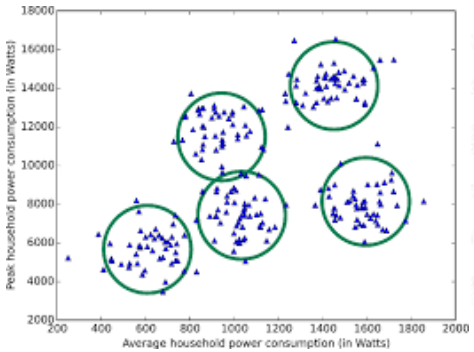
APRENDIZAJE AUTOMÁTICO

Javier Diaz Cely, PhD

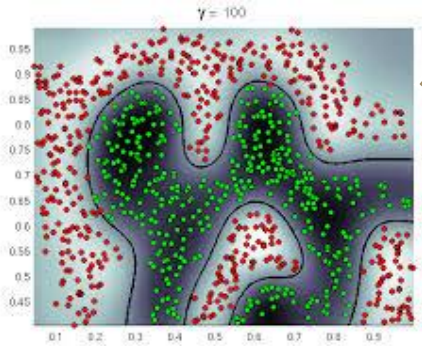
UNIVERSIDAD ICESI

AGENDA

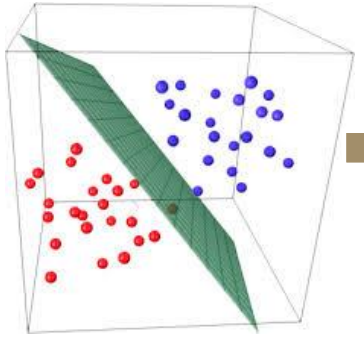
Aprendizaje automático



Aprendizaje no supervisado

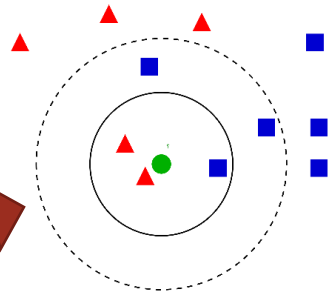


Aprendizaje supervisado

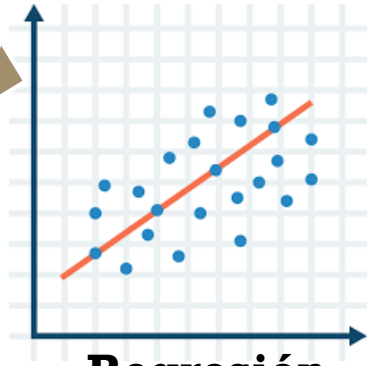


Clasificación

Métricas de Evaluación de la clasificación



KNN



Regresión

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

DATOS:
Materia prima

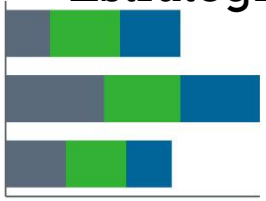
PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

MODELO:
Implementación del modelo de analítica

EVALUACION:
Calidad del modelo
Ajuste del modelo

DESPLIEGUE:
Resultados
Conocimiento
Estrategia



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

UNIVERSIDAD ICESI

Machine Learning

what society thinks I do

what my friends think I do

what my parents think I do

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

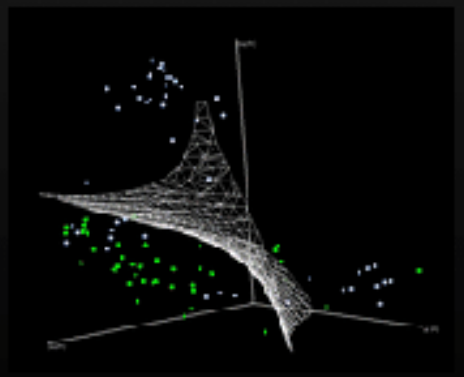
$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\nabla \hat{J}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t)$$

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

what other programmers think I do



what I think I do

```
1  library(ggplot2)
2  library(caret)
3
4  canciones <- read.table('
5  str(canciones)
6  summary(canciones)
7  head(canciones)
8
```

what I really do

UNIVERSIDAD ICESI

APRENDIZAJE AUTOMÁTICO

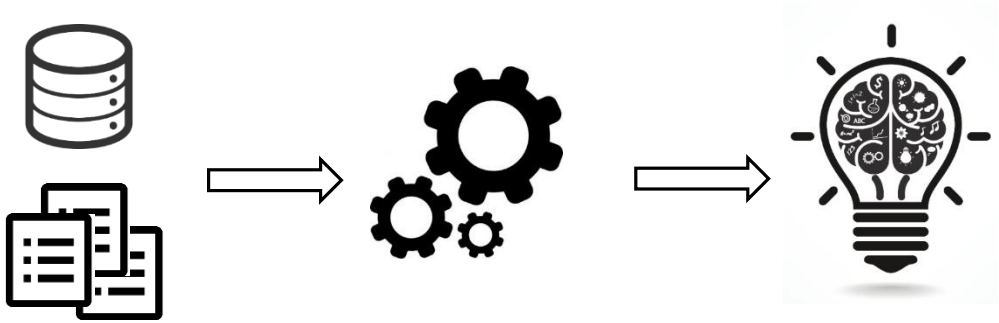
- **¿Por qué es necesario?**
  - Tareas complejas extremamente difíciles de programar
  - Poder computacional disponible para tratar grandes volúmenes de datos

Las máquinas tienen que aprender por sí solas

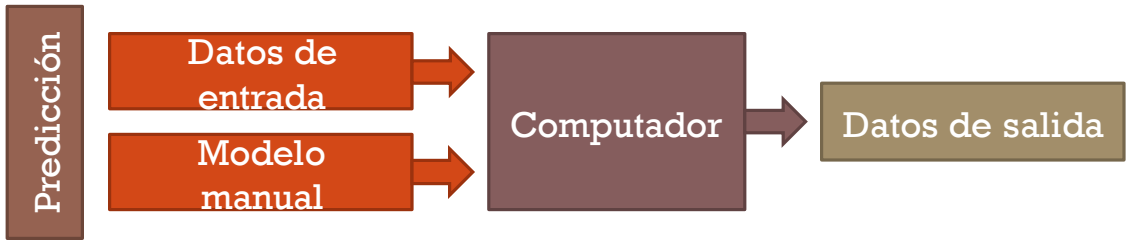
APRENDIZAJE AUTOMÁTICO

- Definición:

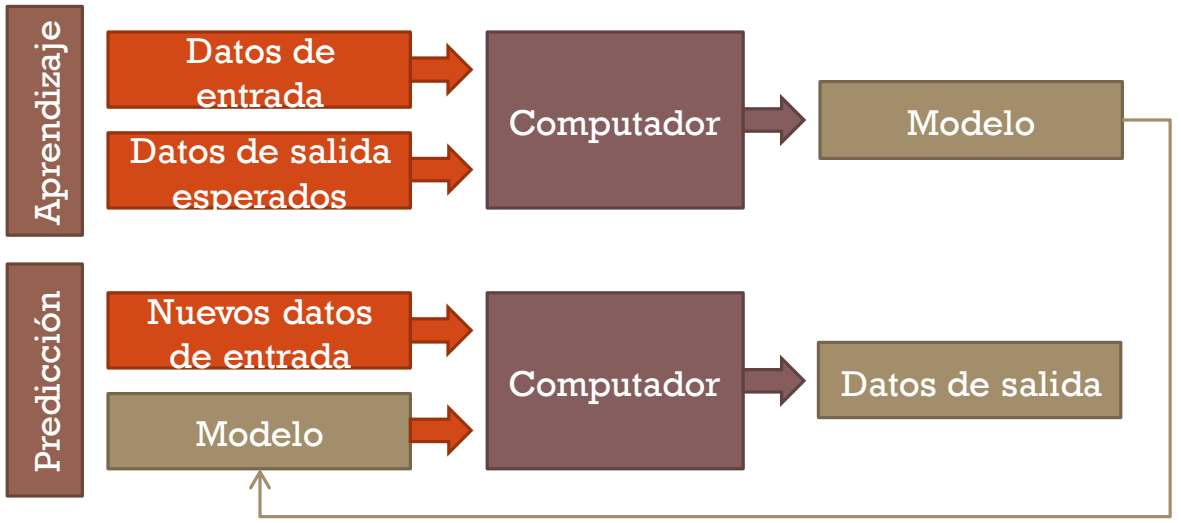
  El aprendizaje automático es la ciencia que permite a los computadores aprender, sin ser explícitamente programados[1]



Modelo tradicional



Ciencia de datos



1. Andrew Ng, Stanford University, 2014

UNIVERSIDAD ICESI

APRENDIZAJE AUTOMÁTICO

Aprendizaje supervisado

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

(x1, x2, …, xn, y)

Predictores, explicativos, independientes

Dependiente, objetivo, salida

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

Aprendizaje no supervisado

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

(x1, x2, …, xn)

- **Meta**: descubrir factores no observados, estructura, o una representación mas simple de los datos

UNIVERSIDAD 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

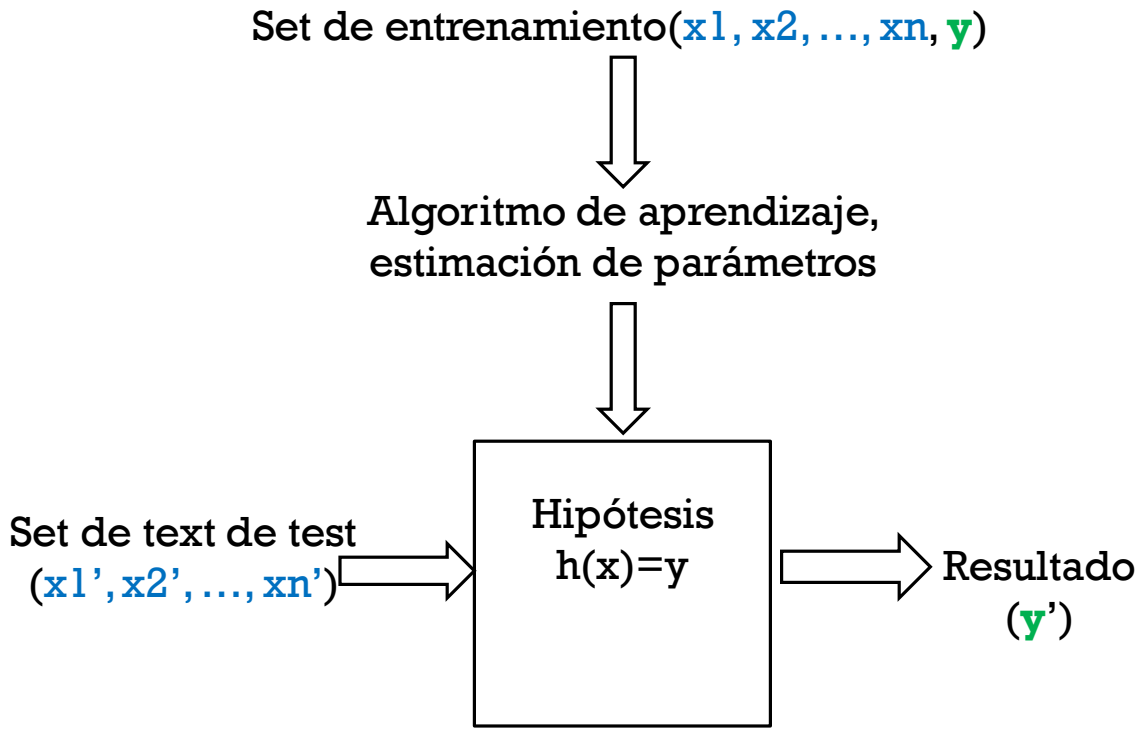
Factores/atributos/variables

¿Se puede encontrar alguna estructura en los datos?

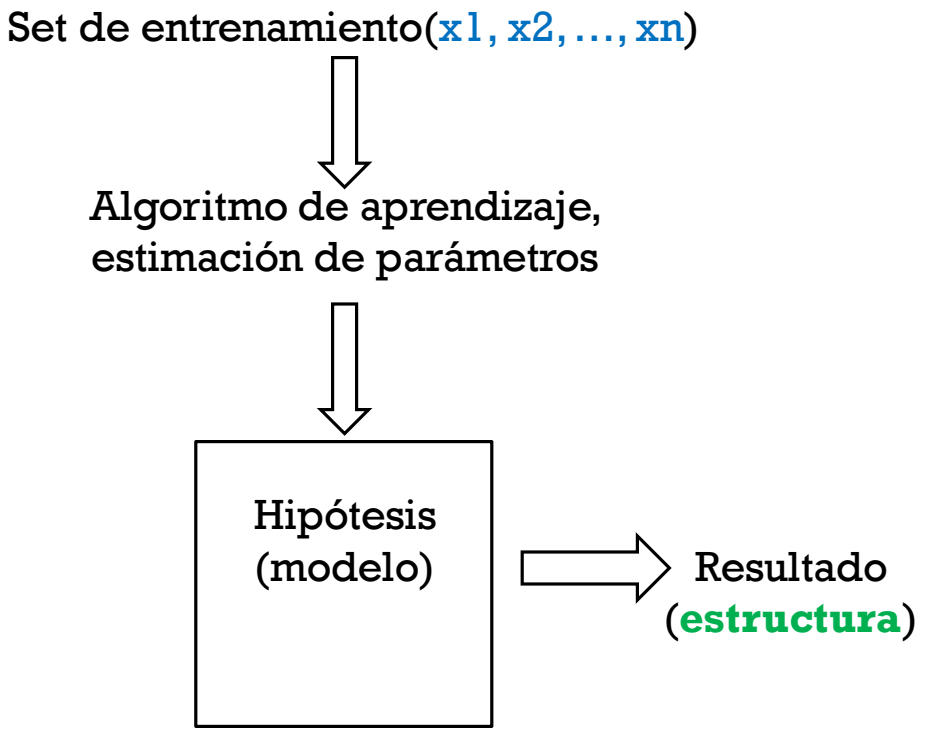
UNIVERSIDAD ICESI

APRENDIZAJE AUTOMÁTICO

Aprendizaje supervisado



Aprendizaje no supervisado



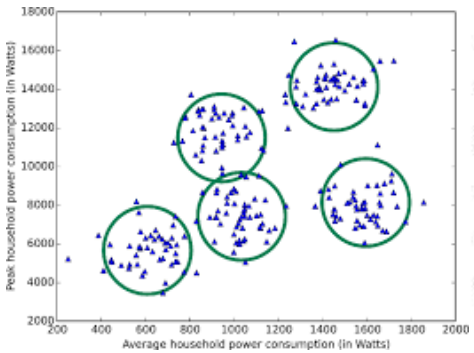
UNIVERSIDAD ICESI

MÉTRICAS DE EVALUACIÓN

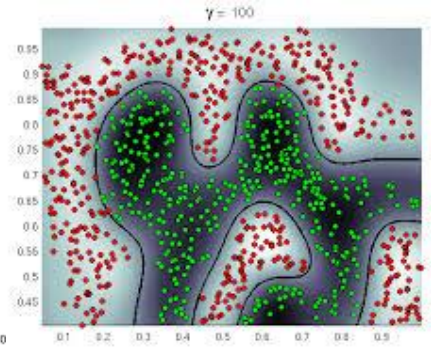
- Necesidad de evaluar la calidad de los modelos de aprendizaje automático
- Diferentes criterios a tener en cuenta:
  - Correctitud de la predicción
  - Simplicidad (parsimonia)
  - Interpretabilidad
  - Tiempo de aprendizaje o de predicción
  - Escalabilidad (importante para Big Data)

AGENDA

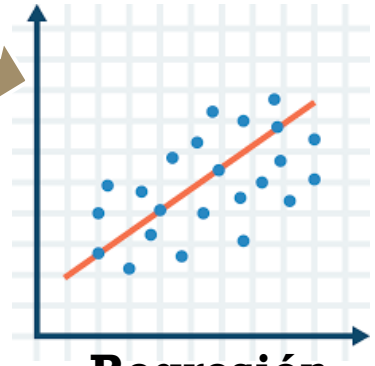
Aprendizaje automático



Aprendizaje no supervisado



Aprendizaje supervisado

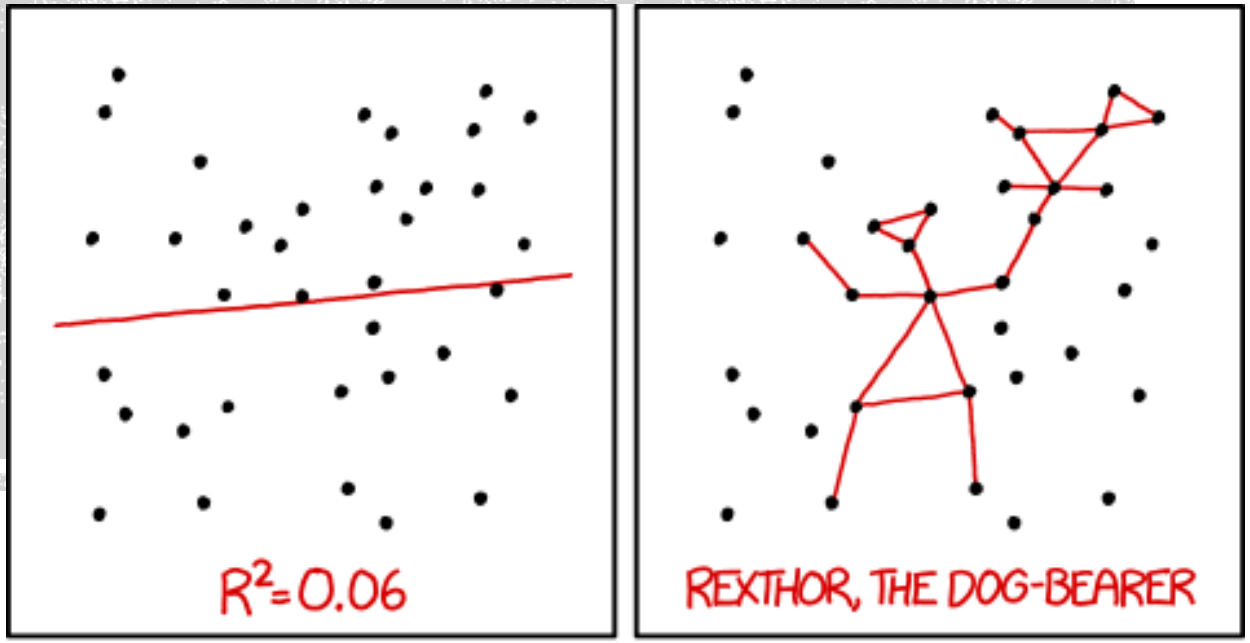


Regresión

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

REGRESIÓN

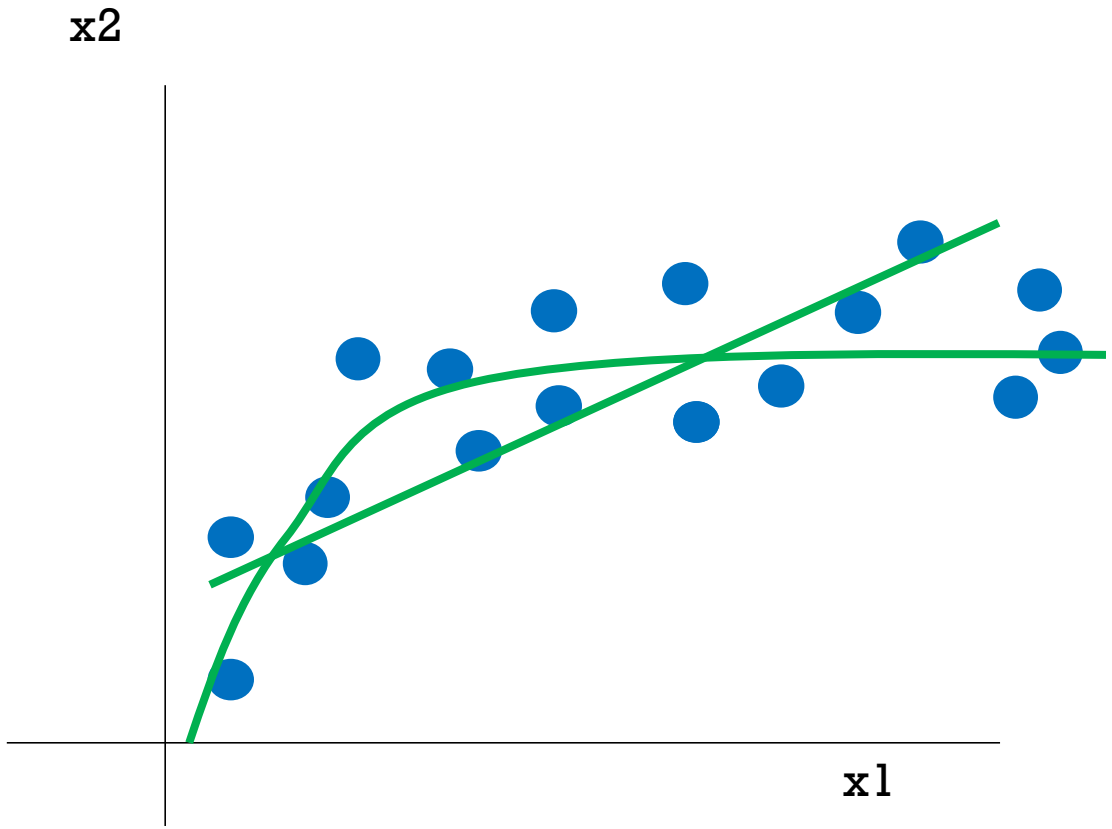


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

UNIVERSIDAD ICESI

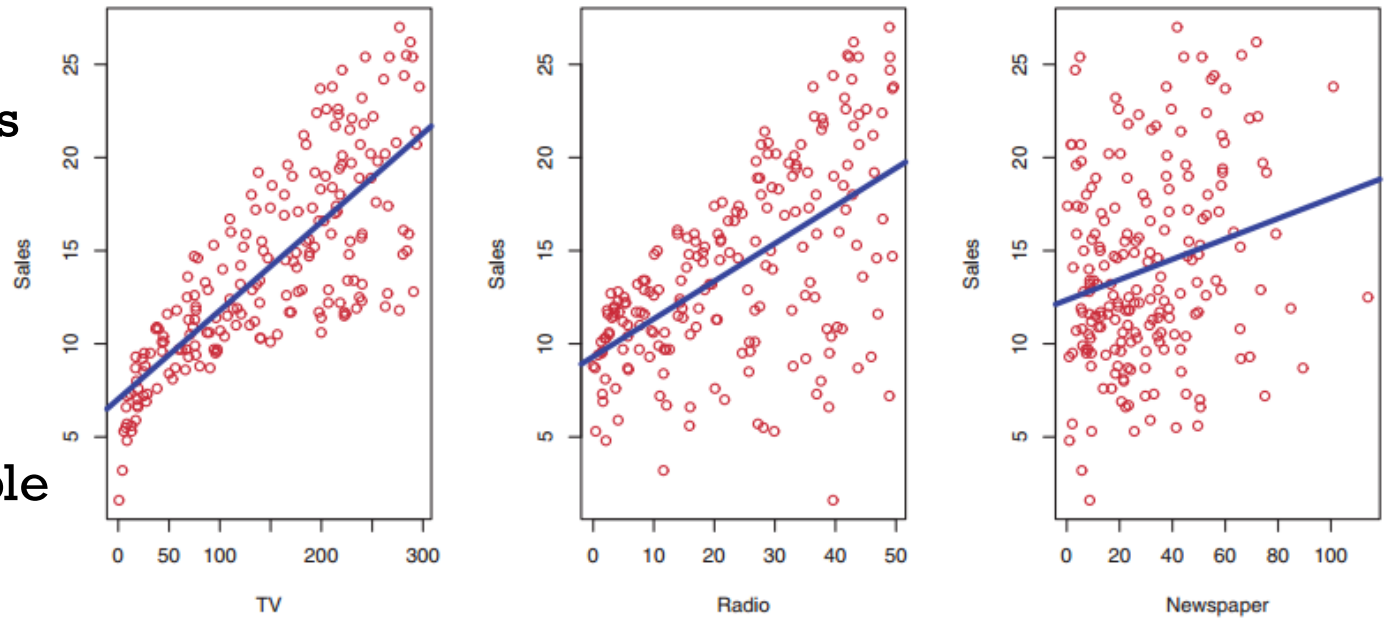
REGRESIÓN

- Encontrar modelos que permitan predecir valores continuos:
  - KNN
  - Regresión lineal
  - Regresión polinómica
  - Árboles de regresión
  - …
- Valores **continuos** de la variable 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 resumir esa relación linealmente o se trata de una relación mas compleja?



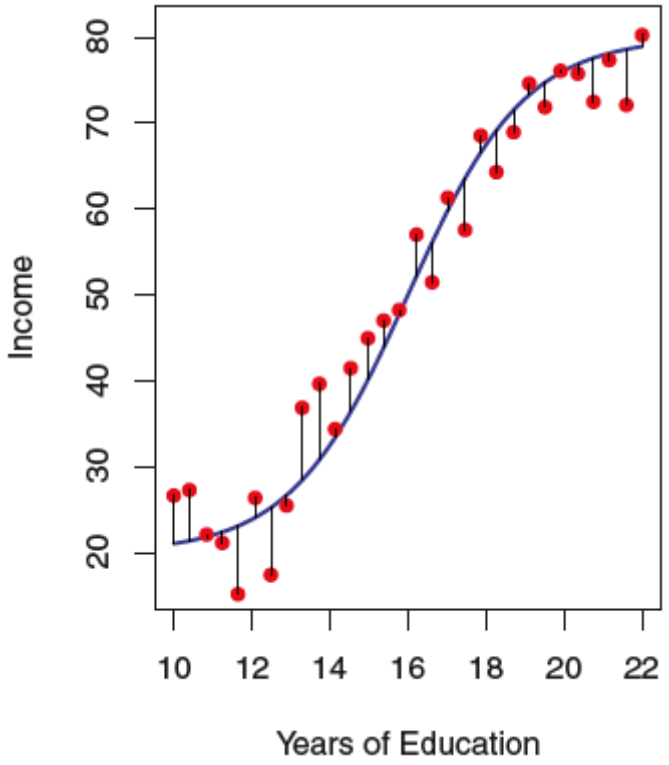
$$Ventas = f(TV, Radio, Periódicos)$$

ISLR, 2013

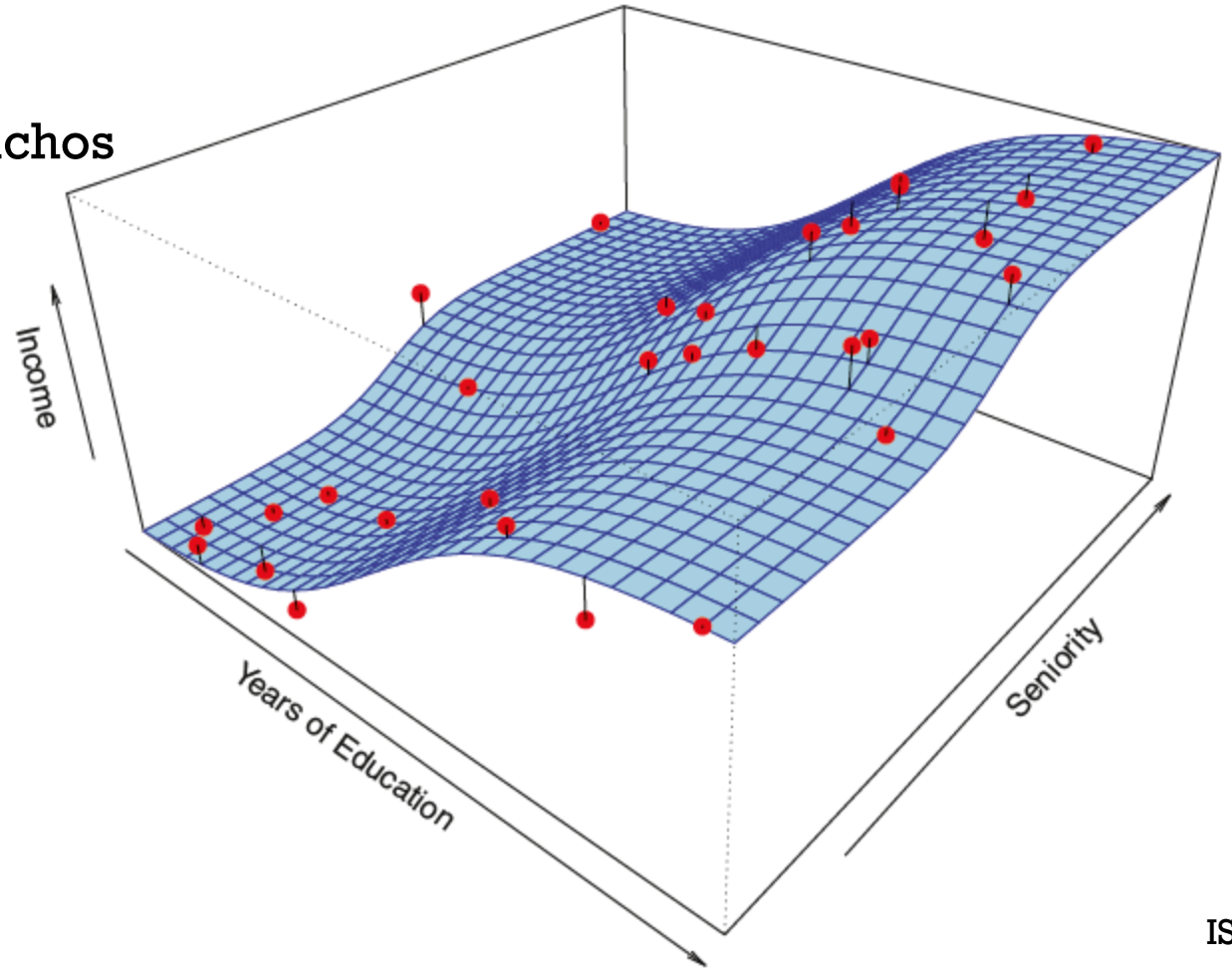
UNIVERSIDAD ICESI

RESIDUOS

Residuos: diferencia entre los valores reales y los valores predichos



Regresión polinómica



Regresión no paramétrica: thin-plate spline

ISLR, 2013

MÉTRICAS DE REGRESIÓN

Coeficiente de correlación (Pearson $\rho \in [-1;1]$): indica la fuerza de la relación linear entre los predictores y la variables objetivo, que puede ser positive 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 ($R^2 = \rho^2$): indica el porcentaje de la varianza que puede 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):

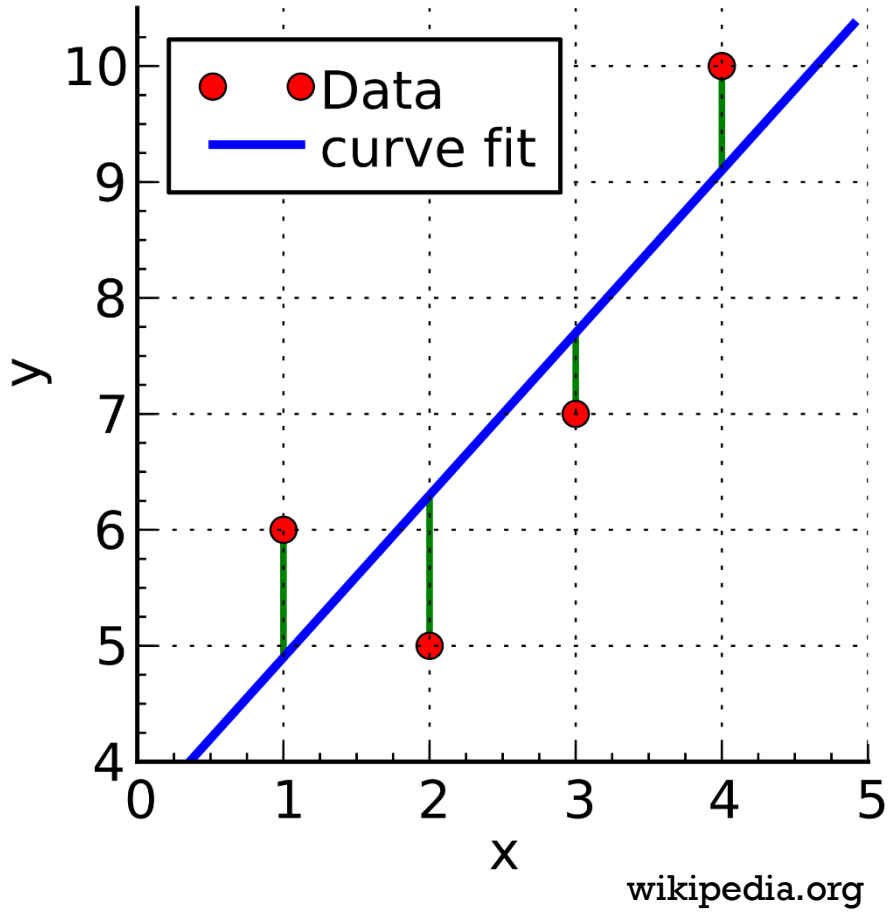
$$\frac{1}{m}\sum_{1}^{m}(h_\theta(x_i)-y_i)^2$$

- 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

VARIABLES CATEGÓRICAS

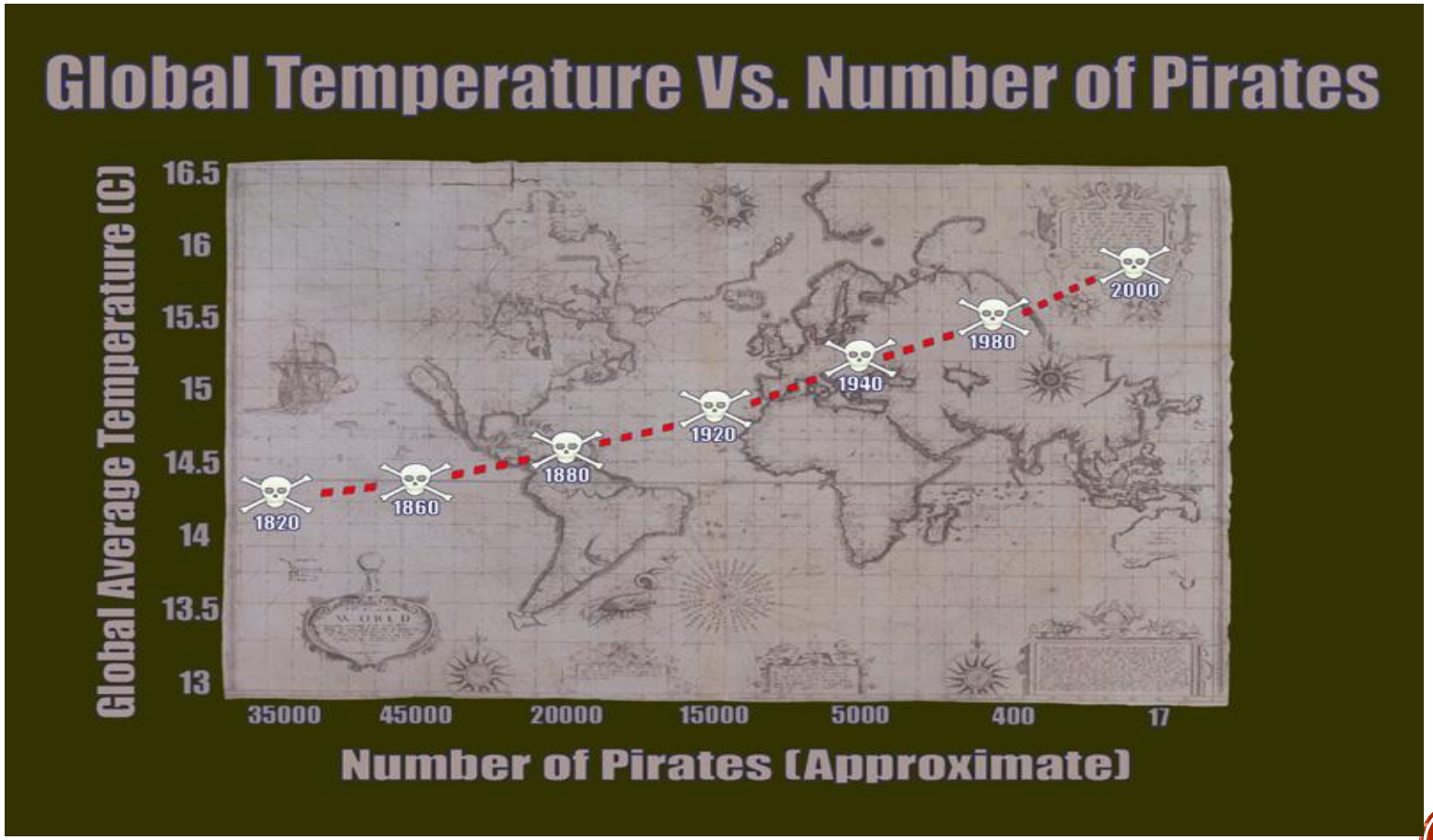
- Las variables predictoras deben ser numéricas.
- Las variables categóricas debes ser convertidas en numéricas:
  - One hot encoding: se crea una variable para cada valor posible de cada variable categórica
  - Contraste o "dummy": se crea una variable para cada valor posible menos 1 de cada variable categórica.

Ejemplo: variable estrato con 3 valores posibles (bajo, medio y alto)

	Estrato_bajo	Estrato_medio
Valor = bajo	1	0
Valor = medio	0	1
Valor = alto	0	0

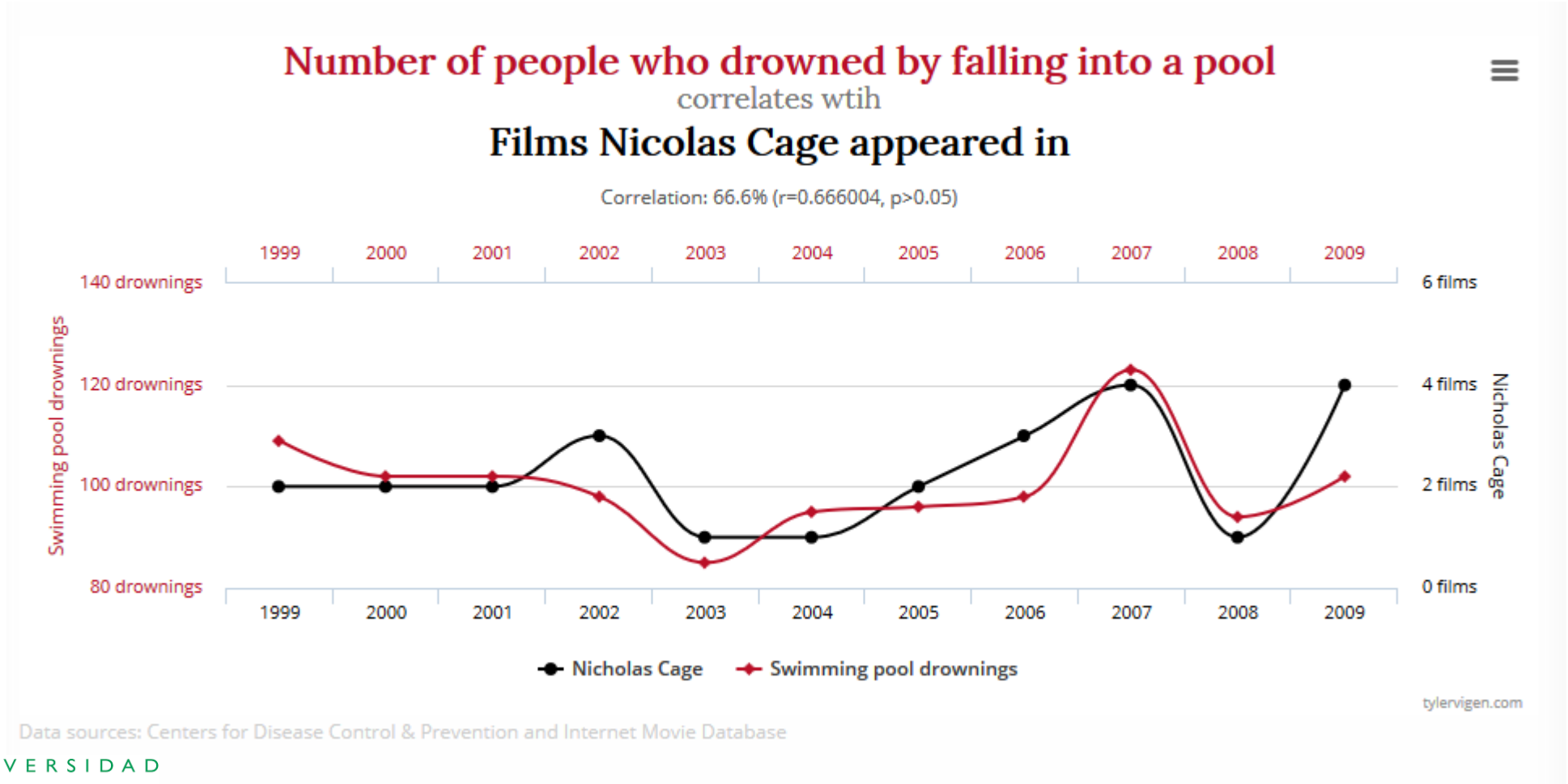
REGRESIÓN – CUIDADO!

Correlación y causalidad son dos cosas muy diferentes



Erika Andersen, Forbes, 2012

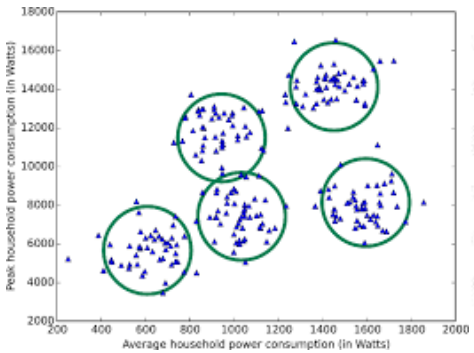
REGRESIÓN – CUIDADO!



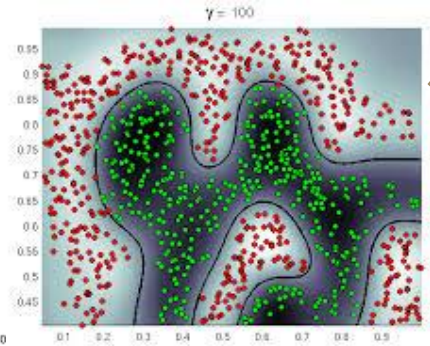
UNIVERSIDAD ICESI

AGENDA

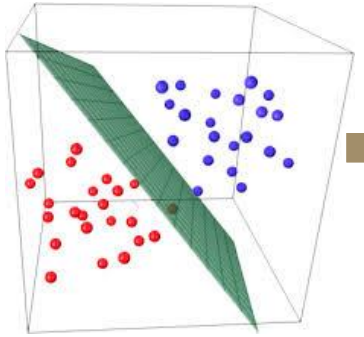
Aprendizaje automático



Aprendizaje no supervisado

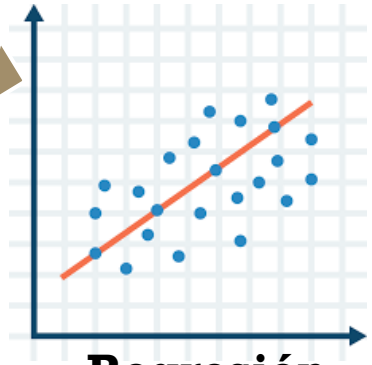


Aprendizaje supervisado



Clasificación

Métricas de Evaluación de la clasificación

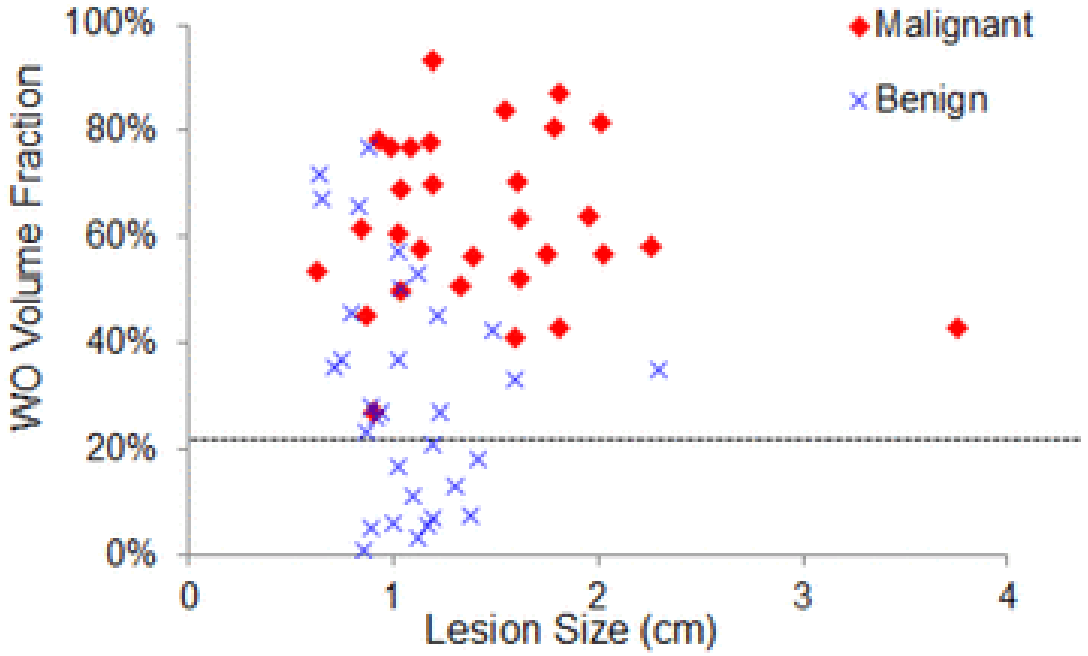


Regresión

Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

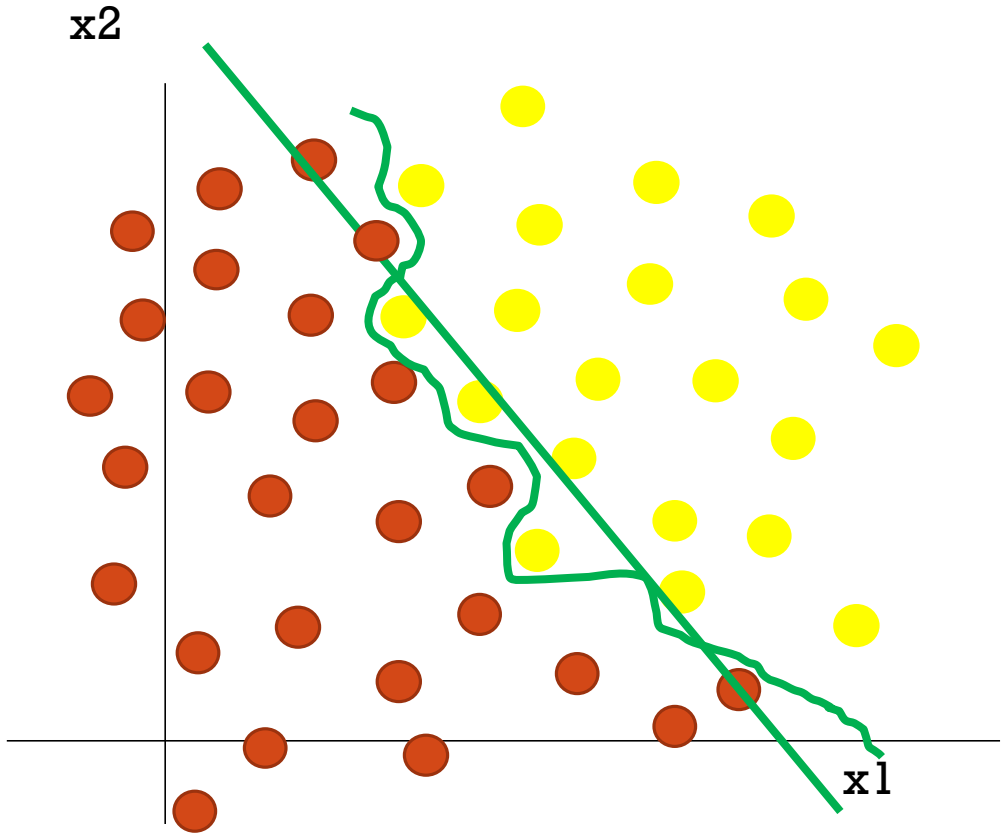
CLASIFICACIÓN



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

CLASIFICACIÓN

- Encontrar modelos que describan clases para futuras predicciones:
  - KNN
  - Árboles de decisión
  - Regresión logística
  - Redes neuronales
  - …
- Valores **discretos** de la variable objetivo
- Incluye la estimación de **probabilidades** de clase
- **Baseline**: medida de evaluación dada por un clasificador que escoge siempre la clase mayoritaria



MÉTRICAS DE CLASIFICACIÓN

- Se usa una **matriz de confusión** para evaluar diferentes métricas de correctitud/error
- 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 graves que otros (costos diferentes)

		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

UNIVERSIDAD ICESI

¿Qué pasa cuando hay mas de dos clases?

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:

UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

- Tasa de correctitud (accuracy) = (VP+VN)/(VP+VN+FP+FN)
- Error de mala clasificación (contrario de *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-)
- Valor de predicción negativa (FPR) = VN / (VN+FN)
- F1-Measure = $2 * \frac{precision * recall}{precision + recall}$ (promedio armónico)

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

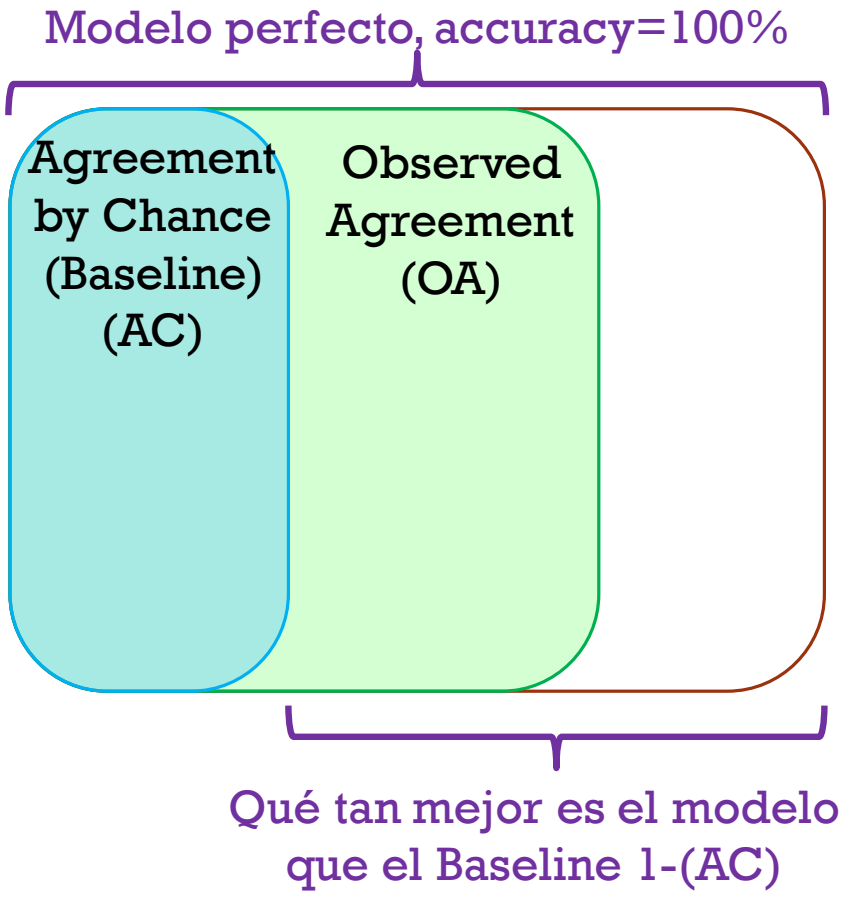
Imaginemos el problema de detección de spam mail e interpretemos cada métrica

Imaginemos el problema de diagnóstico de cáncer e interpretemos cada métrica

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 = $\frac{OA-AC}{1-AC}$



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

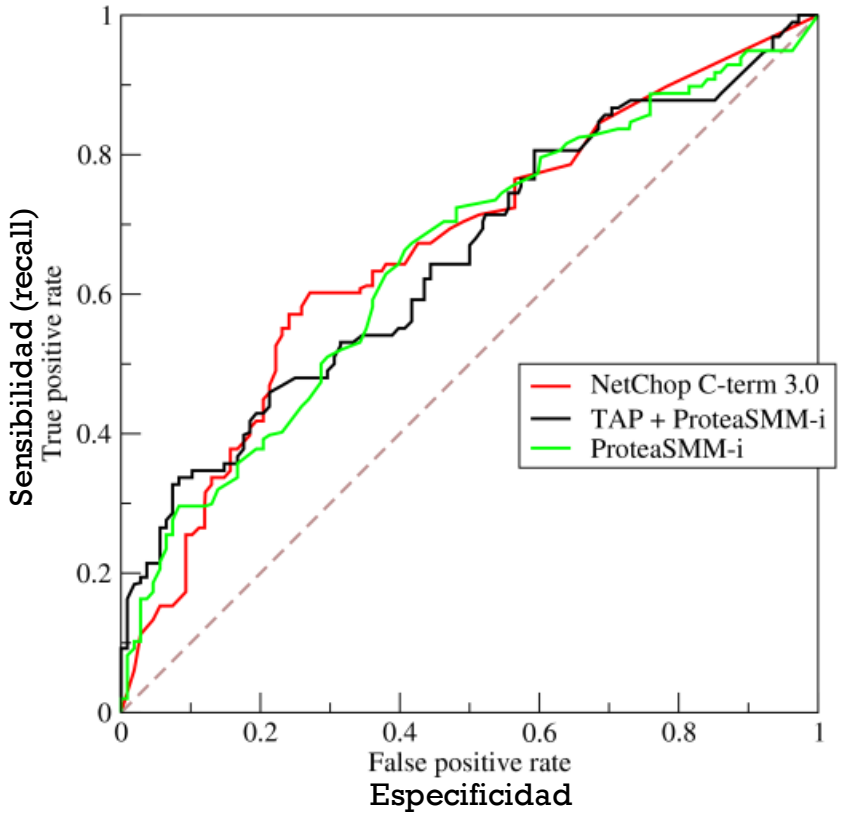
UNIVERSIDAD ICESI

MÉTRICAS DE CLASIFICACIÓN

Comparación de varios modelos:

- Validación cruzada con accuracy o error de mala clasificación
- Medida-F=

  2* (Precisión*Recall) / (Precisión+Recall)
- Área debajo de la curve ROC
- Todas las métricas son insensibles a los diferentes costos del error de las diferentes clases, que depende del contexto. Pueden tener impactos diferentes.



Wikipedia.org

MÉTRICAS DE CLASIFICACIÓN
TALLER: CÁLCULO DE MÉTRICAS

Calcule las métricas de evaluación de un modelo de clasificación cuyos resultados están reflejados en la tabla siguiente

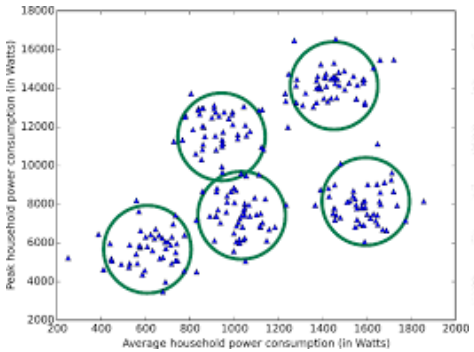
- Error, Accuracy y Kappa global
- Precisión, Recall, especificidad, F-Measure de cada clase

	PREDICCIÓN				
REAL	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

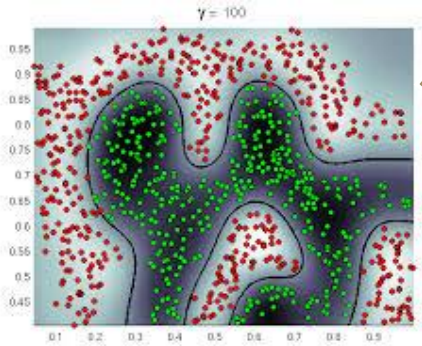
UNIVERSIDAD ICESI

AGENDA

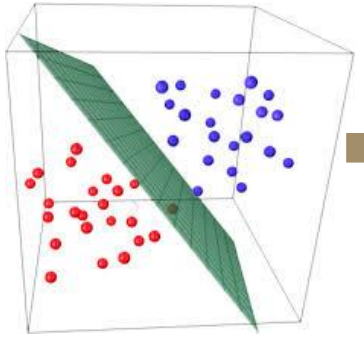
Aprendizaje automático



Aprendizaje no supervisado

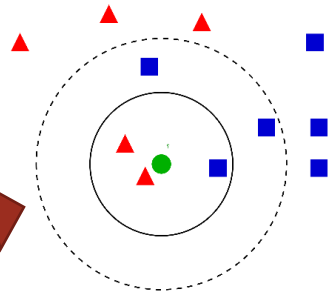


Aprendizaje supervisado

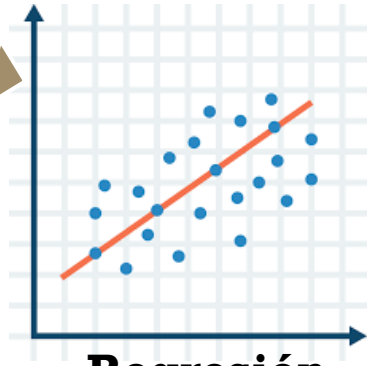


Clasificación

Métricas de Evaluación de la clasificación



KNN



Regresión

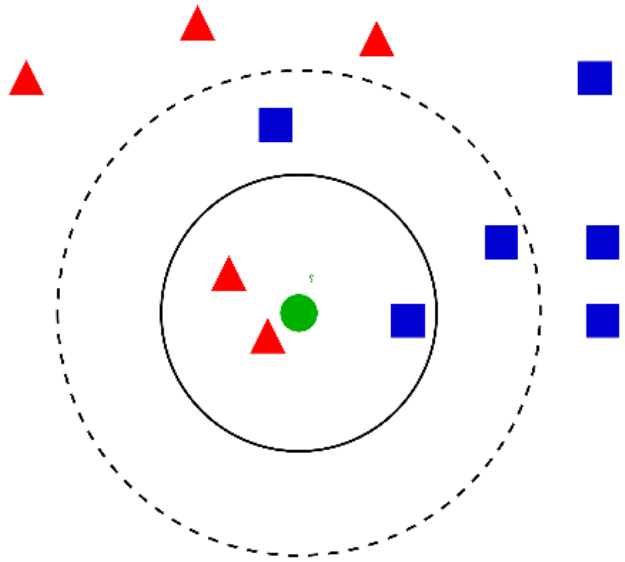
Métricas de Evaluación de la regresión

UNIVERSIDAD ICESI

KNN

KNN (K Nearest Neighbors): K Vecinos más Cercanos

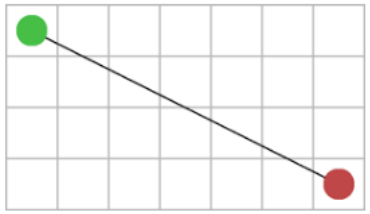
- Algoritmo de aprendizaje supervisado para **clasificación** y **regresión**
- **Sencillo**: asignar la clase o valor agregado de las instancias conocidas que se encuentran mas cerca de la instancia a predecir
- Basado en las **instancias** de aprendizaje, no en un modelo subyacente probabilístico/estadístico
- 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



Wikipedia, 2016

KNN – DISTANCIAS

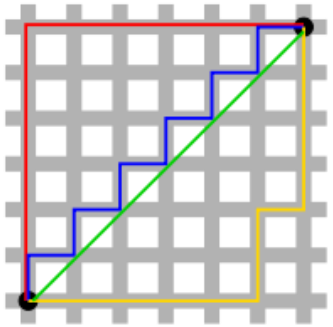
- Basado en una medida de **similitud** o **distancia** que hay que definir para encontrar los vecinos mas cercanos:
  - **Euclidiana**: tamaño del segmento linear que une las dos instancias comparadas.



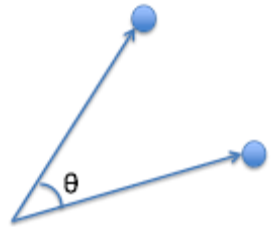
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

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

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



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



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

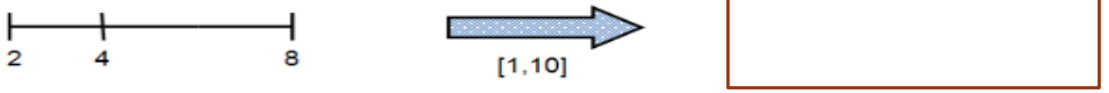
KNN - NORMALIZACIÓN

- Normalización [0, 1]

$$Y = \frac{X - mínimo\_original}{máximo\_original - mínimo\_original}$$

- Normalización [newmin, newmax] → Generalización, cambio de escala a otro intervalo cualquiera, no necesariamente [0,1], ni [oldmin, oldmax]

$$Y = \min + \frac{X - mínimo\_original}{máximo\_original - mínimo\_original}(\max - \min)$$

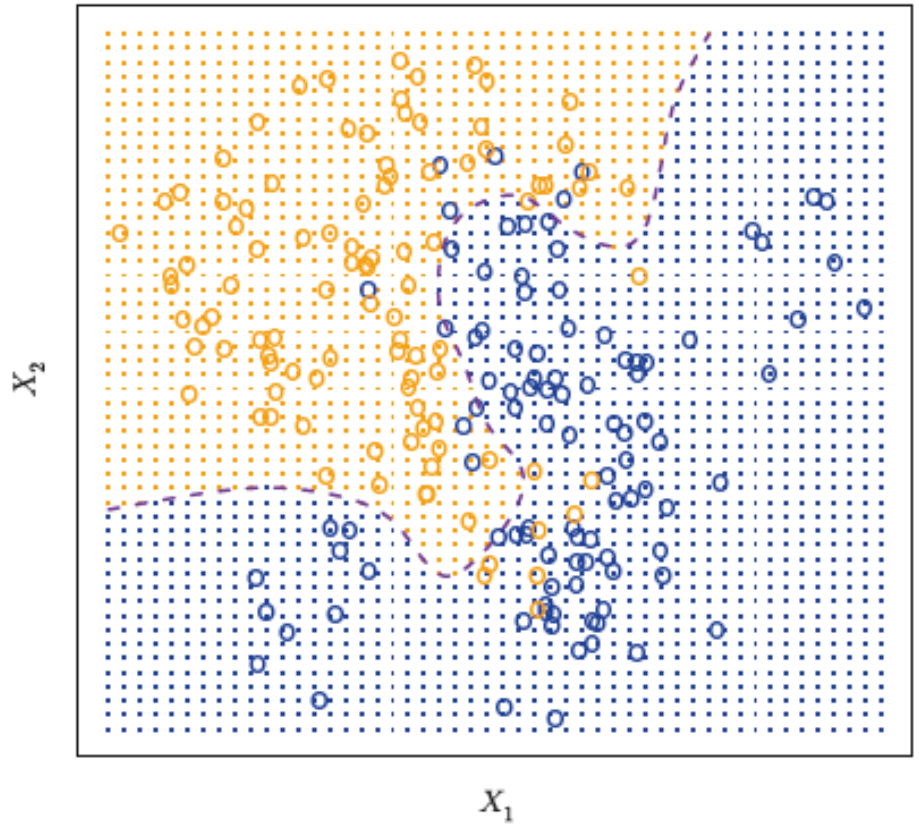


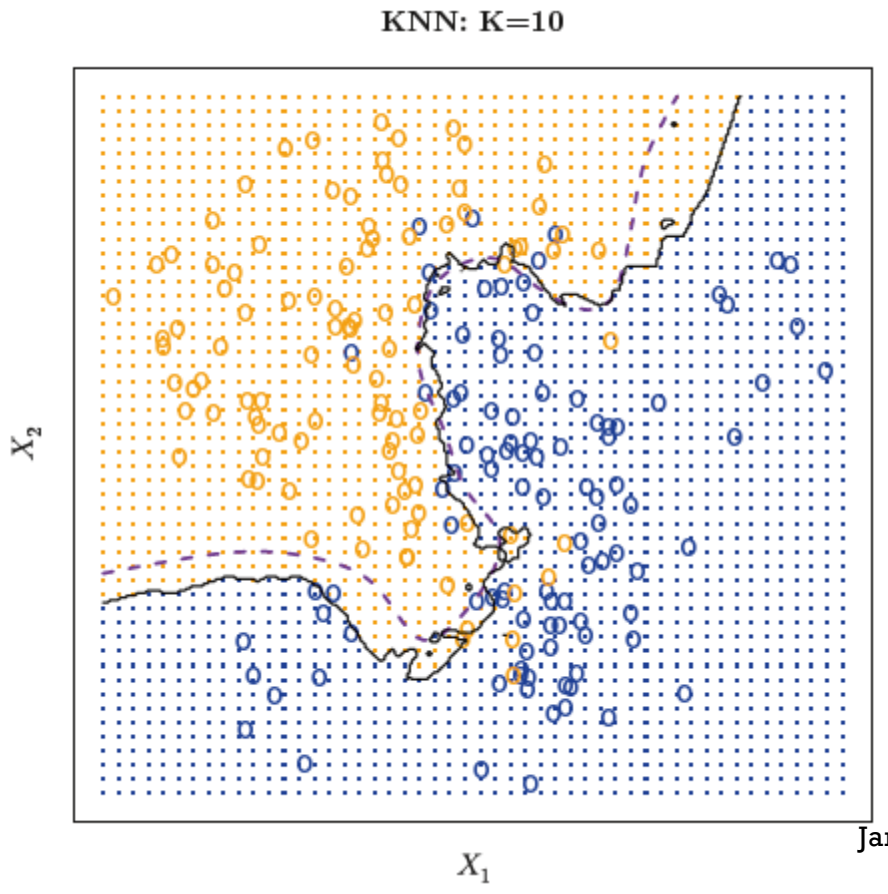
- Normalización z-score (estandarización)
  - Supuesto de distribución normal
  - Sea Z la representación estandarizada del dato
  - X la representación actual del dato
  - μ el valor promedio de los datos
  - σ la desviación estándar del campo

$$Z = \frac{X - \mu}{\sigma}$$

KNN – K

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

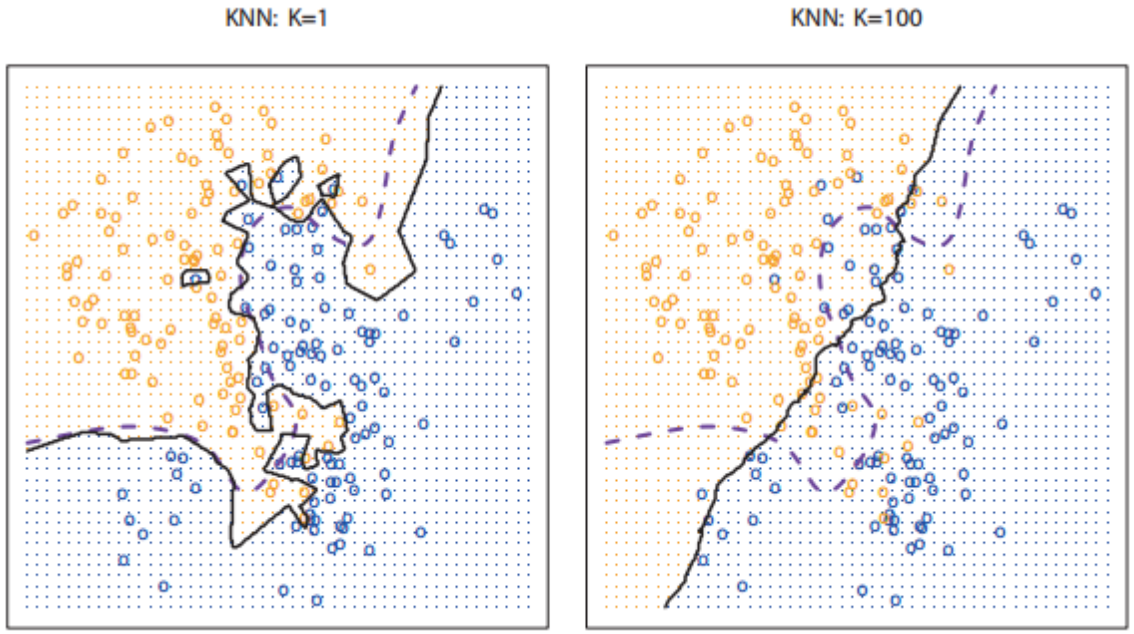




James et al, ISLR, 2013

KNN – K

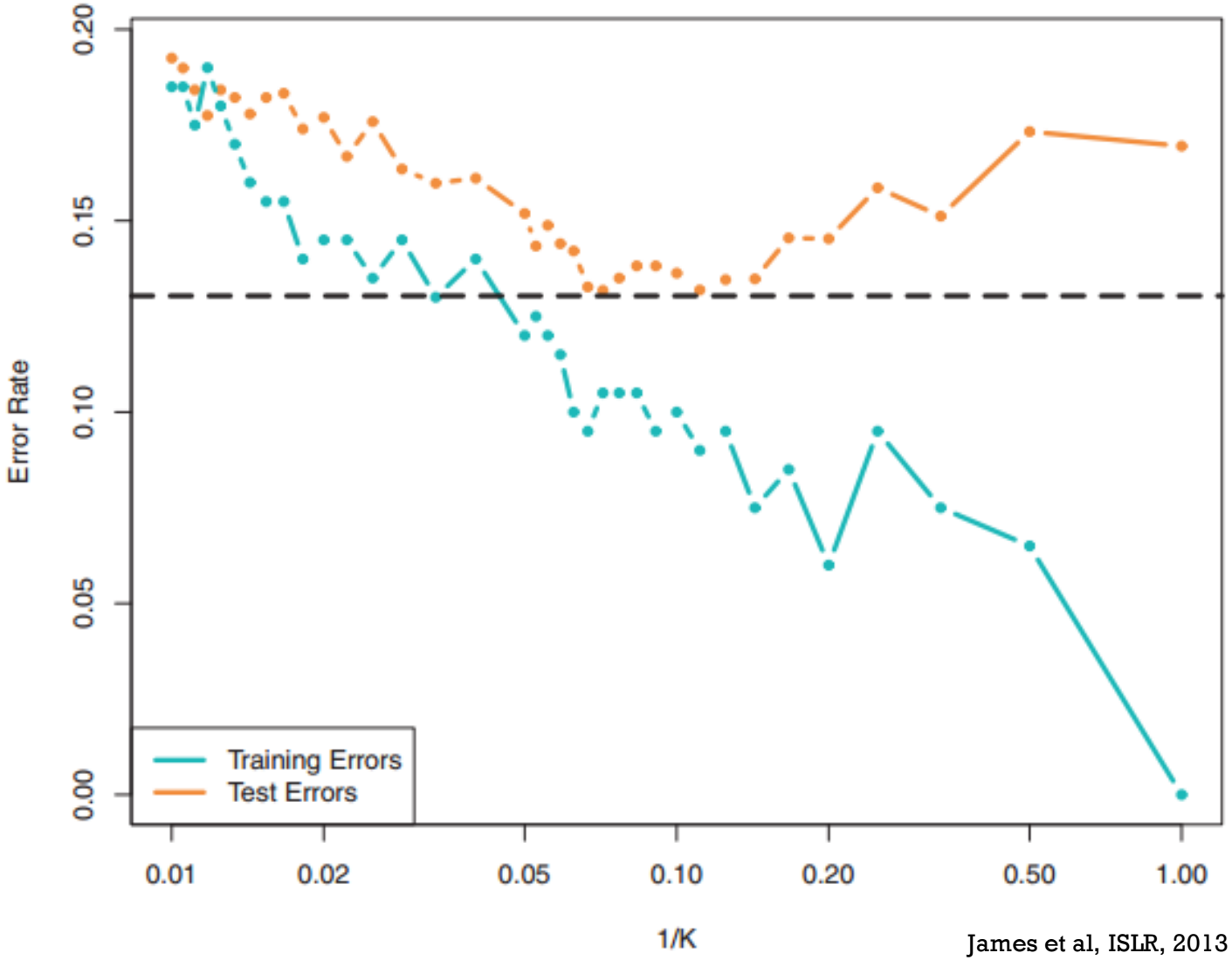
- **Parámetro K**
  - El resultado puede ser drásticamente diferente para diferentes valores de K
  - Un valor de K grande suavizará los límites entre clases/valores (alto sesgo, baja varianza)
  - Un valor de K pequeño resultará en límites muy flexibles (bajo sesgo, alta varianza)
  - El valor de K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



James et al, ISLR, 2013

KNN – K

- **Overfitting**: (sobre aprendizaje) a considerar en el momento de escoger el K.
- Modelos mas sencillos previenen el overfitting → K mas grandes
- Igualmente, cuidado con el **underfitting** (sub aprendizaje)

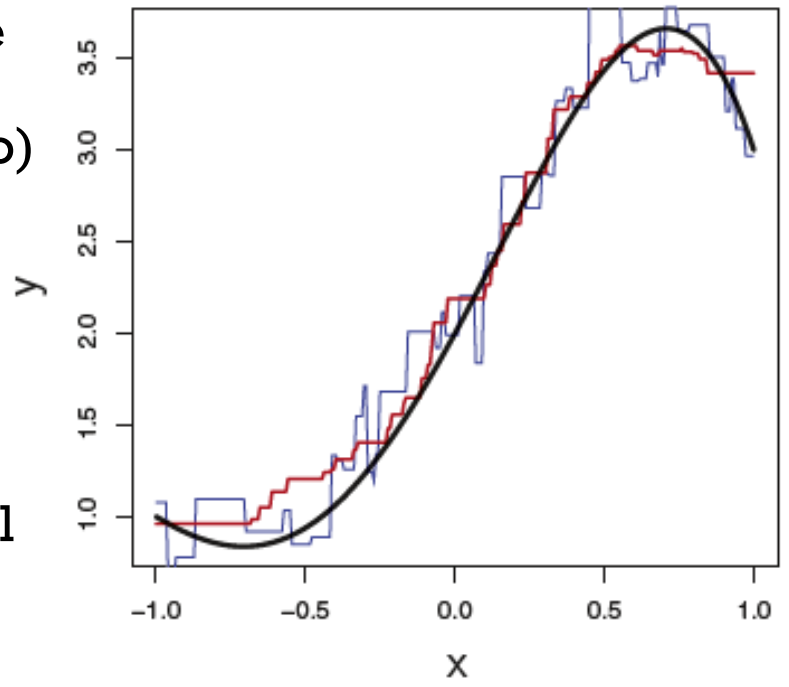


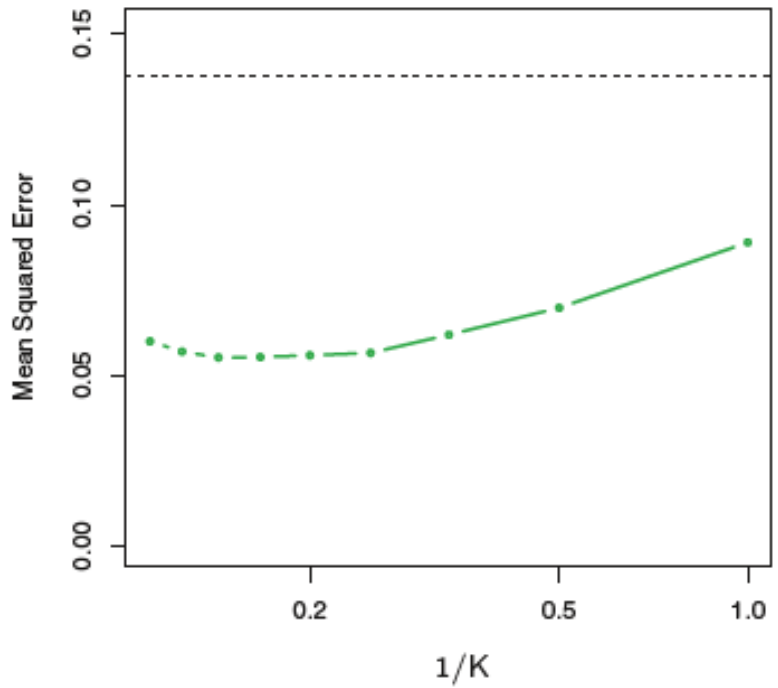
James et al, ISLR, 2013

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

KNN

Consideraciones:

- Perezoso (Lazy learning)
- No paramétrica y no lineal
- **Método local, no generalizable (no hay un modelo construido como tal)**:
  - 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**
- Muy sensible a la **unidad de medida** de los atributos (se deben **normalizar** las variables para evitar diferencias en sus importancias finales), 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**
- **Variaciones**: K-nn ponderado por la distancia, basado en un radio dado.

CNN (CONDENSED NEAREST NEIGHBORS)

- Dificultad de aplicación de KNN cuando se tienen **muchos registros**
- No todos los registros son necesarios para la correcta clasificación
- Aproximación de KNN utilizando un conjunto de datos reducido
- Escogencia de **prototipos** que permitan una clasificación con K=1 lo más parecida al resultado utilizando el dataset completo
- Algoritmo: Siendo **X** el conjunto de datos inicial y **U** el conjunto reducido:
  - Identificar todos los elementos x de **X** cuyo vecino más cercano sea de clase diferente
  - Retirar los x identificados (son prototipos) de **X** y agregarlos a **U**
  - Repetir hasta que no se agreguen más prototipos a **U**

TALLER DE CLASIFICACIÓN CON KNN

- DATASET: 150 ejemplos pertenecientes a 3 especies diferentes de la flor Iris
- 4 Atributos: largo y ancho del sépalo, largo y ancho del pétalo
- Reproducir el taller

Iris setosa

Iris versicolor

Iris virgínica

REFERENCIAS

- *Introduction to Statistical Learning with Applications in R (ISLR), G. James, D. Witten, T. Hastie & R. Tibshirani, 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/
- https://www.ibm.com/developerworks/library/os-weka1/

UNIVERSIDAD ICESI