

Predicción estadística: definición de predicción óptima lineal

Antonio Sala

Dept. Ing. Sistemas y Automatica (DISA)
Universitat Politècnica de València (UPV)

Video-presentación disponible en:

<http://personales.upv.es/asala/YT/V/preli1.html>



UNIVERSITAT
POLITÀCNICA
DE VALÈNCIA

Presentación

Motivación:

En muchas aplicaciones de estadística, el objetivo es “predecir” una variable y dada otra x (o varias), que proporcionan “información” significativa. La predicción puntual óptima $E(y|x)$ es, en general, difícil de calcular o estimar basada en datos.

Objetivos:

Comprender la definición y el significado de los conceptos asociados a la idea de predicción óptima “lineal”.

Contenidos:

Predicción óptima condicional. Predicción óptima lineal. Discusión. Conclusiones.

Predicción óptima condicional

- Predicción de una vble. aleatoria y , **dada otra** x (información):
 - Lo mejor sería estimar/calcular la probabilidad condicional, esto es, función de densidad condicional $f(y|x)$.
 - En su defecto, predicción puntual: media o moda (condicional) + **intervalo de confianza**

A partir de ahora, elegiremos:

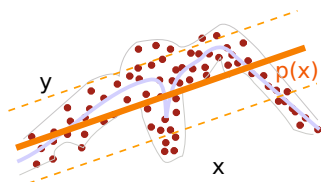
Definición (Predicción óptima puntual)

La mejor predicción de una variable aleatoria y dado un valor x_0 de otra variable aleatoria x es una función $\hat{p}(x_0) = E(y|x_0)$ (media condicional).

Con esa definición, el **error de predicción** $e = y - \hat{p}(x)$ será una variable aleatoria de **media cero**, y **de media independiente** de x (esto es, $E(e|x) = 0$ para todo x). Además tendrá la **mínima varianza** $E[(y - \xi(x))^2]$ de entre todos los posibles predictores $\xi(x)$.

Predicción óptima puntual: difícil!

Se trata de un concepto “ideal”:



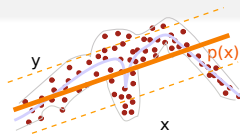
La media $E(y|x)$ puede ser una función muy complicada de x .

Para estimarla bien debería tener muchísimas muestras de y **para cada** x .

* Si los quisiera calcular, la desv. típica o los intervalos de confianza también serían funciones muy complicadas.

En modelos de predicción entre variables continuas, se suelen buscar modelos con funciones más sencillas... p.ej., **regresión lineal**.

Predicción óptima LINEAL



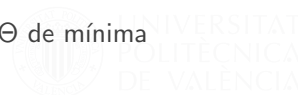
Definición (Predicción óptima **lineal**)

La mejor predicción **lineal** de una variable aleatoria y dado un valor x_0 de otra variable aleatoria x es una función $p(x_0) = \Theta x_0 + \mu$ tal que el error de predicción $e = y - p(x)$ sea una variable aleatoria de **media cero**, **no correlada** con x (esto es, $\Sigma_{ex} = 0$), de **mínima varianza** entre todos los predictores **lineales**.

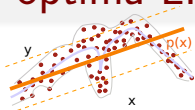
Supongamos (sin pérdida de generalidad) que x se ha trasladado para que tenga **media cero**. Minimizando $J = E[(y - \Theta x - \mu)^2]$, tenemos

$$\frac{dJ}{d\Theta} = -2E[ex] = 0, \quad \frac{dJ}{d\mu} = -2E[e] = -2E[y - \Theta * 0 - \mu] = 2E[y] - \mu = 0$$

esto el error tiene media cero (exige que μ sea la media de y), y el Θ de mínima varianza consigue $E[ex] = 0$, correlación nula.



Discusión: predicción óptima LINEAL vs. IDEAL



El error tiene media **cero** $E[e] = 0$, pero **no** media **cero** **para x fijado arbitrario** $E[e|x] = 0$ que es lo que sería idealmente buscado:

- En algunos x , el error tiene media condicional (variando y , esto es $E(y|x)$) positiva, en otros media negativa... promediando todo (sobre la distrib. de probabilidad de x) sale **cero** $E(E(e|x)) = 0$, pero no es **independiente de x** .

En resumen, en predicción lineal relajamos “error de media cero independiente de x ” a “error media cero que no tenga correlación con x ”.

- ▶ Coincide con la aproximación “determinista” de mínimos cuadrados.
- ▶ Diferencia entre predicción lineal $p(x)$ y la ideal $E(y|x)$: **error sistemático (bias)**.

Conclusiones

- La predicción de una variable puede definirse **condicionada** a conocer el valor de una cierta información.
- La “densidad condicional” sería el concepto “ideal” de dicha predicción. Con suposiciones razonables (mínima varianza), la “**media condicional**” sería el concepto “ideal” de dicha predicción PUNTUAL.
- Ambos conceptos no son, en general, computables, salvo casos sencillos.
- Predicción óptima lineal*: modelo lineal con mínima varianza de error, error media cero, no correlado con la información.

*Coincide con el ajuste “determinista” de mínimos cuadrados, pero presenta **error sistemático** respecto a la media condicional ideal.

