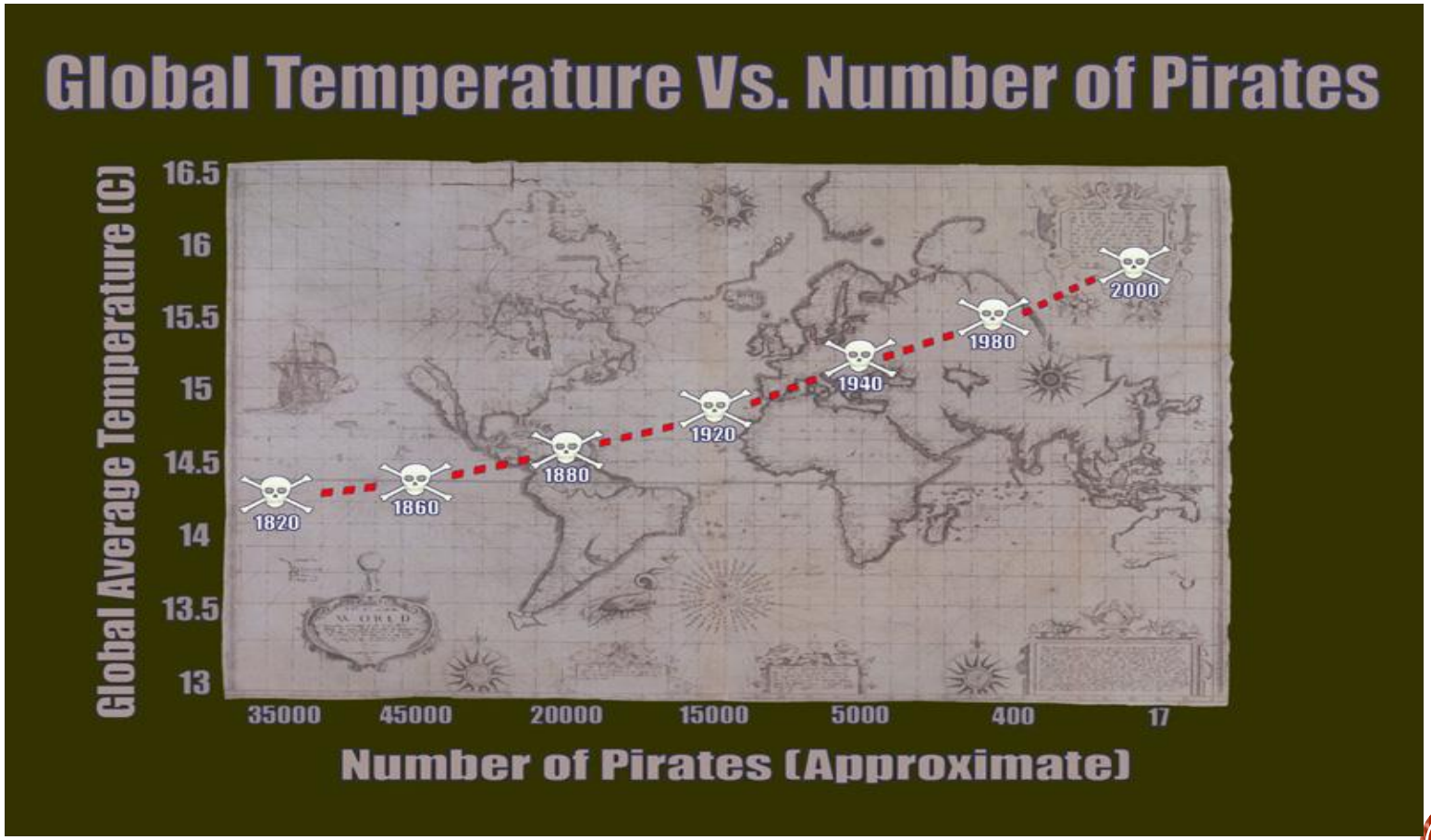




UNIVERSIDAD ICESI

REGRESIÓN – CUIDADO!

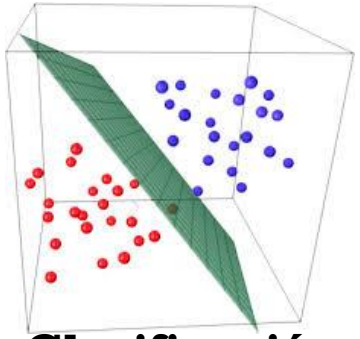
Correlación y causalidad son dos cosas muy diferentes



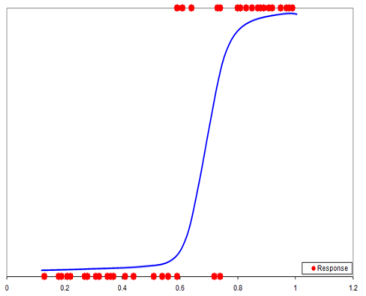
Erika Andersen, Forbes, 2012

UNIVERSIDAD ICESI

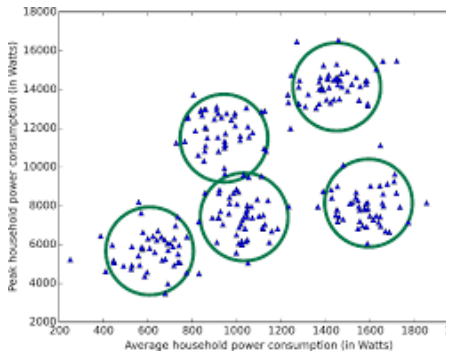
Aprendizaje automático



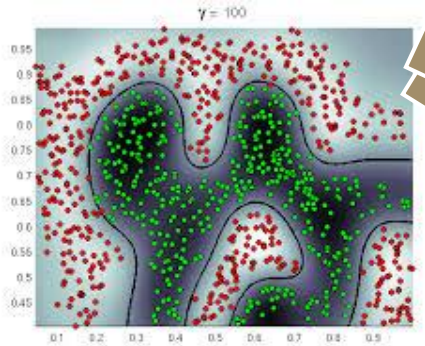
Clasificación



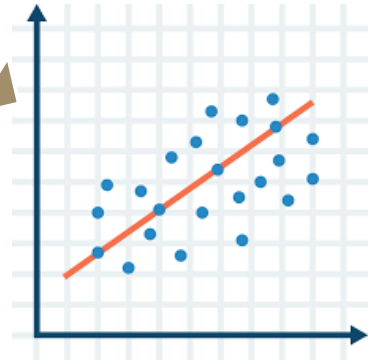
Regresión logística



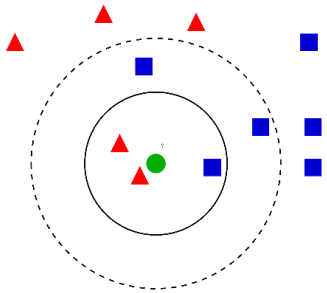
Aprendizaje no supervisado



Aprendizaje supervisado



Regresión

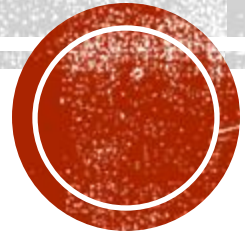


KNN

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

CLASIFICACIÓN: REGRESIÓN LOGÍSTICA, KNN



UNIVERSIDAD ICESI

CLASIFICACIÓN

Vamos a enviar un folleto costoso de promoción a nuestra base de clientes, pero no podemos enviarlo a todos los clientes.

Tenemos los datos históricos de los clientes que en el pasado han comprado el producto en cuestión.

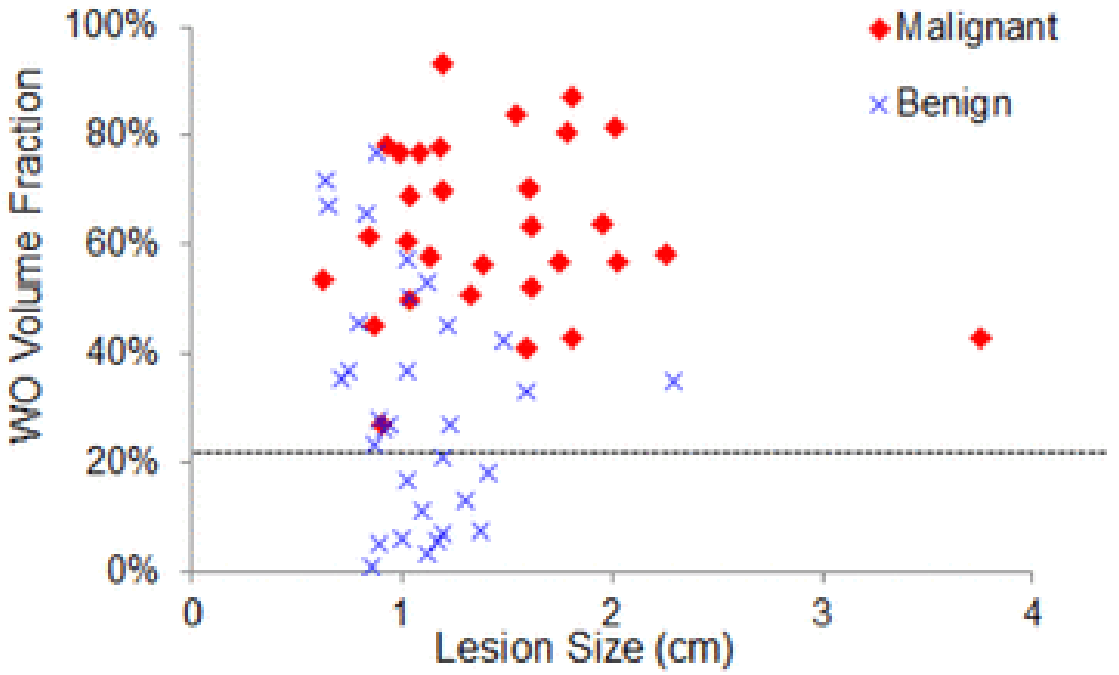
Según nuestra base de datos, le enviamos un folleto a:

- ¿Un hombre de 33 años y estrato 5?
- ¿Una mujer de 42 años y estrato 3?
- ¿Una mujer de 28 años y estrato 3?

Género	Edad	Estrato	CLASE
Hombre	52	4	COMPRA
Mujer	28	3	No compra
Hombre	33	5	No compra
Mujer	26	5	COMPRA
Mujer	35	4	No compra
Hombre	51	3	COMPRA
Mujer	28	3	COMPRA

CLASIFICACIÓN

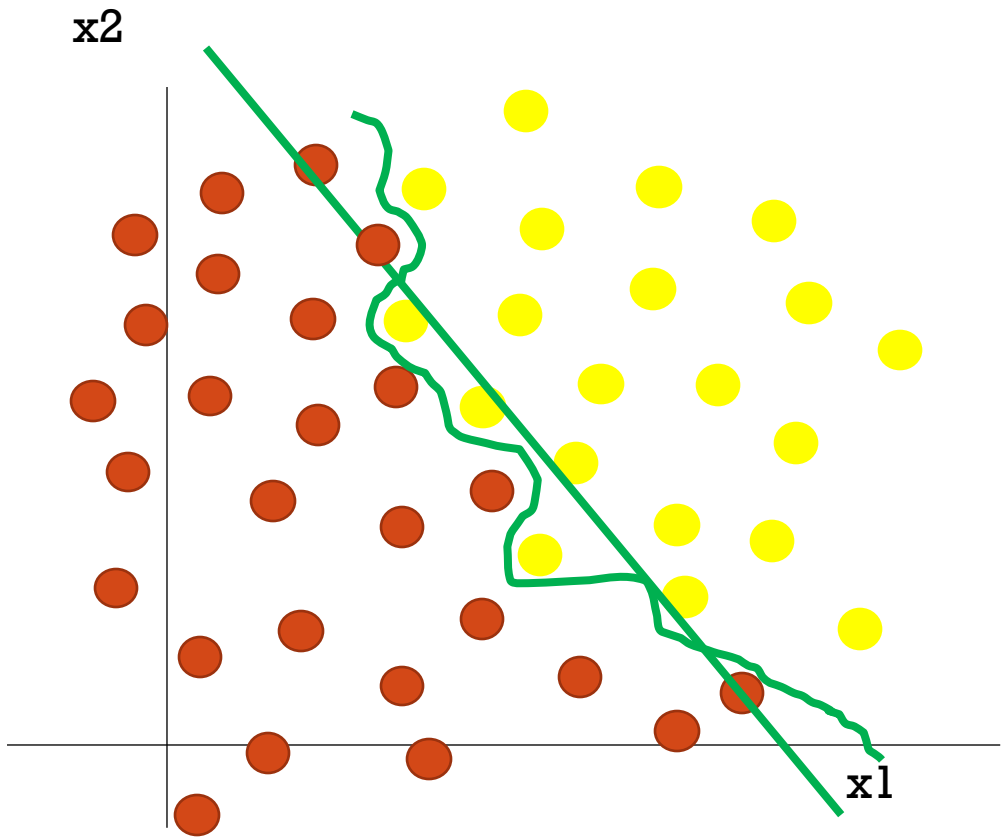
- Encontrar modelos que describan clases para futuras predicciones:
  - KNN
  - Árboles de decisión
  - Regresión logística
  - Redes neuronales
  - …
- Valores **discretos** de la variable objetivo (categóricos)
- Incluye modelos que no solo clasifican sino que estiman las **probabilidades** de cada clase



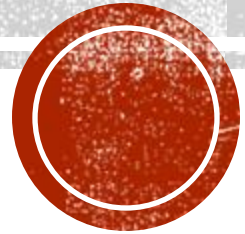
http://www.jacmp.org/index.php/jacmp/article/view/5187/html_374

CLASIFICACIÓN

- **Baseline**: (modelo **nulo**) medida de evaluación dada por un clasificador que escoge siempre la clase mayoritaria.
- **Modelo de Bayes**: (modelo **saturado**) el mejor modelo posible con los datos disponibles (modelo generador de los datos). Límite superior de comparación. En general no se conoce.



REGRESIÓN LOGÍSTICA

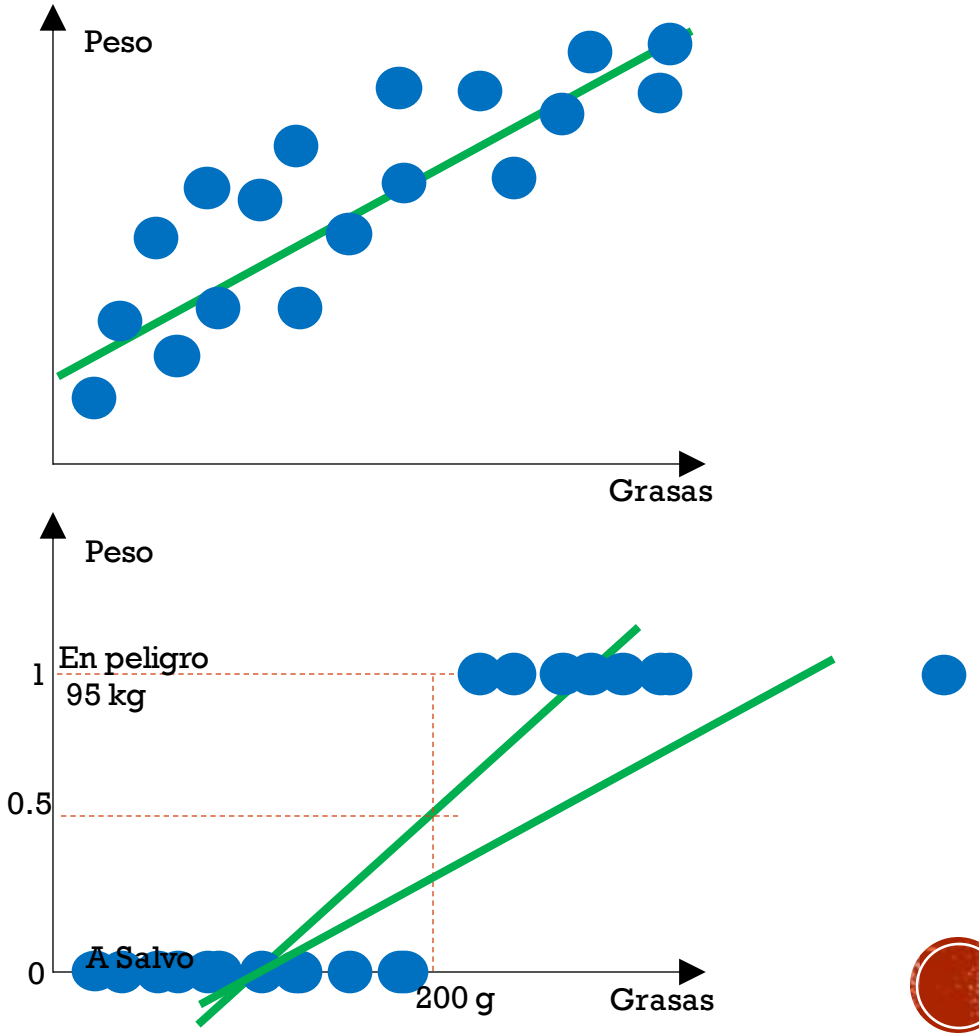


UNIVERSIDAD ICESI

¿REGRESIÓN LINEAL PARA CLASIFICACIÓN?

Ejemplo: tenemos los datos que relacionan la cantidad de grasas consumidas y el peso de las personas → Regresión

- Si un doctor estima que mas de 95kg implica riesgo de diabetes, el problema se convierte en uno de clasificación: 0=a salvo, 1=en peligro
- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes
- No se puede interpretar estas predicciones como probabilidades (valores no están en [0;1])
- Poco robusto.



UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

- Algoritmo de **clasificación**, no de **regresión**
- Parte de la idea de la regresión lineal, cuyo resultado es modificado para poder obtener una salida **binaria**: sólo permite distinguir entre 2 clases.
  - Churn vs. Stay
  - Compra vs. No compra
  - Cliente valioso vs. Cliente no valioso
- Se agrega una transformación del resultado de la regresión lineal a partir de una función de distribución acumulativa logística, también conocida como función **logit** o **sigmoide**.

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

- El modelo pasa de:

$$h_\Theta(\boldsymbol{X}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

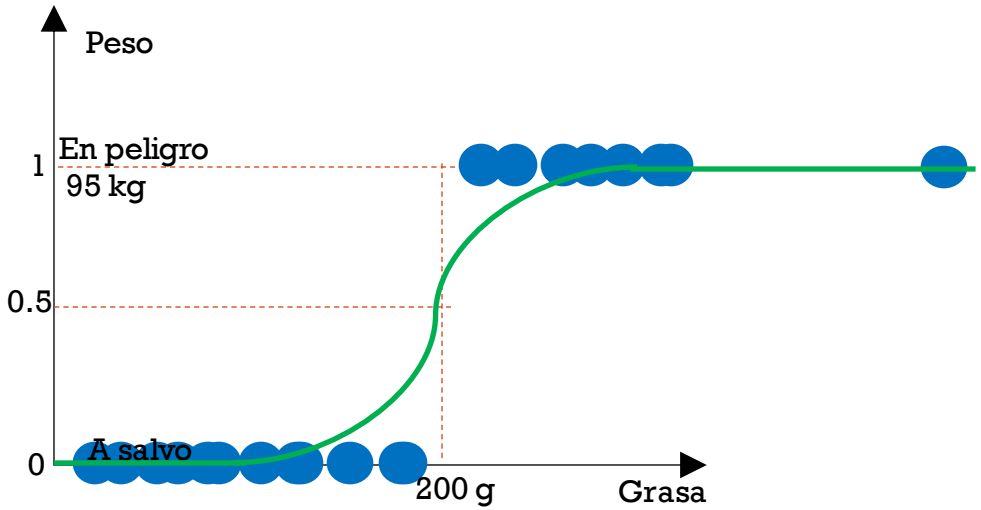
a $$h_\Theta(\boldsymbol{X}) = \boldsymbol{f(z) = \sigma}(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n),$$

con max(f(z))=1 y min(f(z))=0

- $\boldsymbol{\sigma(z)}$ es la función **sigmoide** o **logística**
- Se pueden interpretar los valores de $\boldsymbol{\sigma(z)}$ como **probabilidades** de que una instancia con atributos $\boldsymbol{X}$ pertenezca a la clase Y=1:

$$P(\boldsymbol{Y = 1}|x_1, \ldots, x_n) = p_1(\boldsymbol{X}) = \boldsymbol{\sigma}(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)$$

$$p_1(\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{\Theta^T X}}}$$

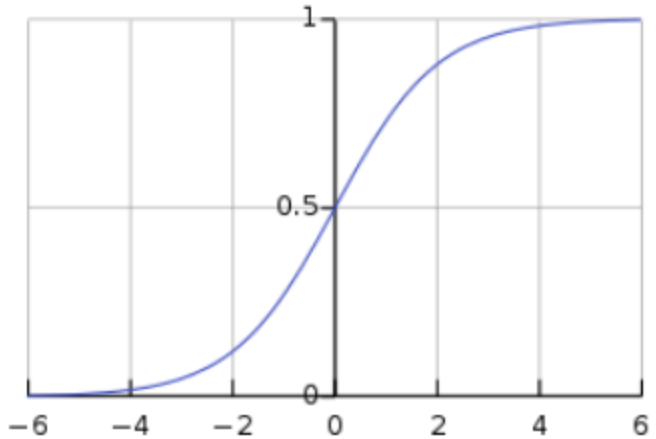


UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

- Comportamiento:
  - Si y=1, queremos que $p_1(\boldsymbol{X}) \approx \mathbf{1}$, luego $\boldsymbol{\Theta}^T \boldsymbol{X} \gg \mathbf{0}$
  - Si y=0, queremos que $p_1(\boldsymbol{X}) \approx \mathbf{0}$, luego $\boldsymbol{\Theta}^T \boldsymbol{X} \ll \mathbf{0}$
- Predicción: se establece un valor de umbral, por ejemplo 0.5
  - Predecir clase 1 si $p_1(\boldsymbol{X}) \geq 0.5$, cuando $\boldsymbol{\Theta}^T \boldsymbol{X} \geq 0$
  - Predecir clase 0 de otra manera
- Se puede establecer un umbral diferente si se quiere ser mas o menos robusto en la clasificación

$$p_1(\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{\Theta}^T \boldsymbol{X}}}$$



Wikipedia, 2019

UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

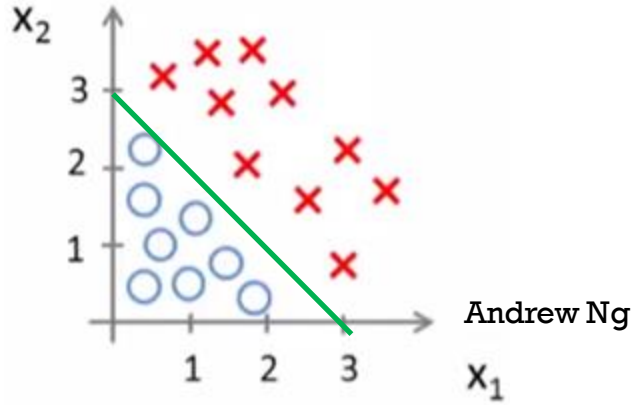
- Coeficientes θ_i:
  - $\log\left(\frac{p_1(X)}{1-p_1(X)}\right) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$
  - Relación **lineal** entre los coeficientes y el logaritmo de la razón de probabilidades (odds ratio)
  - Un crecimiento de una unidad de x_1 indica que el log de la razón de probabilidades va a crecer θ_1 unidades.
  - El **signo** indica la dirección de la influencia.
  - Difícil de interpretar logs, se exponencian los coeficientes $\exp(\theta_i)$, por 1 unidad de $\exp(\theta_i)$, los odds de y=1 aumentan $\exp(\theta_i)$ veces; por 10 unidades de $\exp(\theta_i)$, los odds de y=1 aumentan $\exp(\theta_i)^{10}$ veces.
  - Análisis de sensibilidad de p(y=1) con respecto a una variable, fijando las otras en sus valores promedios.
  - Prueba de hipótesis para evaluar la **significancia** de cada coeficiente (diferencia de 0).

- Razón de probabilidades
  - A probabilidades altas, razón alta y viceversa

$p_1(X)$	odds
1,0	+Inf
0,99	99
0,75	3
0,5	1
0,25	0,33
0	0

REGRESIÓN LOGÍSTICA

- El algoritmo de regresión logística determina una frontera de decisión lineal



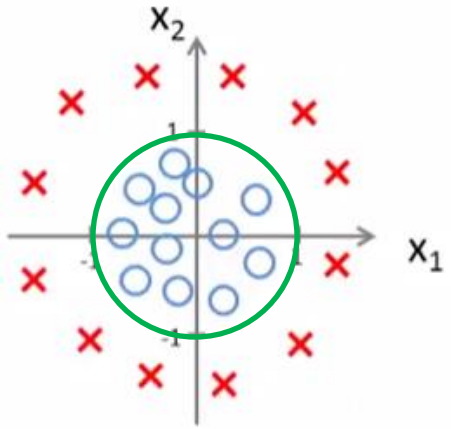
Andrew Ng

$$h_\Theta(\boldsymbol{X}) = f(-3 + x_1 + x_2)$$

Predecir la clase roja de cruz cuando:

- $h_\Theta(\boldsymbol{X}) \geq 0.5$
- $f(-3 + x_1 + x_2) \geq 0{,}5$

- Para fronteras de decisión no lineales: usar polinomios de un mayor orden



$$h_\Theta(\boldsymbol{X}) = f(-1 + {x_1}^2 + {x_2}^2)$$

Predecir la clase roja de cruz cuando:

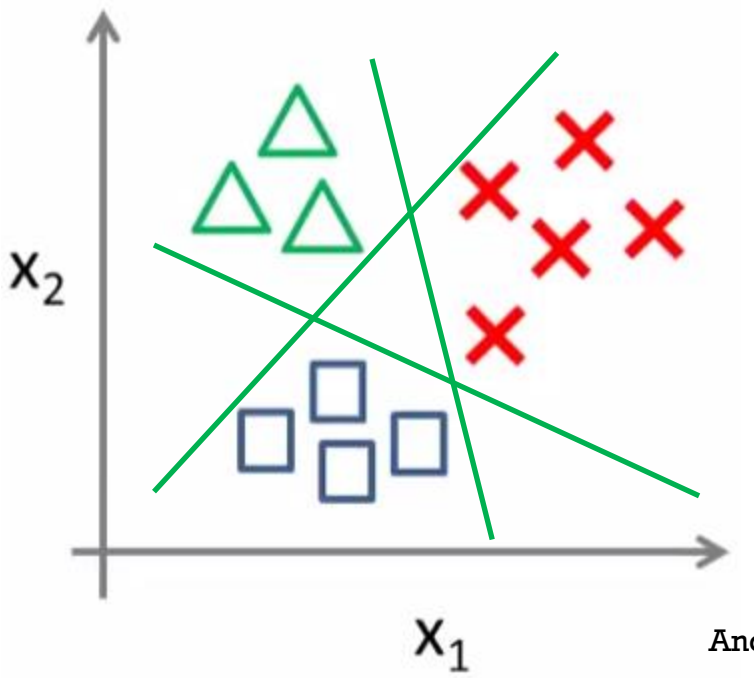
- $h_\Theta(\boldsymbol{X}) \geq 0.5$
- $f(-1 + {x_1}^2 + {x_2}^2) \geq 0{,}5$

UNIVERSIDAD ICESI

REGRESIÓN LOGÍSTICA

¿Qué se puede hacer si se tienen más de 2 clases?

- Para problemas de clasificación con más de 2 clases, es necesario utilizar una aproximación de **1 vs. todos**
- Un clasificador por regresión logística es necesario para cada clase
- Para una nueva instancia, la clase con la mayor probabilidad en su propio modelo es predicha



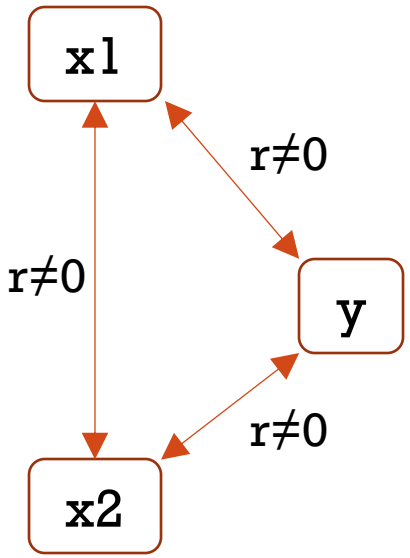
Andrew Ng

→ También se puede hacer regresión logística **multinomial** con la función **softmax**

REGRESIÓN LOGÍSTICA

Confounding

- Problema que ocurre cuando un modelo de regresión no considera variables independientes relevantes (**confounding variables**)
- Posibles efectos en la relación entre variables independientes y dependiente:
  - Sobrestimar / Subestimar la fortaleza de una relación
  - Cambiar la dirección de una relación
  - Esconder un efecto que en realidad existe
- Causas
  - La variable no incluida debe estar correlacionada con la variable dependiente
  - La variable no incluida debe estar correlacionada con al menos una variables independiente del modelo



REGRESIÓN LOGÍSTICA

- Consideraciones
  - Produce estimación de probabilidades
  - No hay parámetros a afinar, solo las variables independientes a considerar.
  - Permite variables independientes numéricas y categóricas
  - Estimación de parámetros eficiente computacionalmente
  - No se ve afectado por situaciones de multicolinealidad leves. Casos importantes se pueden resolver con una regularización L2.
  - Se puede utilizar descenso de gradiente para encontrar los parámetros (mismas ecuaciones de actualización de parámetros que para regresión lineal, cambiando la función de predicción)
  - No es ideal en casos de muchas variables categóricas
  - No es muy flexible (lineal) aunque se puede extender polinómicamente.

OTROS MODELOS PARA CLASIFICACIÓN, BASADOS EN LA IDEA DE REGRESIÓN

- Multinomial logistic regression: regresión con mas de 2 categorías (*mlogit()* de mlogit)
- Robust logistic regression: a prueba de outliers y observaciones influenciadoras (*glmRob()* de robust)
- Ordinal logistic regression: categorías ordenadas (*lrm()* de rms)
- Generalized lineal models (*glm()* de stats): relaja la normalidad de los residuos generalizando de la familia lineal a la familia exponencial de modelos, permite varios tipos de regresiones (lineal, binomial, poisson, gamma, ..)
- Generalized additive models (GAMs) y vector GAMs:
  - Generalización de los modelos lineales que permite remplazar los predictores por funciones suavizadas no lineales de los mismos (splines, polinomios, funciones de saltos)
  - La forma de las funciones no lineales no se especifica, es determinada por los datos
  - Gran capacidad predictiva y de interpolación, resistente al overfitting
  - *gam()* de gam

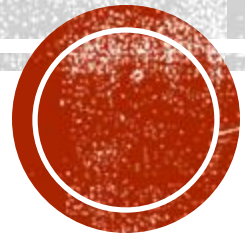
CLASIFICACIÓN: REGRESIÓN LOGÍSTICA

- 03-LogReg-Ejemplo
  - Cuaderno de regresión logística desde cero
  - Uso del método LogisticRegression de sklearn
  - Uso del paquete statsmodels para calcular los coeficientes y la significancia de las variables predictoras.

CLASIFICACIÓN: REGRESIÓN LOGÍSTICA

- Desarrollar la parte de regresión logística, que se encuentra especificada en el documento 03-SAHeartDisease-LogReg+KNN.html

KNN:
K-NEAREST NEIGHBORS



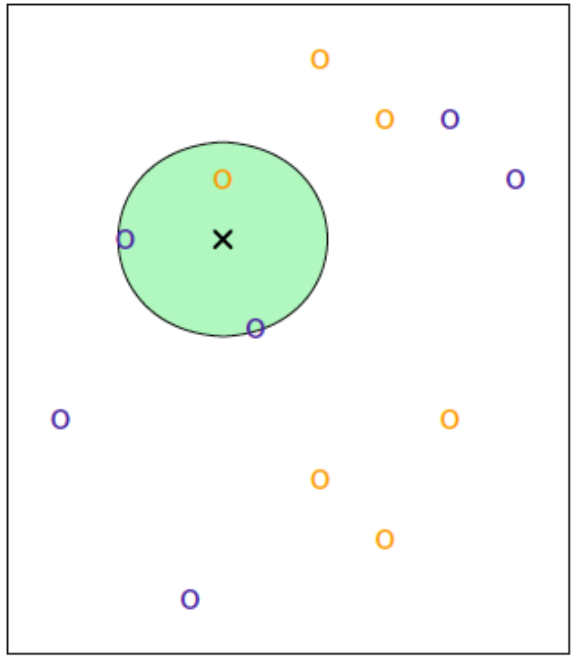
Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

OBJETIVOS DE APRENDIZAJE

- Identificar problemáticas que se puedan resolver a partir de modelos de aprendizaje supervisado, ya sea de clasificación o de regresión.
- Aplicar modelos de K-NN y de regresión logística a conjuntos de datos, para responder a preguntas de negocio involucrando modelos de aprendizaje supervisado utilizando el lenguaje python.
- Comparar modelos de aprendizaje supervisado con respecto a métricas de ajuste, utilizando diferentes protocolos de evaluación

KNN (K NEAREST NEIGHBORS): K VECINOS MÁS CERCANOS

- Algoritmo de aprendizaje supervisado para **clasificación** y **regresión**
- **Simple**: asignar la clase o valor agregado de las instancias conocidas que se encuentran mas cerca de la instancia a predecir
- Basado en las **instancias** de aprendizaje, no en un modelo subyacente probabilístico/estadístico
- Aprendizaje **perezoso**: en realidad el algoritmo solo se ejecuta en el momento que se requiere predecir una nueva instancia a partir de una predicción local
- Depende de la definición de una función de **distancia**, que se escogerá según la cantidad y características de las variables independientes



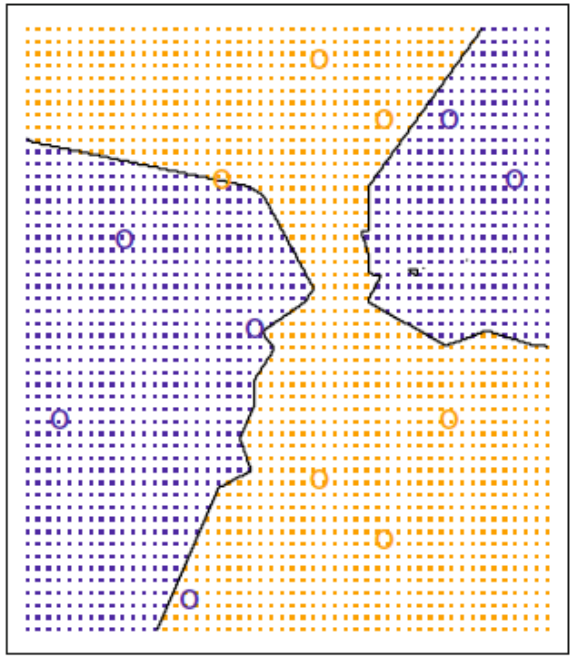
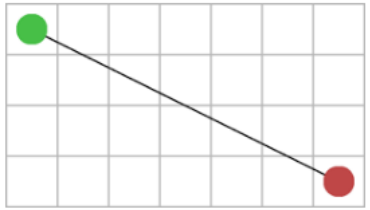


Figura 2.14 SLR

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

KNN – DISTANCIAS

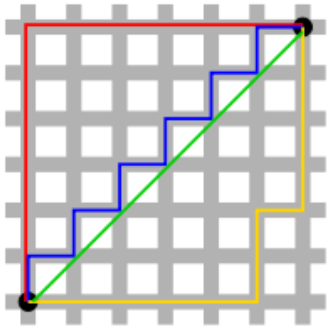
- Ejemplos de medidas de **similitud** o **distancia** utilizadas para encontrar los vecinos mas cercanos:
  - **Euclidiana**: tamaño del segmento linear que une las dos instancias comparadas.



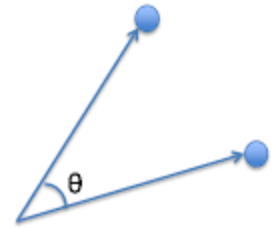
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

  - **Manhattan**: basada en una organización en bloques rectilíneos



  - **Coseno**: coseno del ángulo entre las dos instancias comparadas → Alta dimensionalidad y **big data**

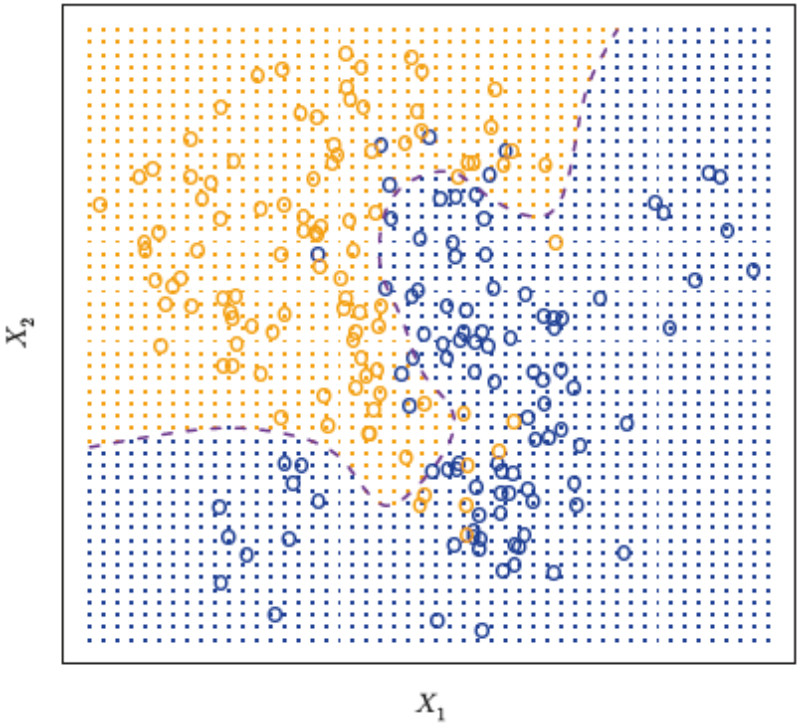


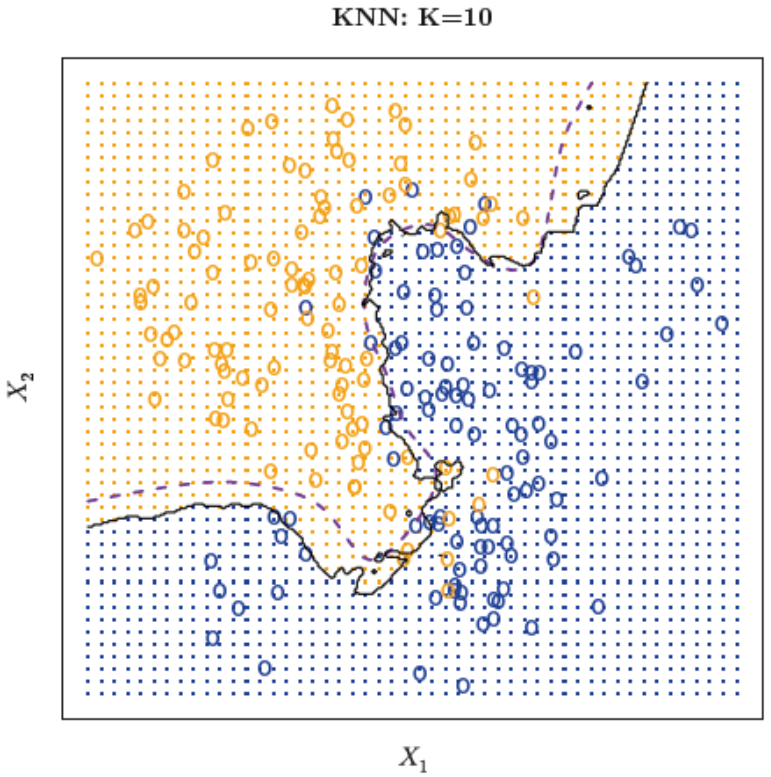
$$sim(\boldsymbol{x},\boldsymbol{y}) = \cos(\theta_{x,y}) = \frac{\boldsymbol{x}.\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\sum_i x_i * y_i}{\sqrt{\left(\sum_i x_i * x_i\right) * \sum_i y_i * y_i}}$$

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

KNN – K

- **Parámetro K**: número de vecinos mas cercanos a considerar para establecer la clase o valor de una nueva instancia

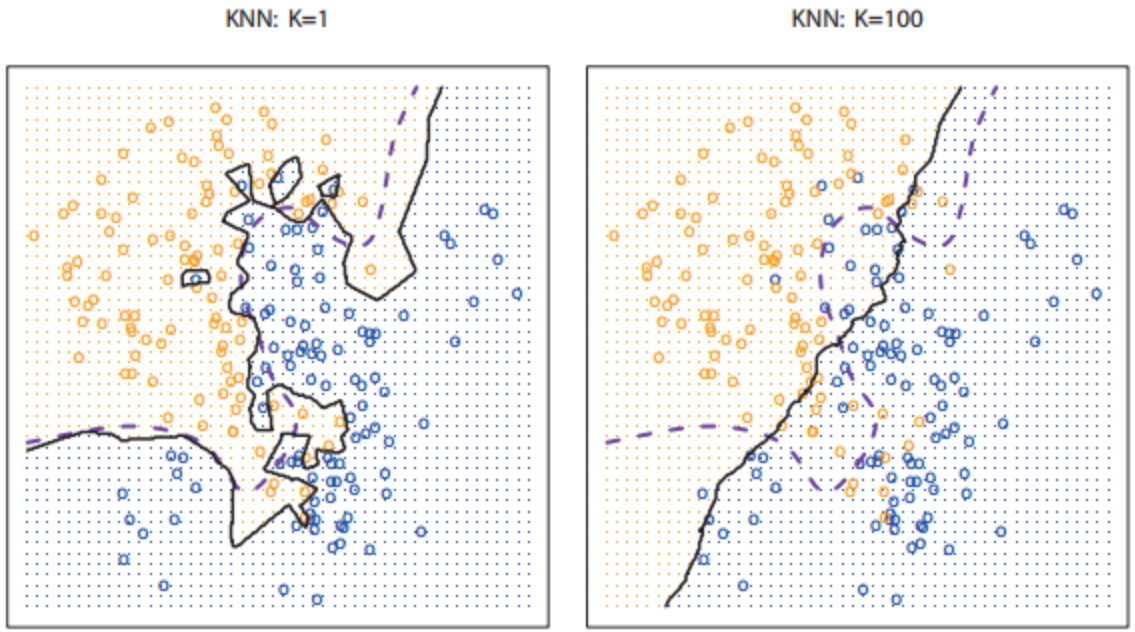




James et al, ISLR, 2013

KNN – K

- **Parámetro K**
  - El resultado puede ser drásticamente diferente para diferentes valores de K
  - Un valor de K grande suavizará los límites entre clases/valores (alto sesgo, baja varianza)
  - Un valor de K pequeño resultará en límites muy flexibles (bajo sesgo, alta varianza)
  - El valor de K óptimo se encuentra empíricamente

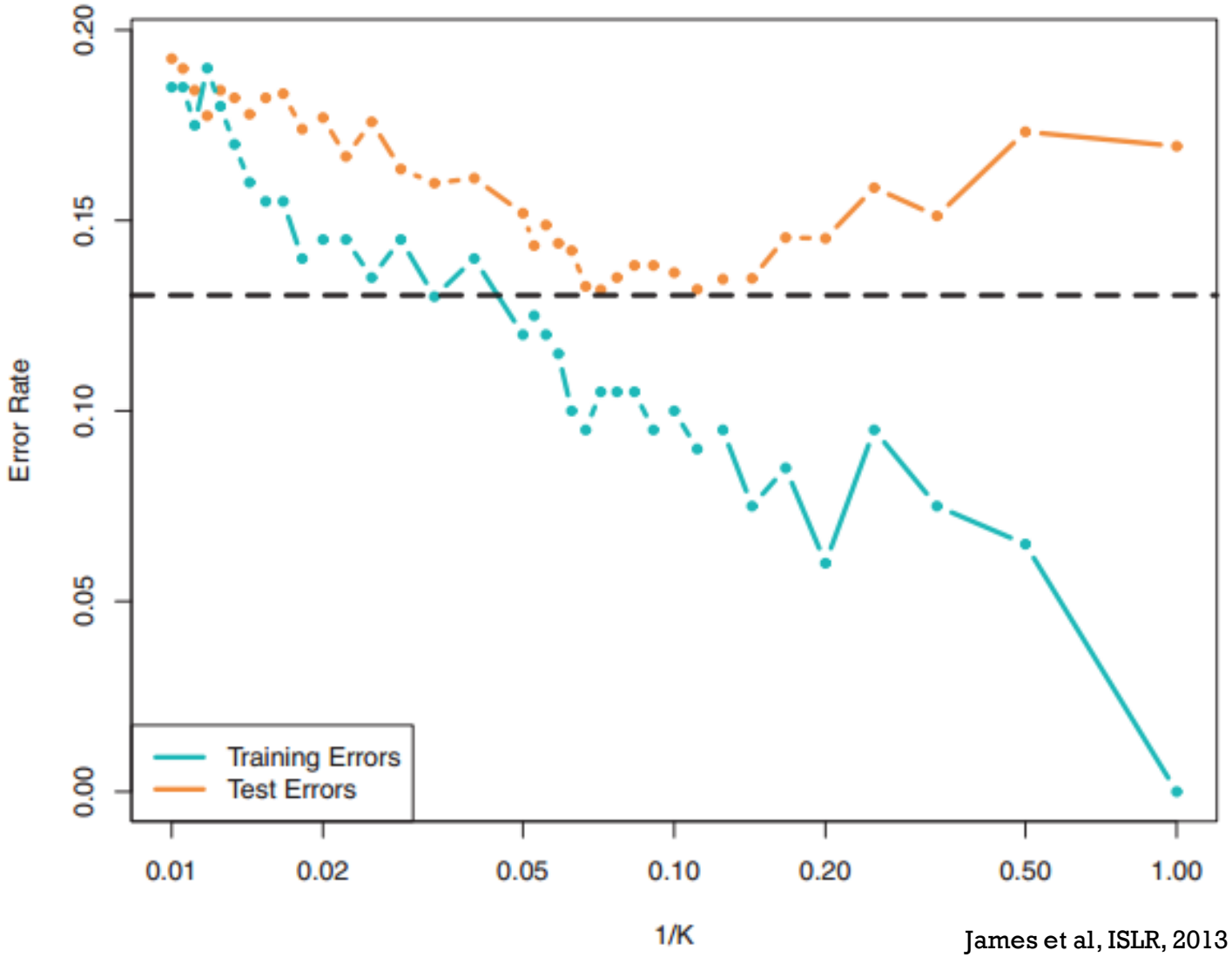


James et al, ISLR, 2013

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

KNN – K

- K controla el **overfitting** (sobre aprendizaje) y el **underfitting** (sub aprendizaje
- Modelos mas **sencillos** (K mas grandes) previenen el overfitting, pero pueden por el contrario irse hacia el underfitting
- Modelos mas **complejos** (K mas pequeños) previenen el underfitting, pero pueden por el contrario irse hacia el overfitting
- El **K ideal** que sirva para todos los casos no existe, depende de cada dataset específico



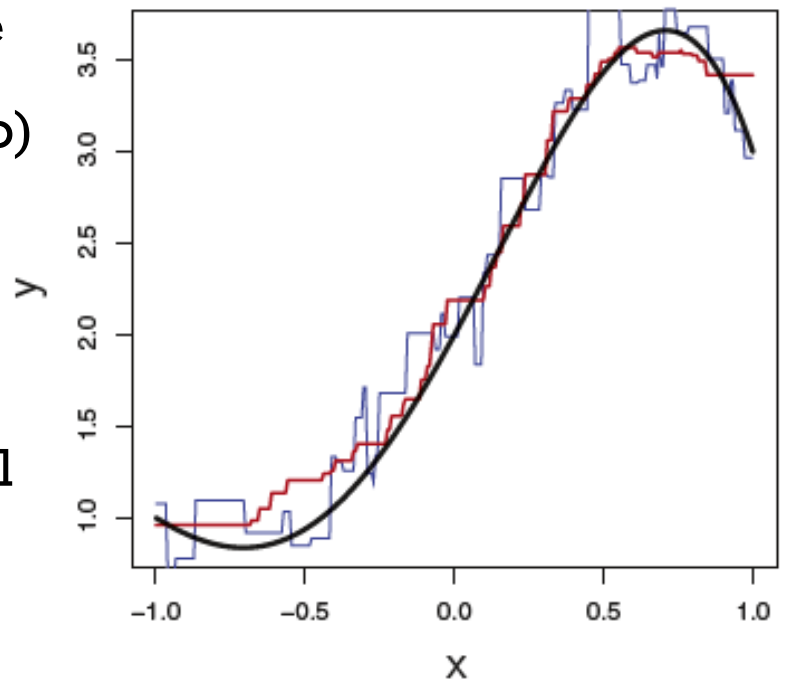
James et al, ISLR, 2013

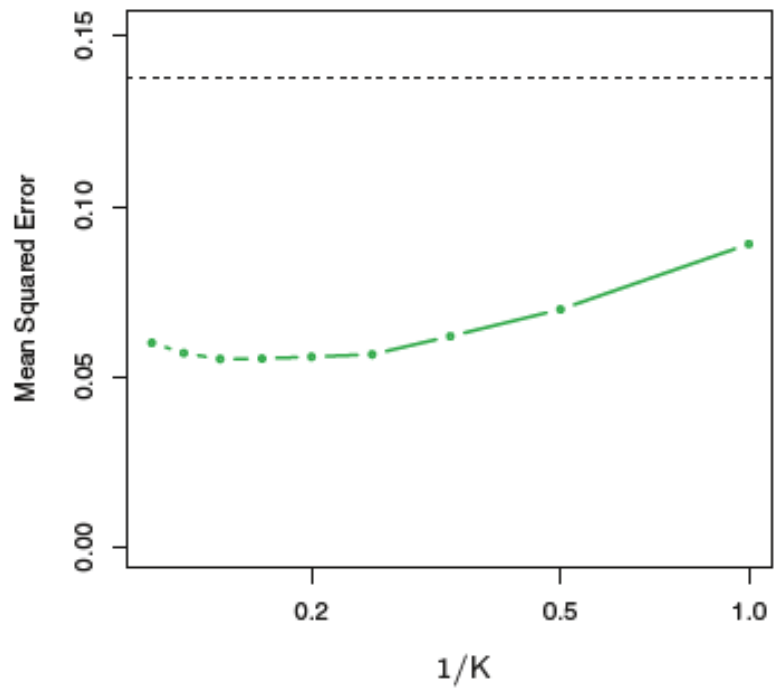
Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

KNN – K

En el caso de la utilización de KNN para la regresión las mismas consideraciones aplican

- En el panel izquierdo: se aplica KNN con un valor de K=1 (azul) y K=9 (rojo)
- En el panel derecho, se puede ver el valor de RMSE para diferentes valores de K (en verde). También se puede ver, por comparación el nivel de error de la regresión lineal simple (punteada en negro)

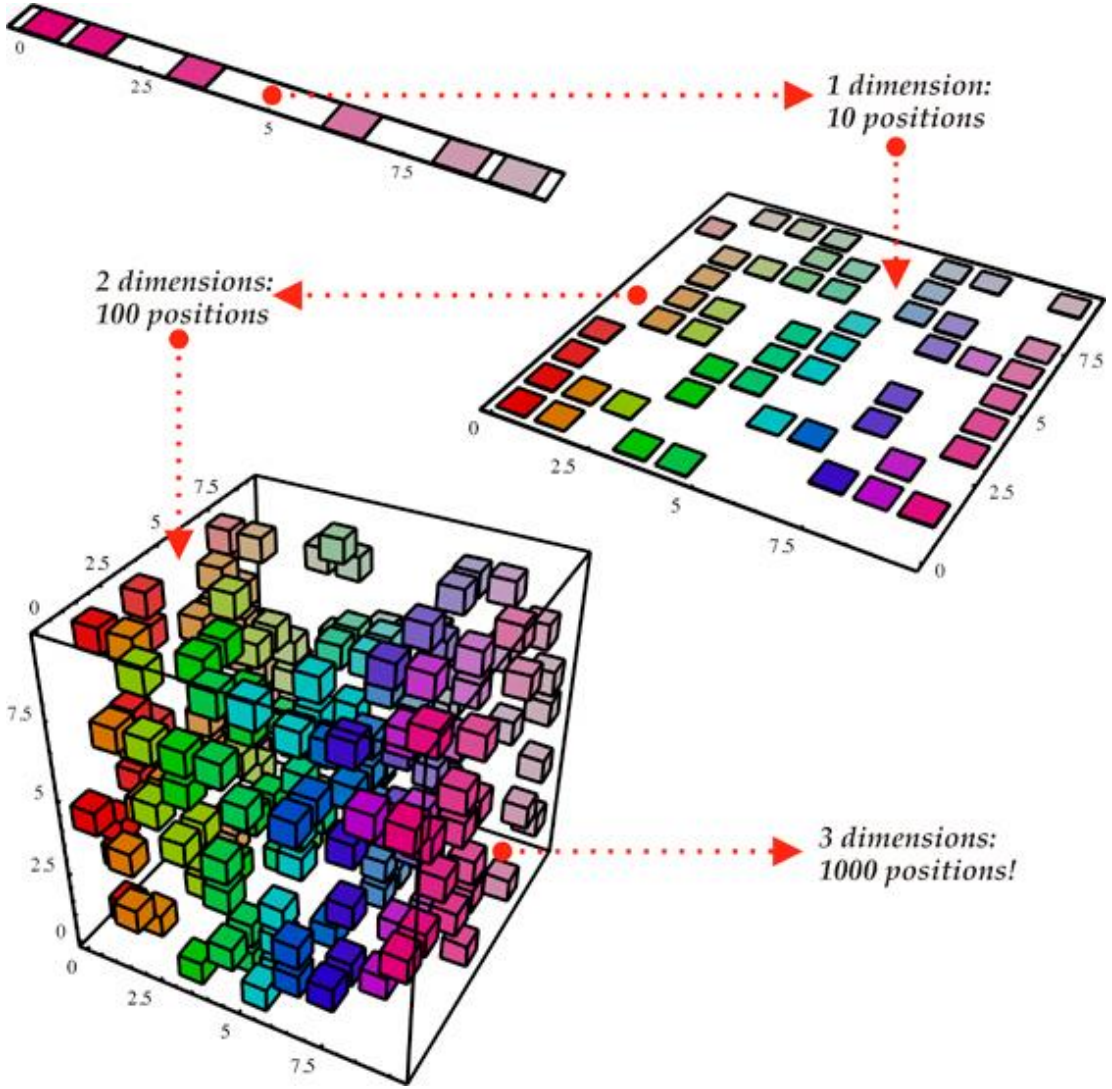




James et al, ISLR, 2013

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

KNN – MALDICIÓN DE LA DIMENSIONALIDAD



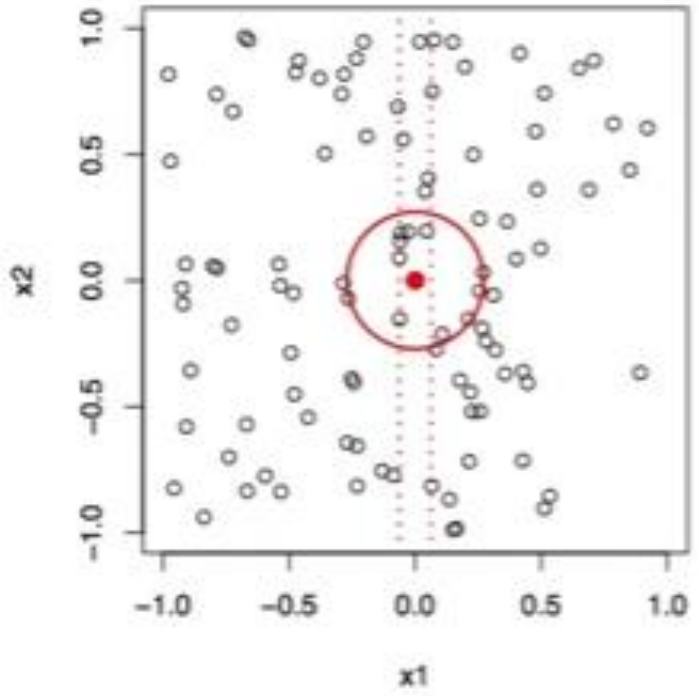
Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

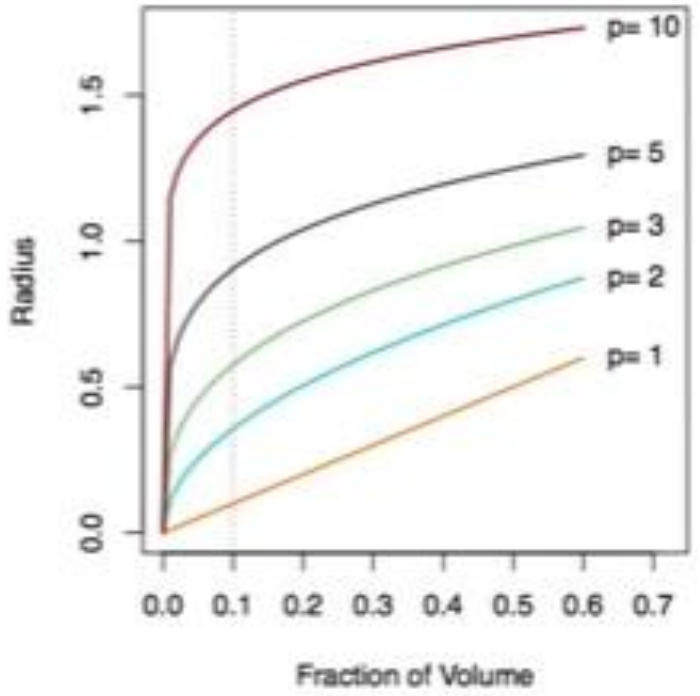
KNN – MALDICIÓN DE LA DIMENSIONALIDAD

- KNN puede llegar a ser un muy buen estimador cuando se considera un pequeño número P de variables predictivas (P <=4, con un buen número de ejemplos).
- KNN puede ser inútil cuando P es grande: todo está mucho mas lejos cuando se consideran altas dimensiones

Ejemplo: considerar el 10% de los vecinos más cercanos.

En altas dimensiones esos puntos no necesariamente son locales





Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

KNN: CARACTERÍSTICAS

- Perezoso (lazy learning), no paramétrico y no lineal
- **Método local**:
  - Puede encontrar particularidades muy específicas a ciertas regiones
  - Su uso (sobre todo en regresión) sólo permite estimaciones en los rangos de las variables del set de aprendizaje (extrapolación no tiene mucho sentido)
- Maldición de la **dimensionalidad**: no utilizar cuando el número de atributos es grande
- Al basarse en la **distancia**, es muy sensible a la **unidad de medida** de los atributos, y a atributos que no aportan poder predictivo (e.g. el color de los ojos no debería considerarse para predecir la edad de una persona)
- No sabe que hacer con los **missing values**, ni con variables **categóricas** (extensión → KnnCat)
- Complejidad temporal cuando hay **muchos registros** (extensión → CNN)

CLASIFICACIÓN: KNN

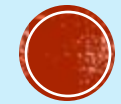
- 03-KNN-Ejemplo
  - Algoritmo k-nn desde cero
  - Uso del método neighbors.KNeighborsClassifier de sklearn

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

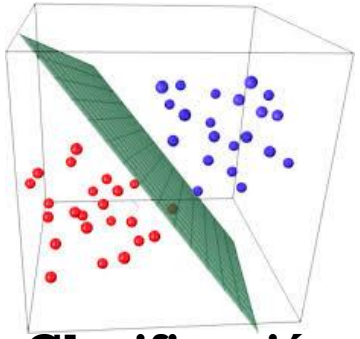
CLASIFICACIÓN: KNN

- Continuar con la parte de k-nn en el cuaderno 03-SAHeartDisease-LogReg+KNN.html

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

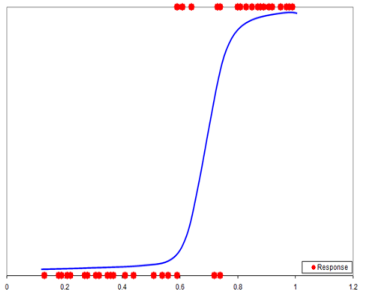


Aprendizaje automático

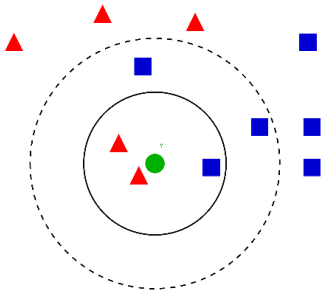


Clasificación

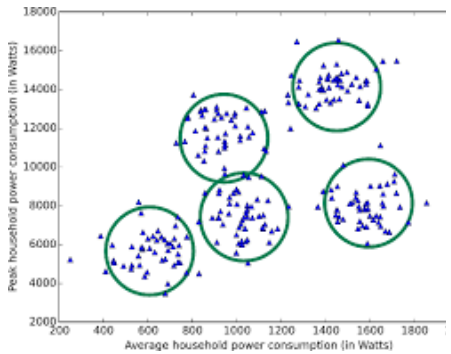
Métricas de Evaluación de la clasificación



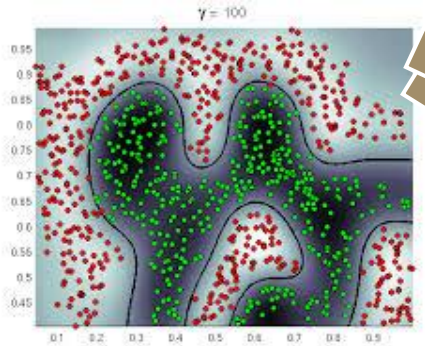
Regresión logística



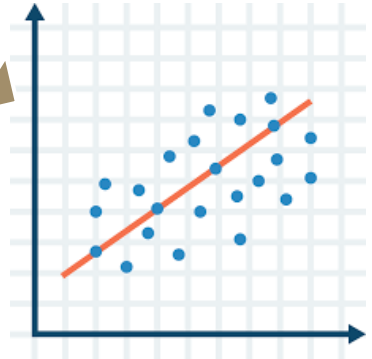
KNN



Aprendizaje no supervisado



Aprendizaje supervisado



Regresión

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

MÉTRICAS DE EVALUACIÓN

- Necesidad de evaluar la calidad de los modelos de aprendizaje automático
- Diferentes criterios a tener en cuenta:
  - Correctitud de la predicción
  - Simplicidad (parsimonia)
  - Interpretabilidad
  - Tiempo de aprendizaje o de predicción
  - Escalabilidad (importante para Big Data)

MÉTRICAS DE CLASIFICACIÓN

- Una **matriz de confusión** evalua diferentes métricas de correctitud, que permiten establecer **objetivos de negocio**
- Se utilizan dos calificadores para describir cada una de sus casillas:
  - Un calificador de la correctitud de la predicción con respecto a la realidad: Verdadero o Falso
  - Un calificador del tipo de la predicción: Positivo o Negativo, con respecto a cada clase de interés (i.e churn)
- Dependiendo del contexto los tipos de error pueden ser mas costosos que otros

		Predicción	
		Churn P	No churn N
Realidad	Churn +	VP	FN - Tipo II
	No churn -	FP - Tipo I	VN

- La diagonal (en verde) muestra las instancias correctamente clasificadas. Las demás casillas resume diferentes tipos de error:
  - Tipo I: Falsos positivos
  - Tipo II: Falsos negativos

¿Qué pasa cuando hay mas de dos clases?

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- **Interpretarían el caso de la detección de un email spam**

  TP, TN:

  FP: , consecuencia:

  FN: , consecuencia:

- **Interpretar el caso del diagnóstico de una enfermedad grave?**

  TP, TN:

  FP: , consecuencia:

  FN: , consecuencia:

		Predicción	
		Churn P	No churn N
Realidad	Churn +	VP	FN - Tipo II
	No churn -	FP - Tipo I	VN

- **Interpretar el caso de la prospección de clientes de un crédito de consumo (baja aceptación)**

  TP, TN:

  FP: , consecuencia:

  FN: , consecuencia:

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- Tasa de correctitud (categorization *accuracy*) = (VP+VN)/(VP+VN+FP+FN)
- Error de mala clasificación (opposite of *accuracy*) = (FP+FN)/(VP+VN+FP+FN): probabilidad de error
- Precisión= VP / (VP+FP): valor de predicción positiva, P(Real+|Predicho+)
- *Recall* (o TPR o sensibilidad)= VP / (VP+FN): qué proporción de todos los positivos reales pude identificar como tal, P(Predicho+|Real+)
- Especificidad (o TNR) = VN / (VN+FP): qué proporción de todos los negativos reales pude identificar como tal, P(Predicho-|Real-)
- Tasa de falsos positivos (o FPR) = FP / (VN+FP)

		Predicción	
		Churn P	No churn N
Realidad	Churn +	VP	FN - Tipo II
	No churn -	FP - Tipo I	VN

Medida-F (F-Score): Promedio armónico entre precisión y recall
2* (Precisión*Recall) / (Precisión+Recall)

Imaginemos el problema de detección de spam mail e interpretemos cada métrica

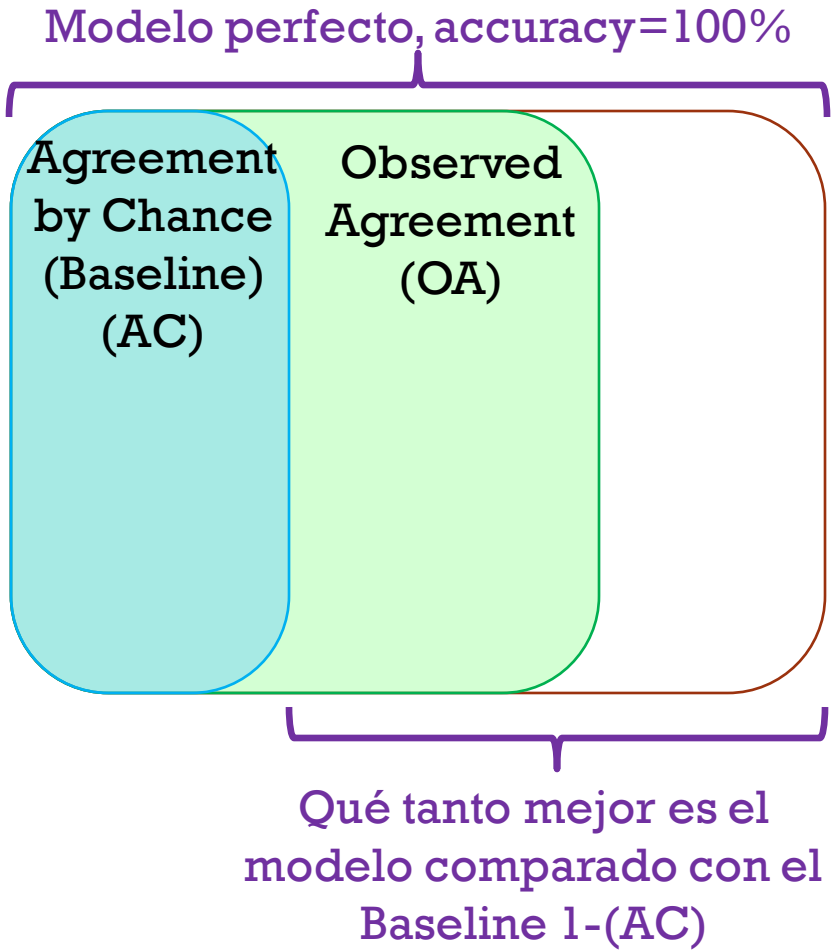
Imaginemos el problema de diagnóstico de cáncer e interpretemos cada métrica

¿Cuál es la relación entre especificidad y FPR?
¿Cómo son la especificidad y sensibilidad del baseline?

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- Coeficiente de concordancia **Kappa**
  - Para datos nominales u ordinales
  - Concordancia entre las predicciones y las clases reales
  - Sustrae el efecto de concordancia por suerte (AC) del valor del **accuracy** (concordancia observada - OA)
  - Valores van de 0 a 1
  - Muy útil sobretodo cuando las clases no están balanceadas
    - Diagnóstico de enfermedades raras
    - Clientes que acepten productos de crédito)
  - Kappa = $\frac{OA-AC}{1-AC}$



Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- Coeficiente de concordancia **Kappa**
  - Para datos nominales u ordinales
  - Concordancia entre las predicciones y las clases reales
  - Sustrae el efecto de concordancia por suerte (AC) del valor del **accuracy** (concordancia observada - OA)
  - Valores van de 0 a 1
  - Muy útil sobretodo cuando las clases no están balanceadas
    - Diagnóstico de enfermedades raras
    - Clientes que acepten productos de crédito)

		Predicciones		
		+	-	TOTAL
reales	+	10	4	14
	-	3	2	5
	TOTAL	13	6	19

OA = 0,63
AC = 0,59
Kappa = 0,11

Accuracy (OA) = (10+2)/19=0,63
(AC) = (13/19 * 14/19) + (6/19 * 5/19) = 0,59
Kappa = (OA-AC)/(1-AC) =0,11

		Predicciones		
		+	-	TOTAL
reales	+	0	3	3
	-	0	97	97
	TOTAL	0	100	100

OA = 0,97
AC = 0,97
Kappa = 0,00

Accuracy (OA) = (0+97)/100=0,97
(AC) = (0/100 * 3/100) + (100/100 * 97/100) = 0,97
Kappa = (OA-AC)/(1-AC) =0

		Predicciones		
		+	-	TOTAL
reales	+	1475	988	2463
	-	556	1981	2537
	TOTAL	2031	2969	5000

OA = 0,69
AC = 0,50
Kappa = 0,38

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN
TALLER: CÁLCULO DE MÉTRICAS

Los clientes son usualmente categorizados en perfiles de comportamiento de compra de productos o servicios.

Suponga que se creó un modelo para clasificar los clientes en una de 4 clases posibles (esporádico, fiel, parcial y promocional), cuya matriz de confusión presentamos a continuación:

	PREDICCIÓN				
REALIDAD	**Esporádico**	**Fiel**	**Parcial**	**Promocional**	**Total**
Esporádico	61	8	1	0	70
Fiel	0	56	17	0	73
Parcial	0	0	15	0	15
Promocional	0	0	0	24	24
Total	61	64	33	24	182

En **grupos de tres personas** calcule las métricas de evaluación de un modelo de clasificación cuyos resultados están reflejados en la tabla siguiente (ver PDF)

UNIVERSIDAD ICESI

TALLER DE CLASIFICACIÓN: CÁLCULO DE MÉTRICAS

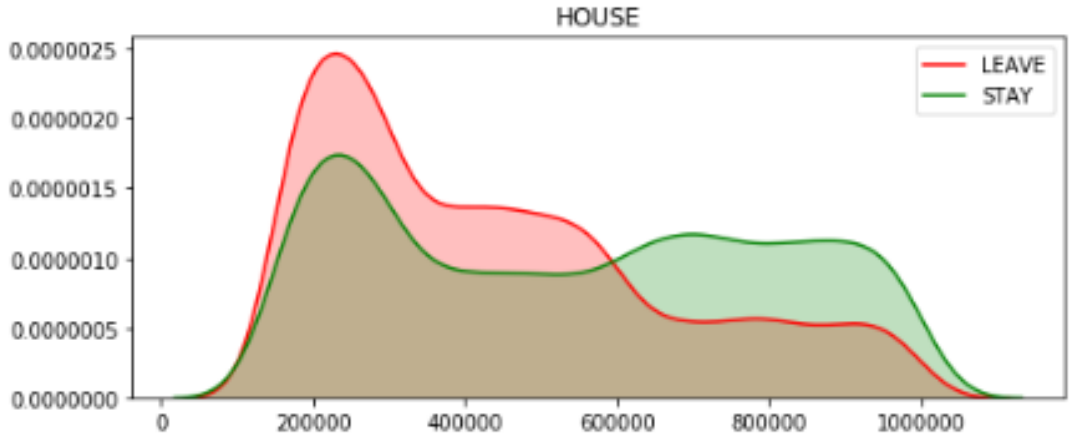
- Trabajar en grupo sobre la tarea para posterior socialización.
- Indicar los integrantes de cada grupo.
- Socialización del taller.

UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN (ROC AUC, LOG-LIKELIHOOD, DEVIANCE, AIC, ENTROPÍA, BIC)

Evaluación de modelos **probabilísticos** de clasificación:

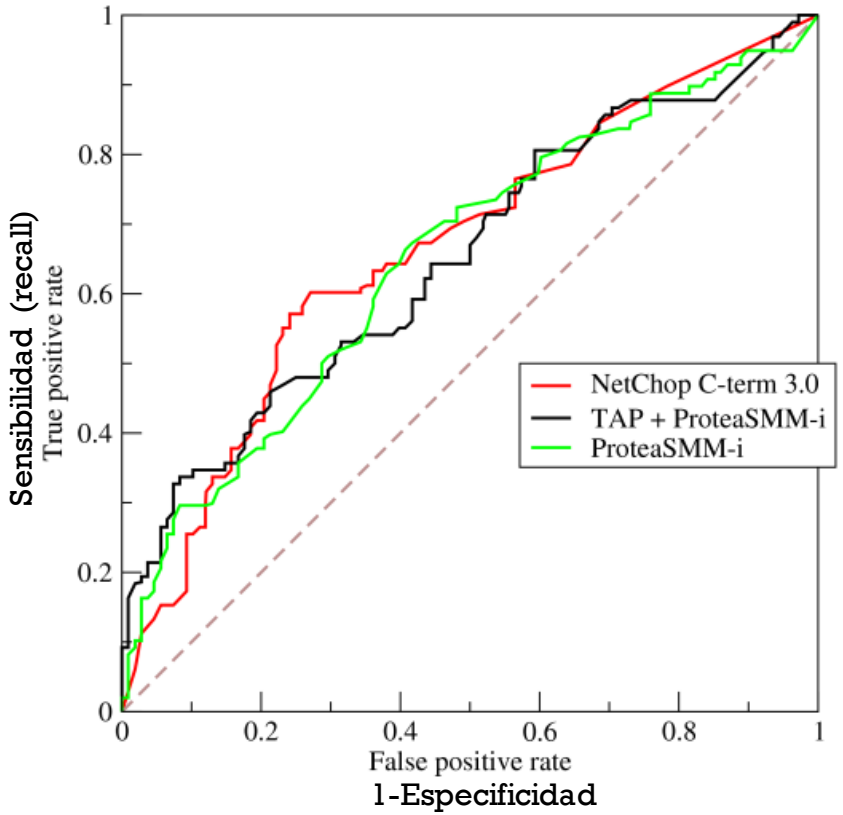
- Métricas que permiten comparar modelos con resultados **probabilísticos** como regresión logística, Naïve Bayes, árboles de decisión, KNN, etc., sobre el mismo dataset
- No deben ser la base para establecer metas de negocio
- Insensibles a diferentes costos del error de las diferentes clases
- Utilidad de los gráficos de **densidad de probabilidad** para entender los efectos de las variables predictivas



Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- **ROC AUC** (Area under the Curve)
  - **Receiver Operating Curve**: Curva de la tasa de los verdaderos positivos (recall o sensibilidad) contra los falsos positivos (1-especificidad)
  - Clasificadores binarios
  - Se hace variar el umbral de definición de predicción positiva y negativa
  - Permite establecer con los clientes posibles compromisos entre los dos tipos de errores



Wikipedia.org

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- **Log-Likelihood (log-verosimilitud)**
  - Si el modelo explica los datos, las predicciones del modelo deben ser verosímiles (plausibles)
  - Logaritmo del producto de las **probabilidades predichas** de las **categorías reales**, por un modelo con unos parámetros Θ dados
  - Siempre negativa, entre mas cercana a 0 mejor
  - Ejemplo:
    - Tenemos 5 clientes cuyas predicciones de un modelo son "correctas".
    - 2 clientes compran y el modelo les otorga las probabilidades $P(Compra = Si|\Theta)$ de 0,8 y 0,9 (resp.)
    - 3 clientes no compran y el modelo les otorga probabilidades $P(Compra = Si|\Theta)$ de 0,2, 0,3 y 0,9(resp.)
    - **¿Cuál es el log-likelihood de este modelo con ese set de datos?**

$$Log\mathcal{L}(\Theta|Compra) = \ln\left(\prod_{i=1}^{5} P(categoriaCompraReal(i)|\Theta)\right)$$

$$= \ln(0{,}8 * 0{,}9 * 0{,}8 * 0{,}7 * 0{,}1) = \ln(0{,}8) + \ln(0{,}9) + \ln(0{,}8) + \ln(0{,}7) + \ln(0{,}1) = \sum_{i=1}^{5} \ln\left(P(i|\Theta)\right)$$

$$= -3{,}093$$

MÉTRICAS DE CLASIFICACIÓN

- **Deviance (“desvío”)**
  - Medida de ajuste de modelos de probabilidad basada en el **log-likelihood**

$$D = -2 * (Log\mathcal{L}(\Theta|Y) - S)$$

  - Supone conocer una constante **S** que representa el log-likelihood del **modelo saturado**, pero como nos interesan las diferencias entre modelos, se cancelan
  - Si suponemos un valor de S=0, el deviance se puede considerar como una versión análoga inversa del R^2 para clasificación con probabilidades, ya que representa cuanta variación falta por explicar.
  - Un menor valor indica un mejor ajuste

MÉTRICAS DE CLASIFICACIÓN

- **AIC (Akaike Information Criteria)**
  - Mide la cantidad de información relativa que se pierde al usar un modelo como estimador en vez del **modelo saturado**, por lo que sirve para comparar y seleccionar modelos estadísticos (también aplica para regresión).
  - No mide la calidad del modelo en sí, el AIC no permite establecer si los modelos comparados son buenos o malos.
  - Es una variante del **deviance** que incluye una **penalización** con respecto al **número de parámetros** con propósitos de **regularización** de la complejidad del modelo.

$$AIC = -2 * \left(Log\mathcal{L}(\Theta|Y)\right) + 2 * numParams(\Theta)$$

  - Entre más datos se tengan, mejor será el desempeño de esta métrica. Para pequeños datasets existe una variante (AICc)
  - Un menor valor indica un mejor ajuste (se pierde menos información). Pueden haber valores negativos.

MÉTRICAS DE CLASIFICACIÓN

- **BIC (Bayesian Information Criteria)**
  - Medida análoga al AIC, con una penalización diferente de la complejidad del modelo (también aplica para regresión).

$$BIC = -2 * \left(Log\mathcal{L}(\Theta|Y)\right) + \ln(n) * numParams(\Theta)$$

    con n siendo el número de registros
  - Un menor valor indica un mejor ajuste. Pueden haber valores negativos.

Javier Diaz Cely, PhD
UNIVERSIDAD ICESI

REFERENCIAS

- *Introduction to Statistical Learning with Applications in R (ISLR),* G. James, D. Witten, T. Hastie & R. Tibshirani, 2014
- *Practical Data Science with R*, Nina Zumel & John Mount, Manning, 2014
- *Data Mining (4th Edition)*, Ian Witten, Eibe Frank, Mark A. Hall & Christopher J. Pal, Elsevier, 2016
- *Machine Learning,* Tom M. Mitchell, McGraw-Hill,1997
- *Data Science for Business*, Foster Provost & Tom Fawcett, O'Reilly, 2013
- *False positives, false negatives and confusion matrices,* Carlos Guestrin, 2017
- http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/