Aprendizaje Automático
Segundo Cuatrimestre de 2016

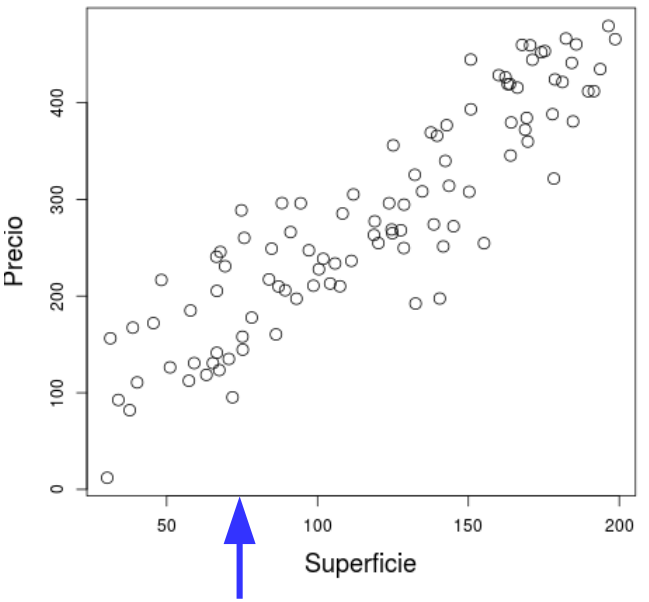
Regresión Lineal

Clase dada en el pizarrón, basada fuertemente en el Capítulo 3 del libro:
James, Witten, Hastie & Tibshirani, "An Introduction to Statistical Learning with Applications in R", 6th ed, Springer, 2015. http://www-bcf.usc.edu/~gareth/ISL/
Esta presentación se ofrece sólo a modo de referencia, y no está completa!

DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Problema de Regresión

- Predecir una **cantidad**:
  - La probabilidad de que un mail sea spam.
  - El tiempo de demora de un vuelo.
  - El precio de una propiedad.



¿Precio de un departamento de 75m^2?

Regresión Lineal Simple

- Consiste en predecir una respuesta cuantitativa *Y* en base a una única variable predictora *X*.

$$Y \approx \beta_0 + \beta_1 \cdot X$$

- Ejemplo:

$$Precio \approx \beta_0 + \beta_1 \cdot Superficie$$

Ordenada al origen (*intercept*)

Pendiente (*slope*)

- β_0 y β_1 son los coeficientes desconocidos que vamos a estimar, o ajustar en base a los datos de entrenamiento. Una vez estimados, los podemos usar para predecir:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

Valor predicho para *Y* cuando *X*=*x*

Estimación de β_0

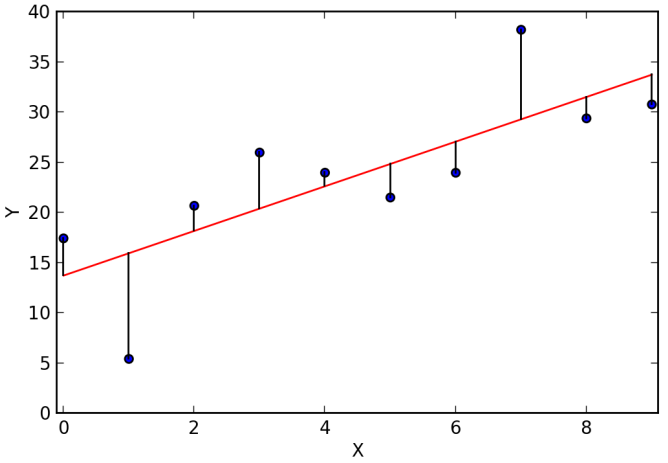
Estimación de β_1

Nueva instancia

Estimación de coeficientes

- Def: Residuo o error de predicción

$$e_i = y_i - \hat{y}_i$$



- Residual sum of squares:

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i\right)^2$$

- Los residuos se elevan al cuadrado para sacar el signo y para que RSS sea diferenciable.
- Hay que tener cuidado con los outliers en los datos, porque RSS penaliza los residuos grandes.

Estimación de coeficientes

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i\right)^2$$

- Para estimar los coeficientes, buscamos minimizar RSS:

$$\frac{\delta \mathrm{RSS}}{\delta \hat{\beta}_0} = 0 \qquad\qquad \frac{\delta \mathrm{RSS}}{\delta \hat{\beta}_1} = 0$$

- Con un poco de análisis matemático, llegamos a estas expresiones (ejercicio 1 de la guía):

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \cdot \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}$$

- Otra opción es minimizar RSS usando la técnica iterativa de Descenso por el Gradiente. (Ejercicio de la guía.)

Predicción de nuevos valores

- Estimados los coeficientes, los podemos usar para predecir:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

- En nuestro ejemplo, podríamos predecir el precio de un departamento de 75m^2:

$$\widehat{Precio} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 75$$

Exactitud de los coeficientes estimados

- Al correr una regresión lineal, es común reportar el error estándar de cada estimador: $\mathrm{SE}(\hat{\beta}_0)$ y $\mathrm{SE}(\hat{\beta}_1)$
- Esto es útil para estimar intervalos de confianza de las predicciones.
- Evaluar la significancia de cada estimador, mediante un test estadístico.
    - *p*-valor bajo (típicamente, $p<0.05$ o $p<0.01$) → es improbable observar al azar una asociación semejante entre *X* e *Y*.
    - *p*-valor alto → es probable que la asociación observada sea sólo consecuencia del azar.

Evaluación del modelo

- RSS: Variabilidad no explicada por el modelo

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

- TSS (Total Sum of Squares): Variabilidad total de los datos

$$\mathrm{TSS} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

- R^2: Proporción de la variabilidad explicada por el modelo.

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}$$

$R^2 \rightarrow 0$ cuando el modelo explica poco de la variabilidad de los datos.

$R^2 \rightarrow 1$ cuando el modelo explica mucho de la variabilidad de los datos.

Regresión Lineal Múltiple

- Consiste en predecir una respuesta cuantitativa *Y* en base a una múltiples variables predictora X_1, X_2, ..., X_p.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- RSS se define igual que para la regresión lineal simple:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

- Los coeficientes se estiman en forma análoga.
- TSS, R^2 también se definen en forma similar.
- ¿Qué pasa si hay variables categóricas? (Ejercicio de la guía.)