Taller #10

Business Intelligence

Carlos Reveco
creveco@dcc.uchile.cl

Cinthya Vergara
cvergarasilv@ing.uchile.cl

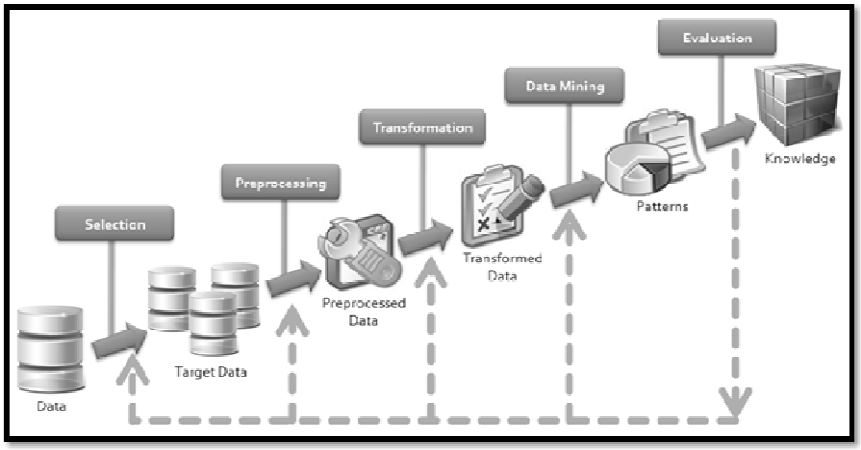
Agenda

- Taller #10 - Regresiones
  - Métodos de Regresión
    - Regresión Lineal
    - Regresión Polinomial
    - Support Vector Regression
    - Redes Neuronales para Regresión
  - Errores en métodos de regresión
  - Actividad práctica:
    - GermanCredit dataset

Proceso KDD

3

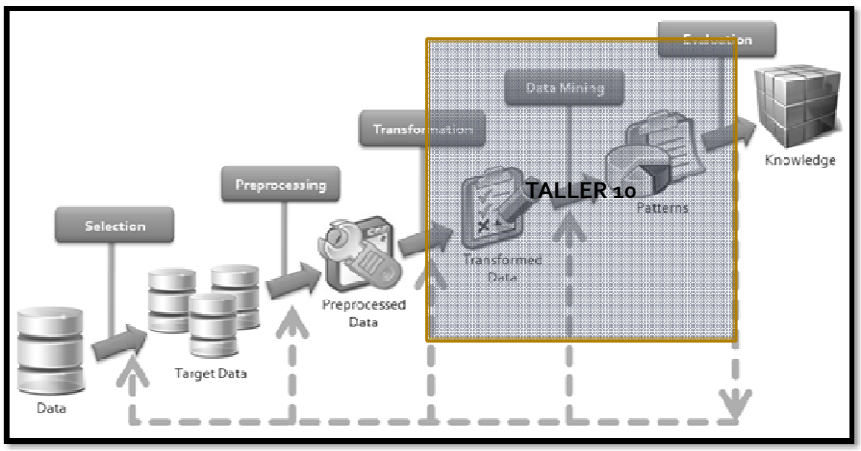
Proceso KDD



Knowledge **D**iscovery in **D**atabases → **KDD**

4

Proceso KDD



Knowledge **D**iscovery in **D**atabases → **KDD**

5

Regresión Lineal y Polinomial

6

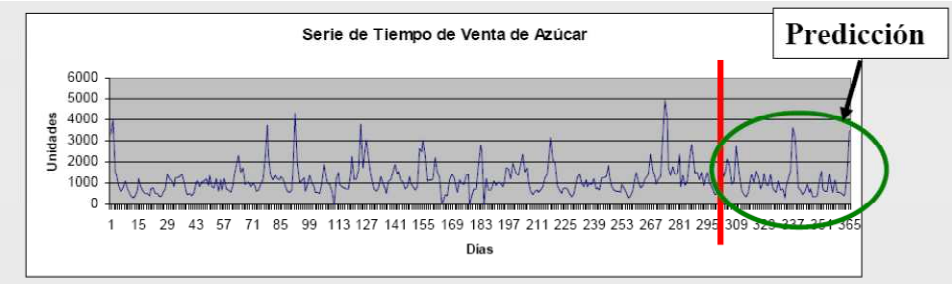
Regresión

- La Regresión (o estimación) trata con problemas donde el valor a clasificar puede tomar valores en un **rango continuo** (ingresos, balance de la tarjeta de crédito, demanda, etc.)

- Ejemplos
  - Estimar el número de hijos de una familia.
  - Estimar el tiempo de vida de un cliente.
  - Estimar la variación de un precio de un producto en el siguiente período.

Procedimiento

- Estudiar el comportamiento temporal y dinámico de alguna variable.
- Encontrar la **mejor función** que describa este fenómeno.
- Aplicar la función encontrada a la predicción de nuevos valores de la serie.



Regresión General

$$Y_i = f(X_i) \mid Y_i \in \mathbb{R}, X_i \in \mathbb{R}^N$$

- Tipos de variables
  - Dependiente (Y): variable a ser explicada o endógena.
  - Independientes (X): Variables explicativas o exógenas.
  - Ficticias: dummies o cualitativas.
    - Se utilizan para incluir variables nominales o para incorporar tendencias u otros factores.

Regresión Lineal

- Regresión Simple y Múltiple
  - **Simple**: 1 variable explicativa y 2 parámetros (una constante más una variable).
  - **Múltiple**: Más de una variable explicativa y/o más de 2 parámetros.
- Regresión Lineal Simple:

$$Y_i = a + b X_i + e_i$$

Regresión Muestral

- En la práctica no contamos con todos los datos, sino que con una muestra de ellos.
- Formalmente

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

  - La **Función de Regresión Muestral** (FRM) es una aproximación de la **Función de Regresión** (FR) poblacional.
  - La intención es hacer esta función lo más cercana a los datos originales como sea posible.

Regresión Muestral [2]

- Estimación de en base a **Mínimos Cuadrados Ordinarios** (MCO)
- Definimos la Función de Regresión Muestral de la forma

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

- Estimadores (construidos a partir de la muestra) de α y β :

$$\hat{\alpha} \text{ y } \hat{\beta}$$

- Residuos (errores):

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$$

Mínimos Cuadrados Ordinarios

- Criterio de MCO: encontrar los estimadores que minimizan la **suma del cuadrado de los residuos**.

$$\underset{\{\alpha,\beta\}}{MIN} \quad \sum a_i^2$$

- Valor de los estimadores:

$$\beta = \frac{\sum_{i=1}^{n}(X_i - X)(Y_i - Y)}{\sum_{i=1}^{n}(X_i - X)^2} \qquad \alpha = Y - \beta X$$

Supuestos Regresión Lineal

1. El **valor esperado de error** en cualquier observación deberá ser **nula**.
2. La **varianza poblacional** del error es **constante** para todas las observaciones
3. La **distribución de la observación** i es **independiente** de la observación j
4. El **error** está distribuído **independientemente** de la variable explicatoria.
5. Usualmente se asume que el **error** está **distribuído de forma normal**.

1. $E(u_i) = 0.$
2. $Var(u_i) = \sigma^2.$
3. $Cov(u_i, u_j) = 0.$
4. $Cov(u_i, x_i) = 0.$

$$u_i \sim NID(\mu, \sigma^2)$$

Ajuste Regresión Lineal

- R^2 representa la suma total de cuadrados explicada por regresión.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

- R^2 está entre o y 1.
  - Si R^2 =1 todos los puntos están sobre la recta estimada. Si R^2 =o, la recta no explica nada.
- Nos permite medir el ajuste de la regresión a los datos.

Propiedades R^2

- Invariante frente a cambios de escala (multiplicar por constantes).
- R^2 representa la **cantidad de varianza** explicada por la regresión.

$$r(Y,\hat{Y})^2 = \frac{\sum_i(\hat{Y}_i - \bar{Y})^2}{\sum_i(Y_i - \bar{Y})^2}$$

- R^2 aumenta a medida que hay más variables en la regresión.
  - Solución: R^2 ajustado:

$$R^2_{aj} = R^2 - \frac{k-1}{n-k}(1 - R^2) \qquad \begin{matrix} n: datos \\ k: atributos \end{matrix}$$

16

Regresión Polinomial

- Una recta no siempre se ajusta bien a los modelos.
- La regresión es lineal en los parámetros **no en las variables**.
- **Solución**: Incorporar un polinomio de las variables en la regresión.

$$y(x_i) = \alpha_1 x_i + \alpha_2 {x_i}^2 + \alpha_3 {x_i}^3 + \cdots + u_i$$

- Las medidas de error son equivalentes.

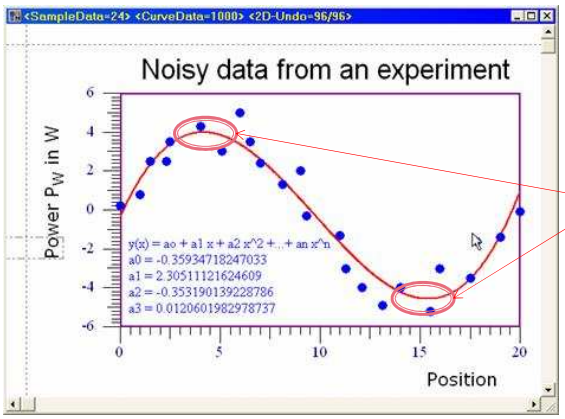
17

Regresión Polinomial (2)

- El problema con la regresión polinomial es que no existe una forma clara de determinar el **grado del polinomio**.
- **Solución**: Graficar el problema (inspección visual).
  - Cantidad de **puntos de inflexión** determinan el grado.
  - Punto de inflexión: Cambios de sentido en la curva.

18

Regresión Polinomial (3)



- Polinomio de grado 3.
- Son dos puntos de inflexión, por lo que el grado corresponde a ellos más uno.

19

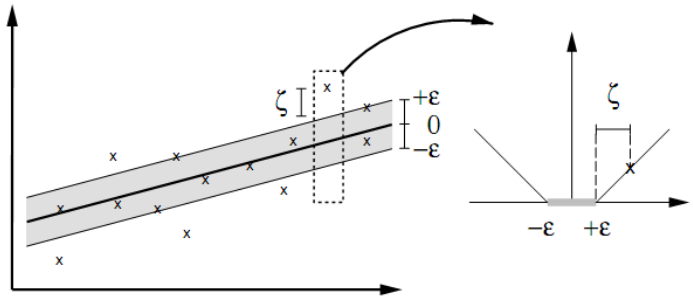
Support Vector Regression y Redes Neuronales

20

Support Vector Regression
SVR

- Es posible **generalizar** el concepto de SVM para clasificación y llevarlo a regresión.
- Se tolera una desviación máxima de ε.
- Se busca una función que se ubique en el medio del "tubo" formado esta desviación máxima permitida. Esto se consigue minimizando la norma euclidiana de w.
- Se penalizan las observaciones que violan la desviación máxima de la misma manera que se penalizan los errores en SVM (holgura en la restricción, se penalizan en la función objetivo).

Support Vector Regression
SVR [2]



$$\text{minimize} \quad \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)$$

$$\text{subject to} \quad \begin{cases} y_i - \langle w, x_i\rangle - b & \leq \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases}$$

Support Vector Regression
SVR - Dual

$$\text{maximize} \quad \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{\ell}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i^*) \end{cases}$$

$$\text{subject to} \quad \begin{cases} \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) & = & 0 \\ \alpha_i, \alpha_i^* & \in & [0, C] \end{cases}$$

$$w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)x_i \qquad f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b.$$

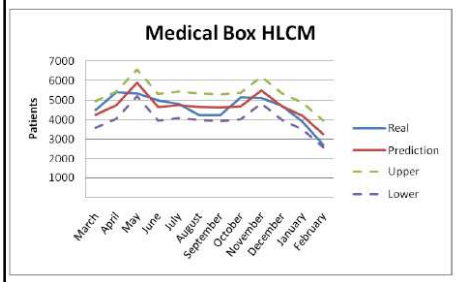
- Para incorporar función Kernel, basta con reemplazar $< x, x_i >$ por $k(x, x_i)$.

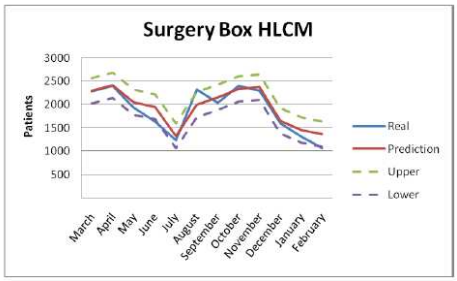
Support Vector Regression
Aplicación del Modelo

- La función de decisión corresponde a:
  - Sin kernel: $f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b.$
  - Con kernel: $f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)k(x_i, x) + b$
- Nuevamente, es complejo computarlo.
- Importante: decidir el kernel a utilizar.
  - RBF sigue siendo una apuesta buena.

24

Support Vector Regression
Aplicación del Modelo





"Demand Forescating and Capacity Planning for Hospitals"
O. Barros et al. (submitted)

25

Redes Neuronales
Regresión

- Las redes neuronales son también pueden ser utilizadas para regresión.
- (Recuerdo) Las redes neuronales son **aproximadores universales**.
  - Una función **continua y acotada** puede ser aproximada con cualquier tolerancia por una red neuronal con **una sola capa oculta**.
  - Una función **continua general** puede ser aproximada con cualquier tolerancia por una red neuronal con **dos capas ocultas**.

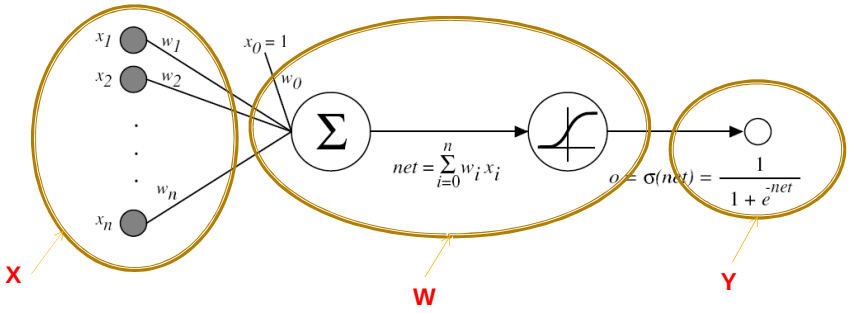
26

Redes Neuronales
Regresión [2]

- Una **red neuronal** es una función

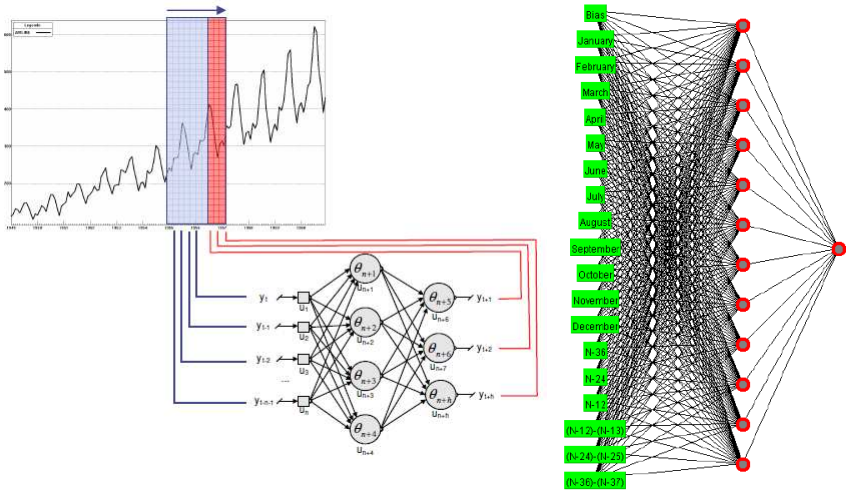
$$Y = F(X, W)$$

donde **Y** es el vector formado por las salidas de la red, **X** es el vector de entrada a la red, y **W** es el conjunto de todos los parámetros de la red (pesos y umbrales), y **F** una función continua no lineal.



27

Redes Neuronales
Regresión [3]



28

Redes Neuronales
La Capa de Salida

- Cuando se utiliza para regresión, una red neuronal tiene **una única neurona en la capa de salida**.
- La decisión importante es qué función de salida utilizar.
  - Debe estar en el rango del problema.
  - Debe incorporar los potenciales valores de forma correcta.

29

Performance de Métodos Regresión

30

Medidas de Performance
Salida Continua

$T = \{t_i = \text{ comportamiento } \mathbf{observado} \text{ para la observación } i, i = \{1, ..., n\}\}$
$O = \{o_i = \text{ comportamiento } \mathbf{estimado} \text{ para la observación } i, i = \{1, ..., n\}\}$

- Mean Squared Error (MSE):

$$MSE(T,O) = \frac{1}{n}\sum_{i=1}^{n}(t_i - o_i)^2$$

  - Ampliamente utilizado para toda regresión.
  - Cota teórica importante (Cramer – RAO) basada en él.

- Mean Absolute Error (MAE):

$$MAE(T,O) = \frac{1}{n}\sum_{i=1}^{n}|t_i - o_i|$$

  - Tiene dimensiones de salida original.

31

Medidas de Performance
Salida Continua [2]

$T = \{t_i = \text{ comportamiento } \mathbf{observado} \text{ para la observación } i, i = \{1, ..., n\}\}$
$O = \{o_i = \text{ comportamiento } \mathbf{estimado} \text{ para la observación } i, i = \{1, ..., n\}\}$

- Root Mean Squared Error (RMSE):

$$RMSE(T,O) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - o_i)^2}$$

  - Ventajas de MSE, pero con dimensiones de input.

- Mean Absolute Percentage Error (MAPE):

$$MAPE(T,O) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{t_i - o_i}{t_i}\right|$$

  - Particularmente importante en series de tiempo.

- Coeficiente de Correlación: $\rho(T,O) = \frac{cov(T,O)}{\sigma_T \cdot \sigma_O}$

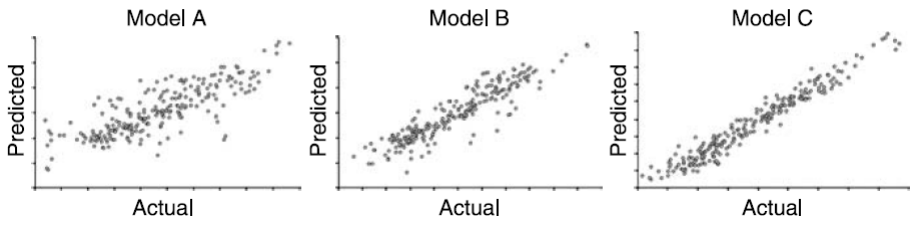
32

Medidas de Performance
Salida Continua [3]

- Existen también medidas "relativas".
  - Estas medidas utilizan la media como predictor simple y comparan en base a "cuánto mejor" es el nuevo modelo.
- Relative Square Error: $\frac{\sum_{i=1}^{n}(o_i - t_i)^2}{\sum_{i=1}^{n}(\overline{t_i} - t_i)^2}$
- Relative Absolute Error: $\frac{\sum_{i=1}^{n}|o_i - t_i|}{\sum_{i=1}^{n}|t_i - \overline{t_i}|}$

33

Comparación de Medidas
Salida Continua



	Model A	Model B	Model C
Mean square error	1.42	0.622	0.176
Mean absolute error	0.874	0.579	0.333
Relative square error	0.342	0.161	0.051
Relative absolute error	0.52	0.346	0.212
Correlation coefficient	0.811	0.916	0.974
Square correlation coefficient	0.658	0.839	0.949

34

Taller 10

35

Taller 10
Regresiones

- Objetivos:
  1. Desarrollar experimentos con regresiones.
     1. Regresión lineal, Regresión Polinomial
     2. Support Vector Regression
     3. Artificial Neural Network Regression
  2. Revisar métricas de evaluación para comparar técnicas de regresiones.
- Base de datos a utilizar:
  - **German Credit** (18 atributos, 800 observaciones)

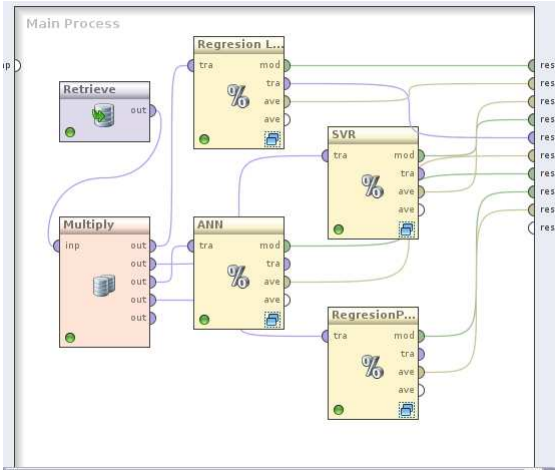
36

GermanCredit dataset
Descripción

- La idea es predecir el monto del crédito.
  - Objetivo: Predecir el monto del crédito a ser solicitado por los clientes.
- Atributos:
  - 18 atributos.
  - Sexo, edad, plazo, ahorro, tiene o no tiene aval, tiene o no tiene casa, entre otros.
  - Base de datos pre-procesada (sin missing values, no es necesaria una selección de atributos)

37

GermanCredit
RapidMiner 5.0 proceso



38

Taller #10

Business Intelligence

Gastón L'Huillier Ch.
glhuilli@dcc.uchile.cl

39