

MATEMÁTICAS ESENCIALES PARA INTELIGENCIA ARTIFICIAL, APRENDIZAJE AUTOMÁTICO Y CIENCIA DE DATOS

INTRODUCCIÓN

- Modelo basado en datos → predecir unas variables a partir de otras, entender y modelizar relaciones entre variables, ...
- Tablas de datos se pueden ver como conjuntos de vectores, la unidad de información más sencilla con la que trabajamos es un vector, un dato

INTRODUCCIÓN

- ¿Qué necesidades matemáticas se tienen en este contexto?
 - Entender la representación de relaciones entre vectores: álgebra vectorial, similitud entre vectores, distancias, etc.
 - Comprender los principios básicos del desarrollo de modelos: elección de datos de entrenamiento y validación, interpretación de los modelos, identificación de variables más relevantes, realización de test estadísticos para detectar sobreentrenamiento y determinar diferencias significativas entre distintos modelos,...
 - Entender los procedimientos matemáticos que se usan para ajustar modelos, desde el ajuste de modelos lineales por métodos algebraicos hasta el de modelos no lineales por métodos numéricos

MODELOS LINEALES

- Un modelo lineal es un modelo que intenta relacionar una variable a partir de una combinación lineal de otras variables. Es decir, usando solamente sumas y multiplicaciones por coeficientes constantes

$$Y = a \cdot X + b$$

- En un modelo lineal Podemos aplicar transformaciones no lineales a las variables de entrada, pero después estas deben combinarse linealmente

$$Y = a \cdot X^2 + b$$

$$Y = a \cdot \sqrt{X} + b$$

$$Y = a \cdot e^x + b$$

$$Y = a \cdot x^2 + b \cdot x + c \cdot \sqrt{x} + d$$

MODELOS LINEALES

- Ejemplo sencillo: EjemploAjusteCurvas

- X es un marcador biológico que se mide con un análisis muy sencillo. Y necesita un análisis mucho más costoso. Nos preguntamos si podemos predecir el valor de Y a partir de X
- Antes de empezar a ajustar modelos y ver si funcionan debemos separar el conjunto de datos en datos de ajuste/entrenamiento y test/validación

x	y
0.0115	50.5
0.012	49
0.012	50.2
0.012	44.5
0.013	48.5
0.0135	47.5
0.026	35
0.032	34.5
0.034	38
0.038	31.5
0.04	28
0.041	38.5
0.084	15
0.086	29.5
0.092	20.5
0.098	17

Ajuste	
0.0115	50.5
0.012	50.2
0.013	48.5
0.026	35
0.034	38
0.04	28
0.084	15
0.086	29.5
0.092	20.5

Validación	
0.012	49
0.012	44.5
0.0135	47.5
0.032	34.5
0.038	31.5
0.041	38.5
0.086	29.5
0.098	17

MODELOS LINEALES: AJUSTE

1. Planteamos el modelo a estudiar:

$$Y = \alpha \cdot X + b$$

2. Determinamos los parámetros del modelo utilizando los datos de ajuste. Para esto planteamos qué ecuaciones deben cumplirse:

$$50.5 = \alpha \cdot 0.0115 + b$$

$$50.2 = \alpha \cdot 0.012 + b$$

$$48.5 = \alpha \cdot 0.013 + b$$

$$35 = \alpha \cdot 0.026 + b$$

$$38 = \alpha \cdot 0.034 + b$$

$$28 = \alpha \cdot 0.04 + b$$

$$15 = \alpha \cdot 0.084 + b$$

$$20.5 = \alpha \cdot 0.092 + b$$

Ajuste	
0.0115	50.5
0.012	50.2
0.013	48.5
0.026	35
0.034	38
0.04	28
0.084	15
0.092	20.5

MODELOS LINEALES: AJUSTE

Se trata de un Sistema donde hay más ecuaciones que incógnitas (sobredeterminado) y que será, previsiblemente, incompatible. Es importante atender a la forma matricial del

Incognitas
 a, b

Sistema:

$$Ax = b$$
$$\begin{bmatrix} 0.015 & | & 50.5 \\ 0.012 & | & 50.2 \\ 0.013 & | & 48.5 \\ 0.026 & | & 35 \\ 0.034 & | & 38 \\ 0.04 & | & 28 \\ 0.084 & | & 15 \\ 0.092 & | & 20.5 \end{bmatrix} \quad \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 50.5 \\ 50.2 \\ 48.5 \\ 35 \\ 38 \\ 28 \\ 15 \\ 20.5 \end{bmatrix}$$

$$50.5 = a \cdot 0.0115 + b$$
$$50.2 = a \cdot 0.0112 + b$$
$$48.5 = a \cdot 0.0113 + b$$
$$35 = a \cdot 0.026 + b$$
$$38 = a \cdot 0.034 + b$$
$$28 = a \cdot 0.04 + b$$
$$15 = a \cdot 0.084 + b$$
$$20.5 = a \cdot 0.092 + b$$

Ajuste	
0.0115	50.5
0.012	50.2
0.013	48.5
0.026	35
0.034	38
0.04	28
0.084	15
0.092	20.5

MODELOS LINEALES: AJUSTE POR MÍNIMOS CUADRADOS

- Ver: Ortogonalidad y proyecciones ortogonales.pdf y Método de los mínimos cuadrados.pdf

Si $Ax = b$ es un Sistema incompatible $\rightarrow b$ no puede obtenerse como combinación lineal de las columnas de A

$$Ax - b \neq 0$$

Buscar que $\|Ax - b\|$ sea lo más pequeño posible.

Encontrar x que $\min \|Ax - b\|$.

$$x = (A^T A)^{-1} A^T b$$

Si usamos la norma euclídea, este problema tiene solución algebraica

MODELOS LINEALES

- Ejercicio 1: EjemploAjusteCurvas.xls
 - Utiliza el procedimiento anterior para ajustar el modelo lineal más sencillo con los datos de ajuste.
 - Una vez ajustado, sustituye los valores X del conjunto de validación en el modelo para obtener los valores Y predichos por el mismo
 - Calcula el error cometido por la predicción como la suma de diferencias al cuadrado entre los valores Y deseados del conjunto de validación y los valores Y' predichos por el modelo
- Ejercicio 2:
 - Repite el proceso para los modelos:
 - ¿Cuál de los tres modelos es mejor? ¿Con qué seguridad puedes afirmar que es mejor?

$$Y = \alpha X^2 + b$$

$$Y = \alpha X^2 + bX + c$$

MODELOS LINEALES

- Preguntas que debemos hacernos cuando entrenamos modelos:
 - ¿Está el modelo sobre entrenado?
 - ¿Qué modelo tiene mejor rendimiento?
 - ¿Son las diferencias de rendimiento significativas

En todos los casos el dato principal que utilizaremos es el error cometido por los modelos en los datos de entrenamiento y validación. Aunque hay muchas formas de calcular este error, utilizar la suma de diferencias al cuadrado entre los valores Y deseados y Y' predichos tiene una serie de ventajas. En realidad, se puede utilizar cualquier producto escalar entre los vectores diferencia

$$(Y - Y') (Y - Y')^T \rightarrow \text{Se puede utilizar cualquier producto escalar}$$

MODELOS LINEALES

- Sobreentrenamiento de modelos

$y_e \rightarrow$ datos deseados de entrenamiento

$y_e' \rightarrow$ datos predichos por el modelo para y_e

$$E_e^z = (y_e - y_e') (y_e - y_e')^\top$$

ERROR
ENTRENAMIENTO

$y \rightarrow$ datos de validación deseados

$y' \rightarrow$ datos predichos por el modelo para y

$$E^z = (y - y') (y - y')^\top$$

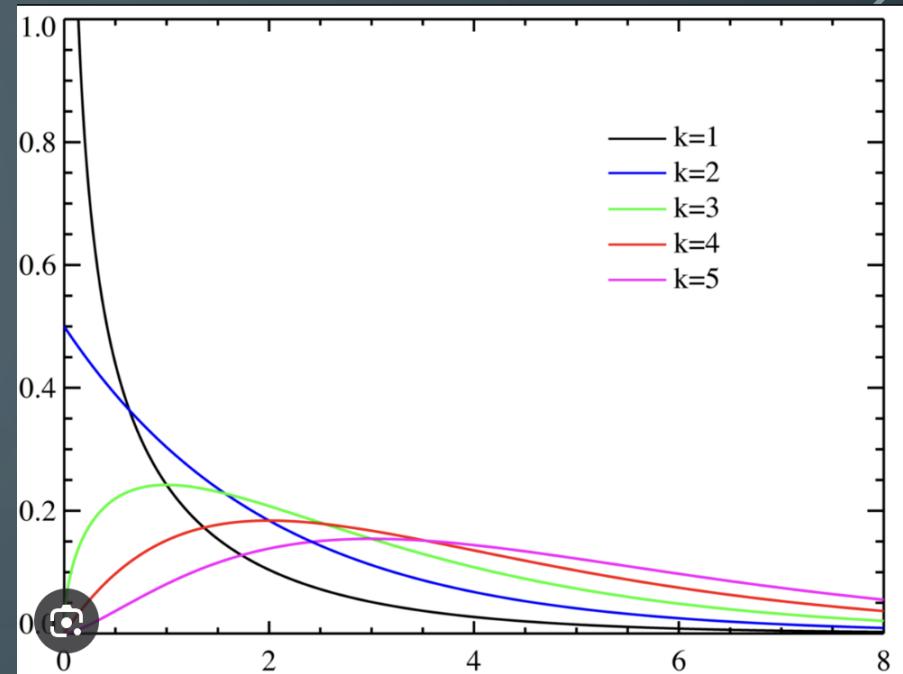
ERROR
VALIDACIÓN

Sobreentrenamiento si

$$E_e^z \ll E^z. \text{ Lo normal es } E_e^z \approx E^z \text{ o } E_e^z \geq E^z$$

MODELOS LINEALES

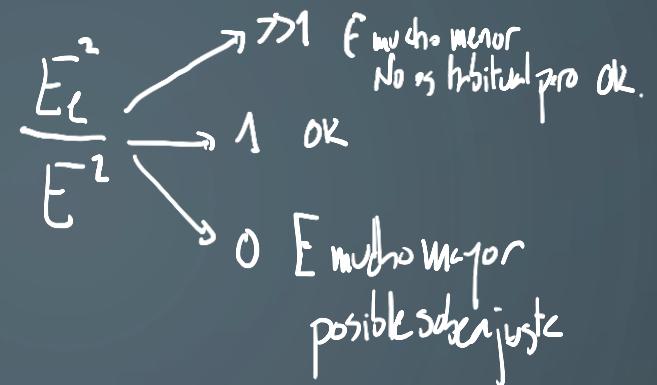
- Sobreentrenamiento de modelos ¿Cuánto es suficiente/demasiada diferencia?
- Aproximación estadística: La suma de errores al cuadrado sigue una distribución Chi-cuadrado de $n-1$ grados de Libertad para n datos
- Compararemos los errores con el cociente
- Lo habitual es comparar el error medio, es decir, el error cuadrático dividido por el número de datos sobre el que está medido. En nuestro ejemplo, Podemos obviarlo porque el número de datos de entrenamiento y validación es el mismo



$\frac{E_L^2}{E^2}$

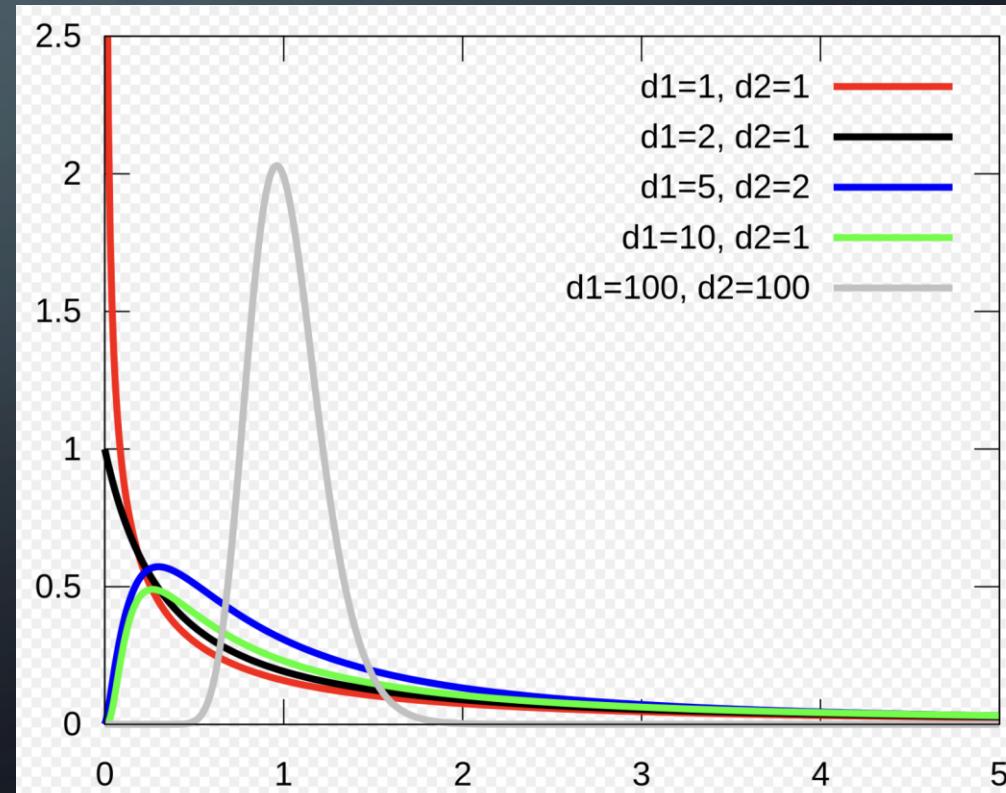
- >1 E_L^2 mucho menor
No es habitual pero OK.
- 1 OK
- 0 E_L^2 mucho mayor
posible sobreajuste

MODELOS LINEALES



- Sobre entrenamiento de modelos: ¿Cuánto es suficiente/demasiada diferencia?
- El cociente de errores al cuadrado sigue una distribución F de $n-1$ y $n-1$ grados de libertad

Nos fijaremos en qué percentil de la distribución están el valor $\frac{E_e^2}{E^2}$. Si está en un percentil muy bajo, es muy probable que tengamos subentrenamiento.



MODELOS LINEALES

```
>> help fcdf
fcdf F cumulative distribution function.
P = fcdf(X,V1,V2) returns the F cumulative distribution function
with V1 and V2 degrees of freedom at the values in X.
```

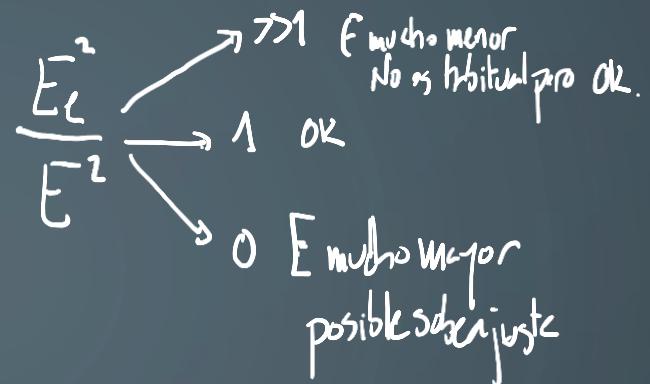
The size of P is the common size of the input arguments. A scalar input
functions as a constant matrix of the same size as the other inputs.

P = **fcdf**(X,V1,V2,'upper') returns the upper tail probability of the
F distribution with V1 and V2 degrees of freedom at the values in X.

See also [finv](#), [fpdf](#), [frnd](#), [fstat](#), [cdf](#).

[Documentation for fcdf](#)

Valores en [0,1] × 100 = Percentil



```
>> fcdf(1,8,8)*100
```

ans =

50.0000

```
>> fcdf(0.1,8,8)*100
```

ans =

0.1907

```
>> fcdf(4,8,8)*100
```

ans =

96.6656

```
>> fcdf(0.25,8,8)*100
```

ans =

3.3344

→ P50, normal OK

→ 2 P1, Sobreajustación
99% probabilidad.

→ P96, Extraino pero OK

→ P3, Sobreajuste 96.6%
probabilidad.

MODELOS LINEALES

- Nivel de confianza: En test estadísticos también es común fijar un nivel de confianza de un determinado % y responder si lo que estamos testeando es o no significativo con ese nivel de confianza → umbralización del percentil anterior.
- El % de confianza fijado representa el % de datos que se consideran normales, mientras que el resto se considera anómalos.

MODELOS LINEALES

- Ejemplo, ¿al 95% de confianza, habría sobre ajuste?
- Equivalente a preguntarse si el estadístico está en el 5% de datos anómalos
- En este caso los datos anómalos son los percentiles más bajos

```
>> fcdf(1,8,8)*100  
ans =  
50.0000  
  
>> fcdf(0.1,8,8)*100  
ans =  
0.1907  
  
>> fcdf(4,8,8)*100  
ans =  
96.6656  
  
>> fcdf(0.25,8,8)*100  
ans =  
3.3344
```

→ \mathcal{N}_0

→ Sí

→ \mathcal{N}_0

→ Sí

MODELOS LINEALES

- Ejemplo, ¿al 98% de confianza, habría sobre ajuste?
- Equivalente a preguntarse si el estadístico está en el 2% de datos anómalos
- En este caso los datos anómalos son los percentiles más bajos

```
>> fcdf(1,8,8)*100  
ans =  
50.0000  
  
>> fcdf(0.1,8,8)*100  
ans =  
0.1907  
  
>> fcdf(4,8,8)*100  
ans =  
96.6656  
  
>> fcdf(0.25,8,8)*100  
ans =  
3.3344
```

→ \mathcal{N}_0

→ s_i

→ \mathcal{N}_0

→ \mathcal{N}_0

MODELOS LINEALES

- Ejercicio: realiza tests para determinar con qué probabilidad existe sobreajuste en los modelos que ajustamos con los datos **EjemploAjusteCurvas.xls**

MODELOS LINEALES

- Comparación de rendimiento entre dos modelos
- Siempre comparamos el error entre datos de validación

$$E_1^2 = (\gamma - \gamma'_1)(\gamma - \gamma'_1)^T$$

$$E_2^2 = (\gamma - \gamma'_2)(\gamma - \gamma'_2)^T$$

- Menor error, mejor modelo. Pero, ¿cuánto deben ser diferentes para considerar la diferencia como significativa?
- De nuevo abordaremos esto desde un punto de vista estadístico: Miraremos el percentil que ocupa en la distribución F el cociente de los errores cuadráticos

MODELOS LINEALES

• Comparación de modelos

```
>> fcdf(1, 8, 8)*100
```

ans =

50.0000

```
>> fcdf(0.1, 8, 8)*100
```

ans =

0.1907

```
>> fcdf(4, 8, 8)*100
```

ans =

96.6656

```
>> fcdf(0.25, 8, 8)*100
```

ans =

3.3344

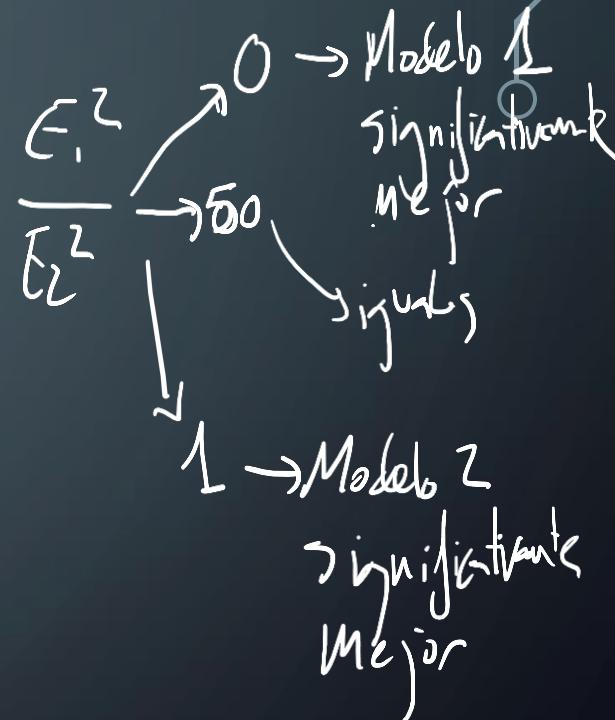
→ iguals

→ Modelo 1 mejor con 98% de prob.

→ Modelo 2 mejor con 96.66% de prob.

→ Modelo 1 mejor con 96.66% de prob.

Percentiles distribución Fd



MODELOS LINEALES

- Fijando nivel de confianza. En este caso se considera anómalo cuando el percentil es muy alto o muy bajo y lo normal está en el centro. Al 90% de confianza quiere decir que los datos anómalos son el primero 5% y el último 5% (test de dos colas)
- Equivalente a preguntarse si el estadístico está por debajo del percentil 5 o más del 95

```
>> fcdf(1,8,8)*100  
ans =  
50.0000  
  
>> fcdf(0.1,8,8)*100  
ans =  
0.1907  
  
>> fcdf(4,8,8)*100  
ans =  
96.6656  
  
>> fcdf(0.25,8,8)*100  
ans =  
3.3344
```

→ Ninguno mejor que el otro al 70%.

→ μ_1 mejor al 70%.

→ μ_2 mejor al 70%.

→ μ_1 mejor al 70%.

MODELOS LINEALES

- Fijando nivel de confianza. En este caso se considera anómalo cuando el percentil es muy alto o muy bajo y lo normal está en el centro. Al 95% de confianza quiere decir que los datos anómalos son el primero 2.5% y el último 2.5% (test de dos colas)
- Equivalente a preguntarse si el estadístico está por debajo del percentil 2.5 o más del 97.5

```
>> fcdf(1,8,8)*100  
ans =  
50.0000  
  
>> fcdf(0.1,8,8)*100  
ans =  
0.1907  
  
>> fcdf(4,8,8)*100  
ans =  
96.6656  
  
>> fcdf(0.25,8,8)*100  
ans =  
3.3344
```

→ Ninguno mejorado 75%.

→ Ni mejorado 95%.

→ Ninguno mejorado 85%.

→ Ninguno mejorado 70%.

MODELOS LINEALES

- Ejercicio: realiza tests para determinar con qué probabilidad los modelos que ajustamos con los datos EjemploAjusteCurvas.xls son mejores o no que los demás utilizando test estadísticos