

Machine Learning

Modelos de regresión 1

M.Sc. Angelo Jonathan Diaz Soto



Data&Analytics
Business Intelligence

Contenido



- Introducción
- Regresión lineal
- Regularización (Ridge, Lasso y Elastic Net) ■
- Regresión polinomial
- Regresión logística

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Introducción



❑ El Machine Learning (ML) es un campo de la informática que surgió de la investigación en inteligencia artificial.

- ❑ La fuerza del **Machine Learning** sobre otras formas de análisis radica en su capacidad para descubrir ideas ocultas y predecir resultados de futuros, insumos ocultos (generalización).
- ❑ A diferencia de los **algoritmos iterativos** donde las operaciones se declaran explícitamente, los algoritmos de **Machine Learning** toman prestados conceptos de la teoría de la probabilidad para seleccionar, evaluar y mejorar los modelos estadísticos.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

¿Qué es el Machine Learning?

El **Machine Learning** es una disciplina dentro del campo de la **Inteligencia Artificial** que, mediante algoritmos, proporciona a las computadoras la capacidad de



identificar patrones a partir de datos masivos para realizar **predicciones**. Este método de aprendizaje permite a los ordenadores realizar tareas específicas de forma **autónoma**, es decir, sin necesidad de ser programados.

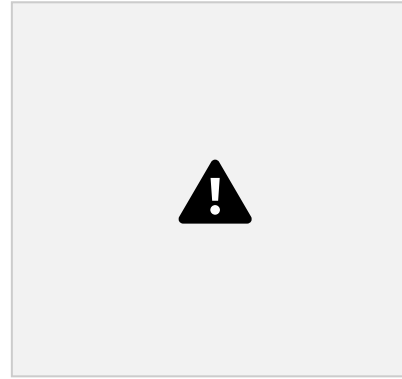
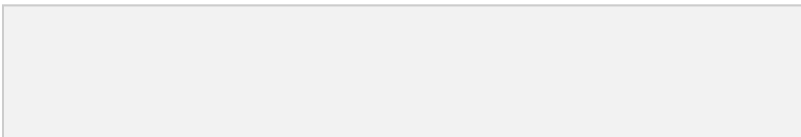


Figure 1: Machine Learning

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

¿Por qué es importante el machine learning?



El resurgimiento del  interés en el aprendizaje

basado en máquinas se debe a los volúmenes y variedades crecientes de datos disponibles, procesamiento computacional más económico y poderoso, y almacenaje de datos asequible.

Todas estas cosas significan que es posible producir modelos de manera **rápida** y **automática** que puedan analizar datos más grandes y complejos y producir resultados más **rápidos** y **precisos** – incluso en una escala muy grande.

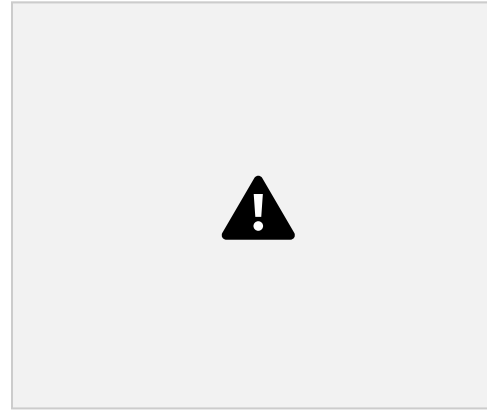
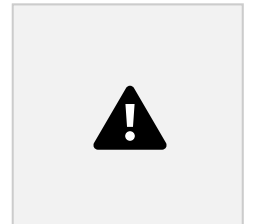


Figure 2: Big data y Machine Learning

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Proceso involucrados en el ML

Recopilación de datos





Preprocesamiento de datos

Elegir modelo

Modelo de tren

Modelo de prueba Sintonizar modelo

Despliegue para predicciones



(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

¿Quién lo utiliza?

La mayoría de las industrias que trabajan con grandes cantidades de datos han reconocido el valor de la tecnología del Machine Learning. Obteniendo **insights** de estos datos – a menudo en tiempo real.

Servicios financieros: Prevenir el fraude, identificar

oportunidades de inversión, identificar clientes con perfiles de alto riesgo o bien utilizar la ciber vigilancia para detectar signos de advertencia de fraude.

Atención a la salud:

Identificar tendencias o banderas rojas que puedan llevar a diagnósticos y tratamientos mejorado

Marketing y ventas: Los





sitios Web que le recomiendan artículos que podrían gustarle con base en compras anteriores

Gobierno: Ayudar a detectar fraude y minimizar el robo de identidad **Petróleo**

y gas: Cómo encontrar nuevas fuentes de energía. Análisis de minerales del suelo. Predicción de fallos de sensores de refinerías.

Transporte: Identificar rutas más eficientes y anticipar problemas potenciales para incrementar la rentabilidad.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com



Tipos de Machine Learning

Dos de los métodos de aprendizaje basado en

máquina más ampliamente
adoptados son aprendizaje **supervisado**
y aprendizaje **no supervisado**, pero
existen también otros métodos de
machine learning. Ésta es una
descripción de los tipos más populares.

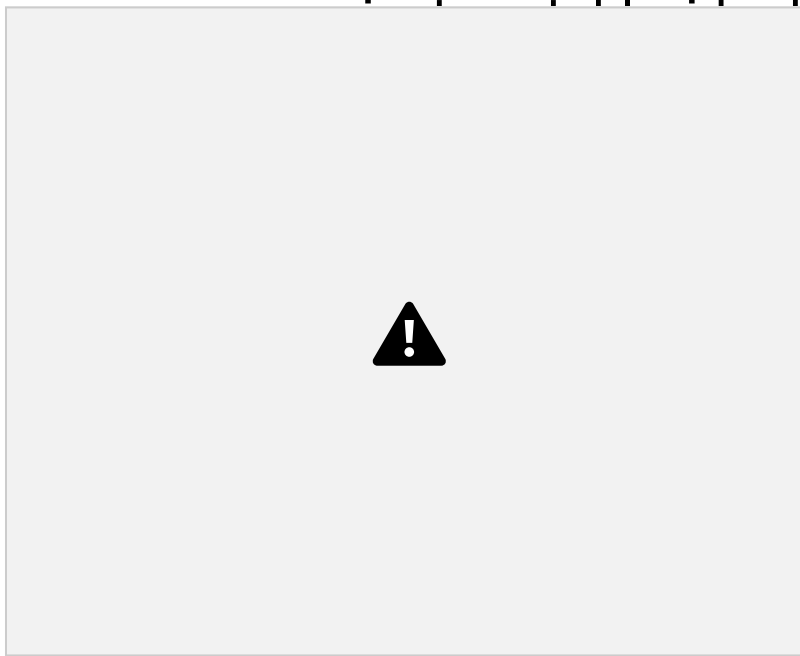
Figure 4: Métodos de Machine
Learning

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Tipos de aprendizaje



1 **Aprendizaje supervisado:** Comienza típicamente con un



de datos y una cierta comprensión de
s datos. Estos datos tienen
adas que definen el significado de los

datos.

2

Aprendizaje no supervisado: Se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar. El algoritmo debe descubrir lo que se muestra.

3

Aprendizaje por refuerzo: Se utiliza a menudo para robótica, juegos y navegación. Con el aprendizaje con refuerzo, el algoritmo descubre a través de ensayo y error qué acciones producen las mayores recompensas.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Machine Learning



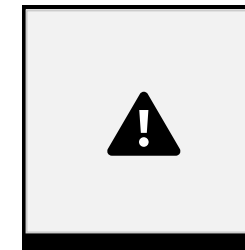
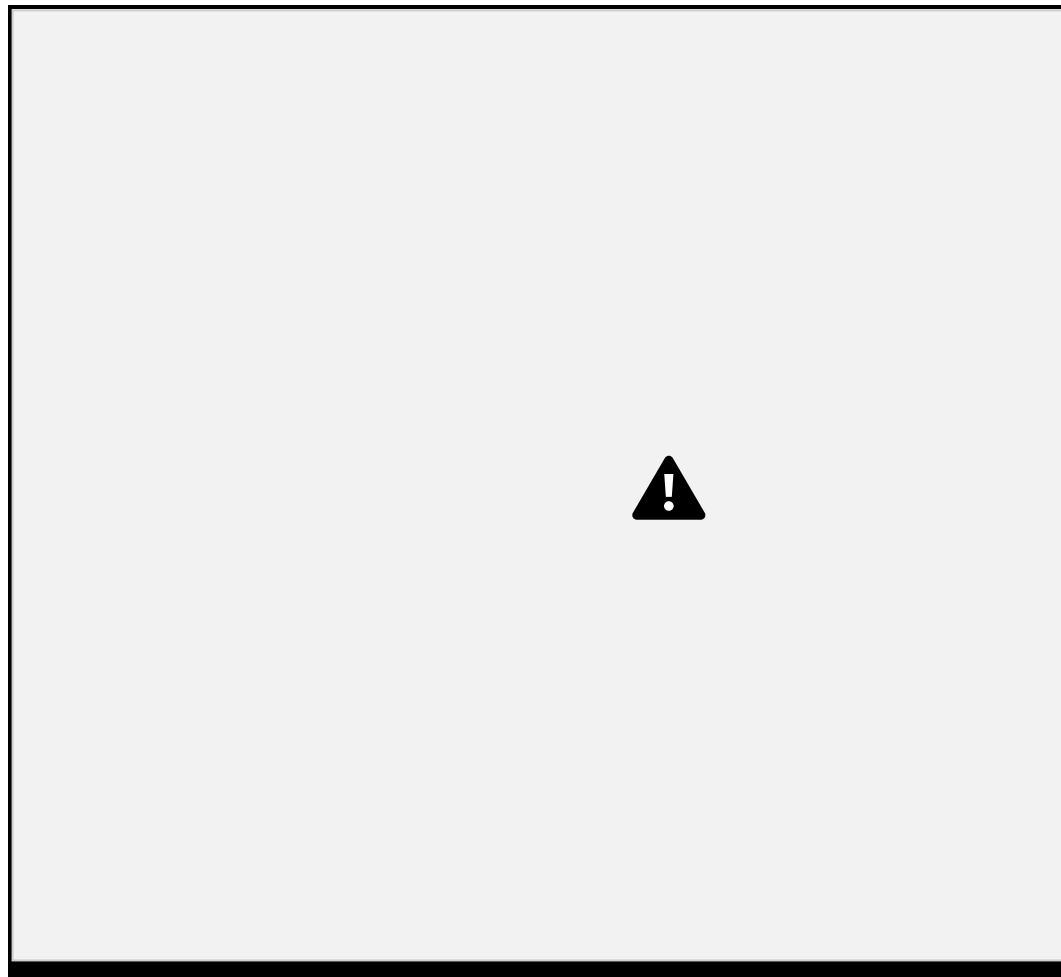


Figure 5: Esquema
generalizado sobre Machine
Learning

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com



Diferencias entre el Data Mining, ML y el Deep Learning



✓ **Minería de datos:** La minería de datos puede ser considerada un súper conjunto de muchos métodos diferentes para extraer insights de datos. Podría implicar métodos estadísticos tradicionales y machine

learning.

- ✓ **Machine Learning:** La diferencia principal con el **ML** es que, al igual que los modelos estadísticos, el objetivo es entender la estructura de los datos – ajustar distribuciones teóricas a los datos que son bien entendidos.
- ✓ **Deep Learning:** El aprendizaje profundo o mejor conocido como Deep Learning, combina avances en **poder de cómputo** y tipos especiales de **redes neuronales** para aprender patrones complicados en grandes cantidades de datos.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Data Mining vs Machine Learning



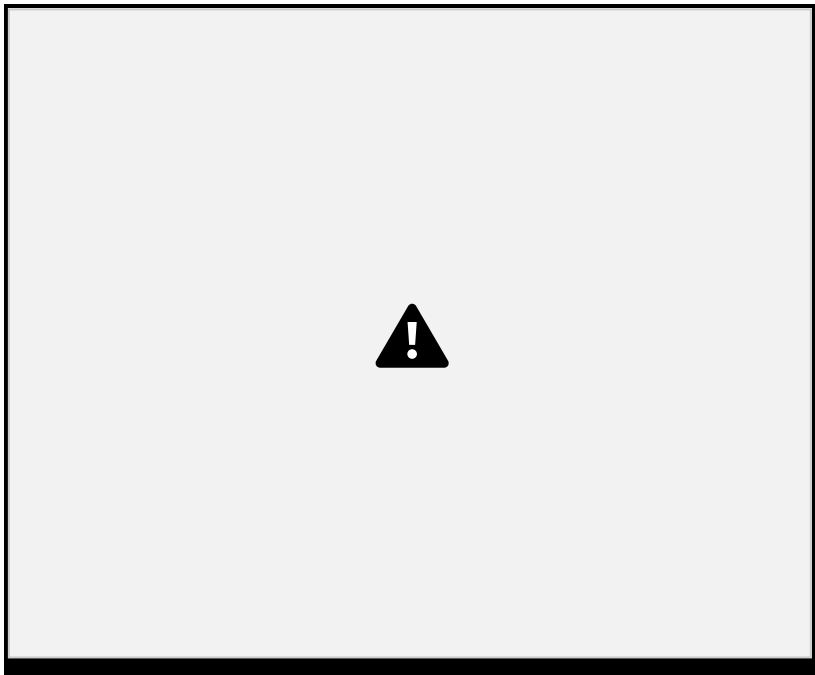
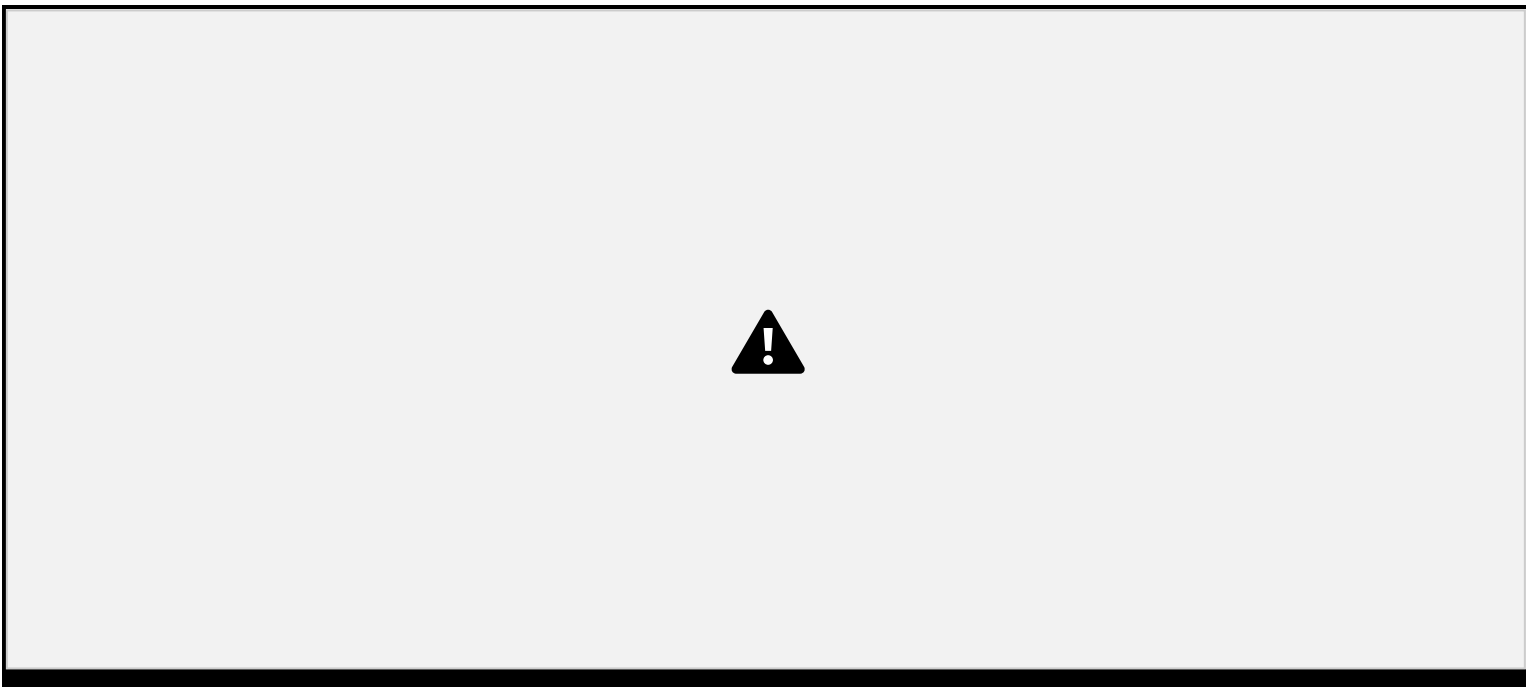
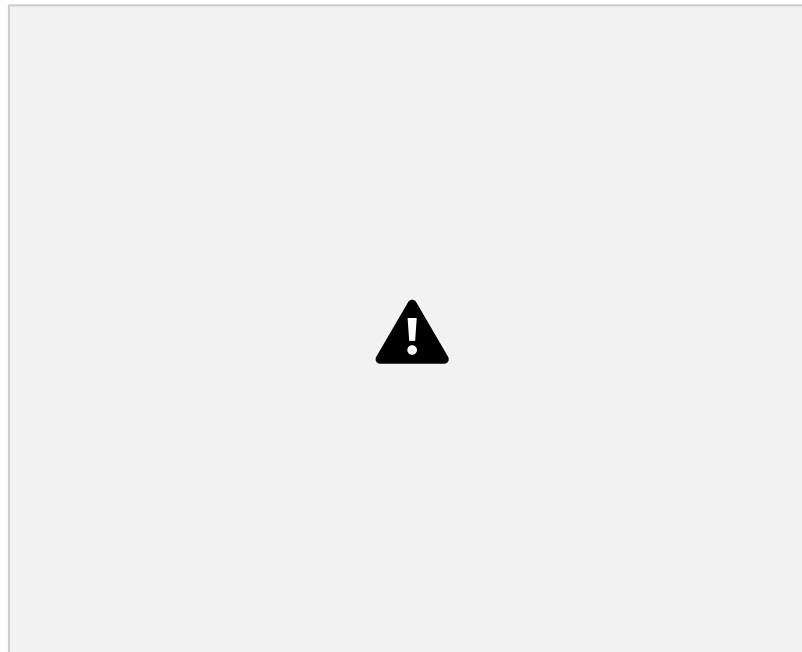


Figure 6: Data Mining vs Machine Learning

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Machine Learning y Otras disciplinas





disciplinas

Figure 7: Machine Learning vs Otras

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Tipos de aprendizaje





En Machine Learning hay dos tipos de aprendizaje: el **aprendizaje supervisado** y el **aprendizaje no supervisado**. ✓ El aprendizaje supervisado los modelos son entrenados a partir un conjunto de datos en el que

de la respuesta correcta es conocida.

- ✓ El aprendizaje no supervisado el conjunto de datos empleado en el entrenamiento no contiene la respuesta correcta.
- ✓ Los algoritmos de aprendizaje supervisado se dividen en dos tipos de problemas: problemas de **regresión** y problemas de **clasificación**.

Tipos de aprendizaje



Las técnicas
se dividen
en dos
grandes
grupos, el
aprendizaje

de Machine Learning



supervisado y el
aprendizaje
no supervisado.

Dentro del aprendizaje
supervisado, encontramos
modelos de
clasificación y modelos de

regresión. Figure 8: Técnicas de Machine Learning

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Modelo predictivo

Un modelo predictivo es un sistema que emplea datos y estadísticas para predecir resultados a partir de un



conjunto de datos. Estos modelos se pueden utilizar para predicciones de todo tipo; desde resultados médicos, deportivos económicos, ecológicos, etc.

Figure 9: Modelo predictivo

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Algoritmos de Machine Learning para regresiones



Regresión lineal

En el caso de los algoritmos de regresión, podemos decir que se

trata de un subcampo del aprendizaje

automático supervisado que tiene el fin de crear una metodología para **relacionar** un cierto número de características y una variable objetivo-**numérica**.

Regresión no lineal

Regresión logística

Árboles de regresión

Modelo KNN para regresión Random Forest para regresión Support Vector Regressor Regularización-Lasso Regularización-Elastic Net Regularización-Ridge Deep Learning.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Algoritmos de Machine Learning para regresiones.

Los algoritmos



de Regresión modelan la
relación entre distintas variables (features)
utilizando una medida de error que se
intentará minimizar en un proceso iterativo
para poder realizar predicciones “lo más
acertadas posible”.



(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

**Métricas de
Evaluación**





$$\sum_{j=1}^n$$

$$\hat{y}$$

$$\sum_{j=1}^n (\hat{y}_j - y_j)^2$$

$$\sum_{j=1}^n$$

3. Raíz del Error cuadrático medio (RMSE)

$$\sqrt{\frac{1}{n} \sum_{j=1}^n (\hat{y}_j - y_j)^2}$$

$$(\hat{y}_j - y_j)^2$$

$$\sum_{j=1}^n (\hat{y}_j - y_j)^2$$

$$\frac{SCT - SCR}{n} = SCE$$



Regresión Lineal



La **regresión lineal** es el modelo básico de modelamiento estadístico que consiste en determinar relaciones de dependencia de tipo lineal entre una variable dependiente o respuesta, respecto de una o varias variables explicativas o independiente.

$$y = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \epsilon_i(1)$$



Regresión Lineal



Supuesto lineal: La regresión lineal asume que la relación entre la variable dependiente e independiente es lineal. **Eliminar colinealidad:** La regresión lineal sobre-ajustará sus datos cuando tenga variables de entrada

altamente correlacionadas.

Distribuciones gaussianas: La regresión lineal hará predicciones más confiables si sus variables de entrada y salida tienen una distribución gaussiana.

Reescalar las entradas: La regresión lineal a menudo hará predicciones más confiables si reescala las variables de entrada usando la estandarización o la normalización.

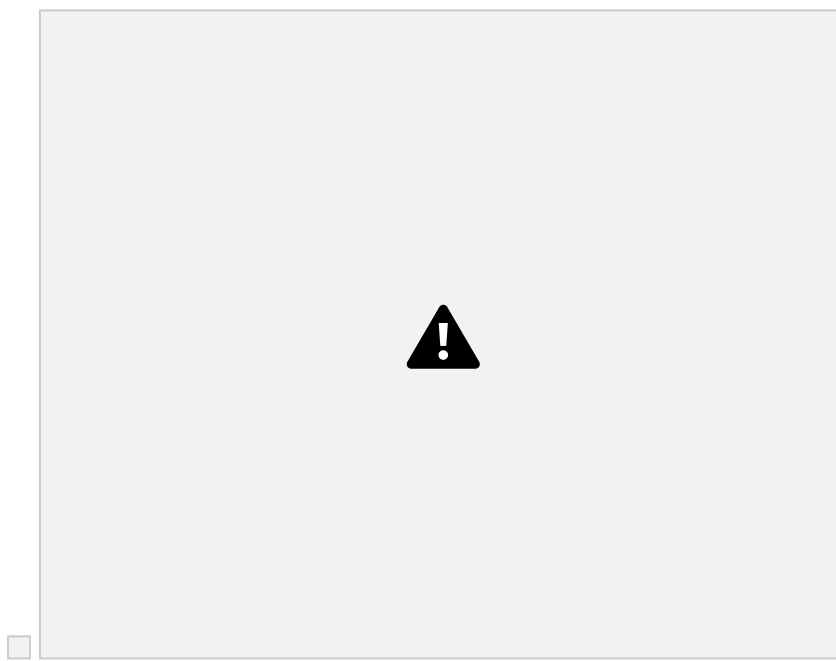
(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Regularización



El objetivo de la regularización es la reducción de la varianza del modelo agregando penalizaciones a los coeficientes estimados.

¿Por qué regularizar? La metodología de mínimos cuadrados ordinarios funciona



bien cuando se cumplen una serie de supuestos:

Relación lineal entre las variables

Normalidad multivariada

No autocorrelación

Homocedasticidad (Varianza constante del error)

Hay más observaciones que variables ($n > p$)

☐ multicolinealidad
No

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Regularización

Una buena opción es la regresión regularizada o también llamada regresión de penalización o métodos de contracción para controlar los parámetros estimados. La idea de la regresión regularizada es poner restricciones de magnitud a los coeficientes y contraerlos.



progresivamente hacia cero. Esta restricción ayuda a reducir la varianza del modelo.

La función objetivo de la regresión regularizada es similar a MCO pero con una penalización adicional:

$$\min\{RSS + p\} (2)$$

Al aplicar la penalización a la suma cuadrada de los errores se restringe el tamaño de los coeficientes tal que la única manera en que los coeficientes puedan aumentar es que la inclusión de una variable sea comparable con una disminución significativa del error.

(+51) 976 760 www.datayanalytics.com info@datayanalytics.com



Regresión de Ridge

Para la
estimación de los

coeficientes en mínimos cuadrados debemos minimizar la suma de los errores al cuadrado. Para generar una regresión tipo ridge agregamos la penalización y minimizamos la expresión:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Donde la primera expresión es la suma de los errores al cuadrado y es un parámetro que debe ser tuneado.



A diferencia
único cambio
la
absoluto en
Esto tiene

☐ No
muy

☐ En la
tienden
Lasso los

cero, lo que implica que la regresión Lasso tiene otro efecto y es que automáticamente depura las variables que no agregan poder predictivo al modelo.

de Ridge, matemáticamente el
es que ahora los coeficientes de
penalización están en valor
vez de elevados al cuadrado.
efectos distintos a la Ridge:

penaliza de la misma manera a los coeficientes
grandes.

regresión Ridge los coeficientes
hacia cero, en la regresión
coeficientes puede volverse





Regresión de Elastic Net

Es una
Lasso.
peso se



combinación de Ridge y
Se decide entonces qué
le da a cada método de
penalización y se
implementa la
regresión:



(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Regresión polinomial



La Regresión Polinomial es un caso especial de la Regresión Lineal, enriquece el modelo lineal al aumentar predictores adicionales,

obtenidos al elevar cada uno de los predictores originales a una potencia. Por ejemplo, una regresión cúbica utiliza tres variables, como predictores.

Este enfoque proporciona una forma sencilla de proporcionar un ajuste no lineal a los datos.



(+51) 976 760 www.datayanalytics.com info@datayanalytics.com

Regresión Logística

En muchas investigaciones se está interesado en relacionar una variable dependiente **dicotómica** y variables independientes de tipo categórico y cuantitativo. Recordemos que una variable



dicotómica



es una variable que puede tomar sólo uno de dos valores mutuamente excluyentes, por lo general se codifican como $Y = 1$ para éxito e $Y = 0$ para fracaso.



(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Aplicaciones de clasificación binaria

■ **Detección**
spam o no.

de spam: Predecir si un correo electrónico es

■ **Fraude de tarjeta de crédito:** Predecir si una transacción de tarjeta de crédito es fraudulenta o no.

Marketing: Predecir si un usuario determinado comprará o no ■ un producto de seguro.

Salud: Predecir si una masa de tejido es benigna o maligna.

■ **Banca:** Predecir si un cliente incumplirá un préstamo.

■

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

¿Qué es la regresión logística?



■ La regresión logística binaria es un tipo especial de regresión donde la variable de

respuesta binaria está relacionada con un conjunto de variables explicativas que pueden ser discretas y o continuas.

- En la regresión logística, la probabilidad de que la respuesta tome un valor particular se modela en función de la combinación de valores tomados por los predictores.

La regresión logística binaria pertenece a la familia de los

- modelos lineales generalizados (GLM).

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Supuestos del modelo

- Los modelos
una

lineales generalizados (GLM) no asume

relación lineal entre variables dependientes e independientes. Sin embargo, asume una relación lineal entre la función de enlace y las variables independientes en el modelo logit.

La variable dependiente no necesita tener distribución normal.

■ No utiliza MCO (Mínimos cuadrados ordinarios) para la estimación de parámetros. En su lugar, utiliza la estimación de ■ máxima verosimilitud (EMV).

Los errores deben ser independientes pero no distribuidos normalmente.

■

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

**Estimando las
probabilidades**



□ Dado el vector de entrada $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$ La probabilidad estimada del modelo de regresión logística está dada por

$$P = \frac{\sigma(X_i \beta)}{1 + \sigma(X_i \beta)}$$

Donde: $\sigma(t) = 1 / (1 + e^{-t})$
Función logística

□ Predicción del modelo de regresión logística:

$$Y = 0, \text{ si } P < 0.5$$

$$Y = 1, \text{ si } P > 0.5$$



El score t es frecuentemente llamado logit, esto proviene del hecho que $t = \log(P / (1-P))$.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Función de costo de la regresión logística





(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Modelo logit binario (Análisis estadístico)

Regresión Logística Binaria, desarrollada por **David Cox en 1958**, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una o varias variables cuantitativas y/o cualitativas.

Una de las principales aplicaciones de la regresión logística es la ☐ de clasificación, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

Es importante tener en cuenta que, aunque la regresión logística permite clasificar, **se trata de un modelo de ☐ regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo.**

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Definición del modelo logit

- ☐ **Componente aleatorio:** Sean Y_1, \dots, Y_n v.a. dicotómicas independientes. Asumiendo que $y_i = 1$ tiene probabilidad π_i y $y_i = 0$ con probabilidad $1 - \pi_i$:





$$y_i \sim \text{Bernoulli}(\pi_i)$$

Componente sistemático:



$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

$$x_i = (1, x_{i1}, \dots, x_{ip})'$$

donde η_i es denominado como predictor lineal y $x_i = (x_{i1}, \dots, x_{ip})'$ es un vector de covariables, donde x_{i1} igual a 1 corresponde al intercepto.

■ Función de Enlace:

$$g(\pi_i) = \eta_i$$

donde $g(\cdot)$ es una función monótona y diferenciable.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Definición del modelo logit



■ Función de Respuesta:



donde $h(\cdot)$ es una función de distribución acumulativa monótona estrictamente creciente sobre la

recta de los números reales. Esto asegura que $h(\eta) \in [0, 1]$ y $g =$

h^{-1}

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Modelo logit: Funciones de enlace



Enlace Probit

$$\Phi(\pi(x))^{-1} = \beta_1 + \beta_2 x_2 + \dots +$$



$\beta_p x_p$

donde $\Phi(.)$ es la f.d.a. de la normal estándar.

■ **Enlace log-log complementario (cloglog)**

$$\log\{-\log(1 - \pi(X))\} = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Modelo logit: Funciones de enlace





(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Gráfico de una función

logit



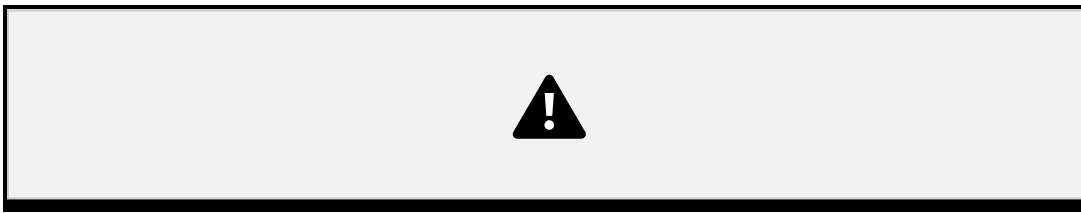


Figure 3: Función logit con dos variables explicativas continuas

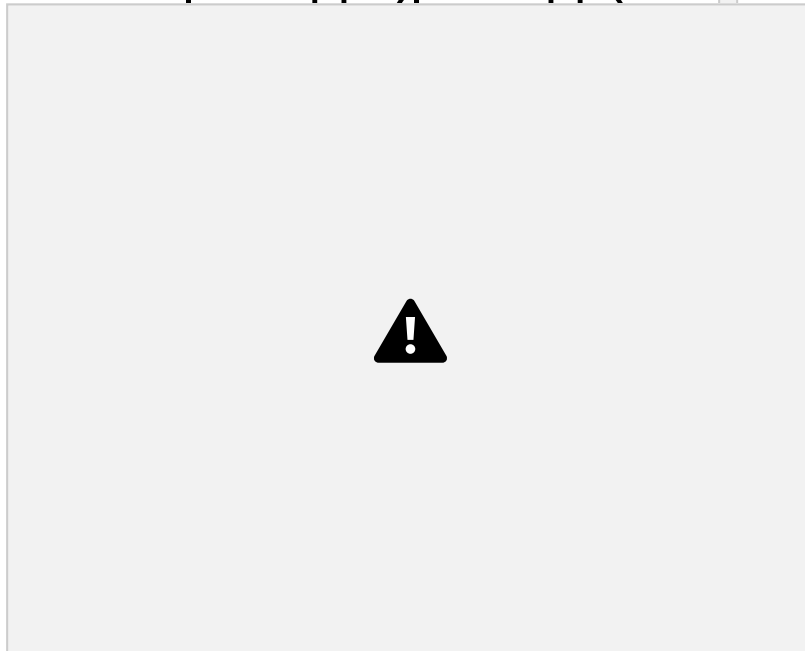
(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Definición del modelo logit





donde $x = (1, x_2, \dots, x_p)^T$ contienen los valores observados de las variables explicativas.



Usando la transformación con la



$$1 - \pi(X) = \exp(\beta_1) \cdot \exp(\beta_2 x_2) \cdot \dots \cdot \exp(\beta_p x_p)$$

$$\exp(\beta_1) \cdot \exp(\beta_2 x_2) \cdot \dots \cdot \exp(\beta_p x_p)$$

lo cual implica que los efectos de las covariables afectan los odds $\pi(X)$



Estimación de parámetros

Considere un modelo
binaria



$$F(t) = \frac{1}{1 + e^{-t}}$$

entonces
tenemos que la **función de verosimilitud** será dada
por





(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Estimación de parámetros



Considerando como distribución a priori para β una **distribución normal multivariada** con vector de medias $\mathbf{0}$ y matriz de covarianza Σ .



Luego la **distribución a posteriori** viene dada por $y_i x_i^T \beta$

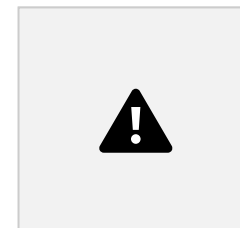
$$f(\beta | y) \propto \prod_{i=1}^T \frac{e^{\beta y_i}}{1 + e^{\beta}} \prod_{i=1}^T e^{-\beta y_i}$$

Como podemos observar en esta caso la distribución a posteriori no presenta una forma conocida por lo que se hace necesario utilizar métodos **MCMC**.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Evaluación del modelo: Caso explicativo

de Wald



1 Test



Test Hosmer-Lemeshow (bondad de ajuste) Test de
significancia global

Criterios de Información (AIC, BIC, etc.)

Un Test de bondad de ajuste, en general, lo que hace es

comprobar si el modelo propuesto puede explicar lo que se observa.

El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Evaluación del modelo: Caso Predictivo

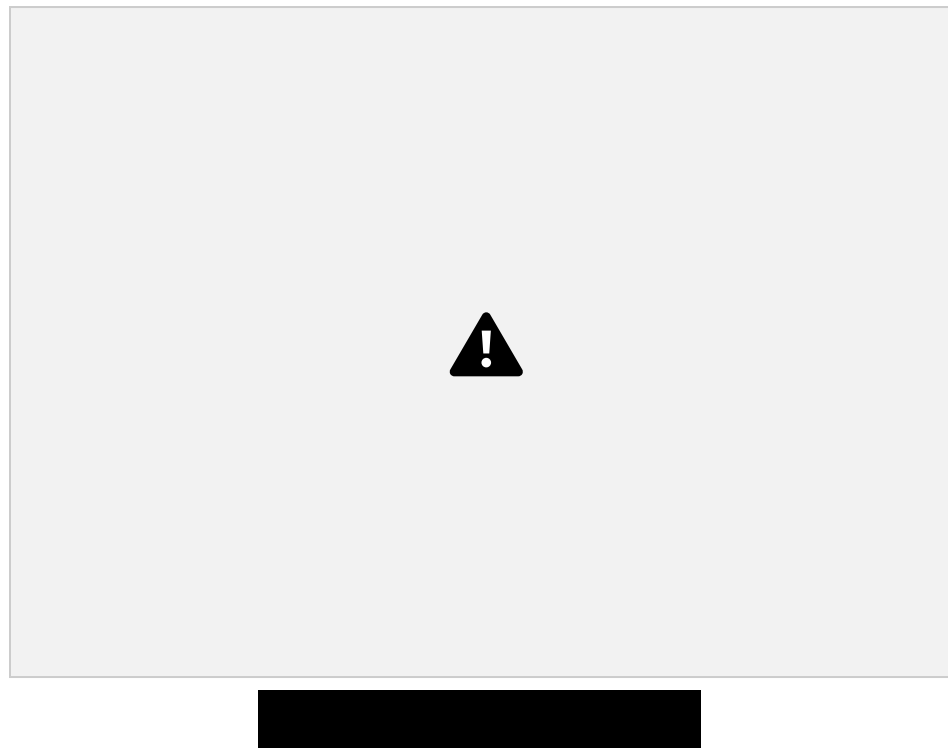
Matriz de confusión Curva

ROC



AUC

 Otras métricas (Kappa)



(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Referencias Bibliográficas

■ Agresti, Alan. (2002). Categorical Data Analysis. New York: Wiley-Interscience.

■ Hosmer, David W.; Stanley Lemeshow (2000). Applied Logistic Regression, 2nd ed. New York; Chichester, Wiley.

□). Handbook of the Logistic
Dekker, Inc. ISBN 978-0824785871.



Albert, J. (2007). Bayesian Computation with R, New York:



Springer.

Robert, C. P. y Casella, G. (1999). Monte Carlo Statistical



Methods, New York: Springer.

Casella, G. y Berger, R. L. (2002). Statistical Inference,



Duxbury: Pacioc Grove, CA.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com





(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com