Bases de datos y modelizción !!de datos

J.M. Puddu
FAMAF - Universidad Nacional de Córdoba - Argentina

15 de octubre de 2025

Resumen

funcion de luminosidad de galaxias **Palabras clave**:

1. Introducción

Hastie [2009]

1.1. Ajuste de datos

Para un conjunto de datos observados muchas veces conviene condensarlos en un modelo con parámetros ajustables, el cual nos pueda decir la información que estos contienen. Para definir cual modelo se va a utilizar en general se recurren a funciones que resultan útiles o a funciones sugeridas por la teoría.
A partir de estos se define una **función de mérito** (FM) que mide el acuerdo entre los datos y el modelo. Existen distintas enfoques de plantear esta función de mérito:

- **Enfoque recuentista:** se buscan los parámetros adecuados para que la FM muestre valores pequeños, reflejando el acuerdo
- **Enfoque bayesianos:** la FM será una probabilidad de los parámetros para los datos, tal que valores altos representarán un buen acuerdo.

En este informe, se trabajara con un enfoque bayesiano, donde para conocer la precisión de los parámetros buscamos la
Se denominan d son los datos medidos, m es la función matemática o modelo y θ son los parámetros libres del modelo. El enfoque bayesiano se basa en el **teorema de Bayes**, el cual nos dice que:

$$P(\theta|d,m) = \frac{P(d|\theta,m)P(\phi|m)}{P(d|m)} \tag{1}$$

Donde:

- $P(\theta|d,m)$: "probabilidad a posteriori" de los parámetros dados los datos y el modelo.
- $P(d|\theta,m) = L(d|\theta,m)$: Función de likelihood. Considerando algún modelo de error, nos dice qué tan bien reproducen los datos a las predicciones del modelo.
- $P(\phi|m)$: "Probabilidad a priori o priors" para los parámetros (es lo que sabemos de antemano). Nos habla de los valores permitidos para los parámetros de este modelo - conocimiento previo.
- $P(d|m)$: "Evidencia", nos dice qué tan bien el modelo ajusta los datos. La integral que define la "evidencia" es muy costosa computacionalmente, pero hay ciertas técnicas que se emplean, como el MCMC. la Evidencia sólo tiene sentido cuando se comparan modelos con arquitecturas muy distintas.

Para seleccionar un modelo

1.1.1. Elección del modelo

Para encontrar el modelo que resulta más adecuado para representar los datos hay que: Elegir una familia de modelos en base a la teoría o a partir de la exploración descriptiva de los datos. Un buen criterio para elegir el modelo es seleccionar aquel que logra un balance entre la **varianza** y el **sesgo**

- **Sesgo**: El sesgo o *bias* se produce por que el modelo asume una serie de simplificaciones,si el sesgo es alto entonces el modelo supone demasiadas simplificaciones.
- **Varianza**: la varianza indica como cambia la función de residuos (la diferencia entre el modelo y los datos)

1.1.2. Sesgos en los parámetros - Ejercicio 7

1.1.3. Principio de máxima probabilidad - Ejercicio 9

Partiendo del teorema de Bayes y suponiendo que solo se trabaja con un modelo y que no se combinan

distintos conjuntos de datos se define el principio de máxima probabilidad como:

El principio de máxima probabilidad establece que los parámetros θ deben ser elegidos de manera que la probabilidad del conjunto de datos, especificados los parámetros, sea máxima.

Tomando una muestra de datos, podemos decir que la probabilidad de que estos sigan una función f viene dada por el producto de las probabilidades individuales de cada punto, asumiendo que observamos la función dentro de un diferencial pequeño, definimos la función de probabilidad del parámetro como:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

Es a esta funcion L a la cual buscamos maximizar respecto a θ, obteniendo asi los valores críticos θ_c. Ahora maximizar esta función implica calcular las derivadas primeras de la productoria lo cual es poco eficiente, por lo que aprovechamos la monotonía de la función logaritmo y la propiedad $log(a*b) = log(a)+log(b)$ para simplificar el proceso, maximizando l donde:

$$l(\theta) = log(L(\theta))$$

1.1.4. Cuadrados mínimos

Constituye el caso más sencillo de ajuste de datos bayesiano, suponiendo que tenemos una muestra de datos con dos variables aleatorias en principio idependientes entre sí (x_i, y_i) con sus ccorrespondientes errores $(\sigma_{xi}, \sigma_{yi})$ podemos tomar que en un gráfico x vs y los datos de la muestra siguen una función lineal, es decir que podemos relacionar $y(x) = ax+b$ donde a y b son los parámetros a los que les vamos a aplicar el principio de máxima probabilidad.

Para esto tenemos que desarrollar cual va a ser nuestro L, para lo cuál (y en la versión más básica de los cuadrados mínimos) vamos a suponer que $\sigma_x = 0$, $\sigma_y = \sigma =$ cte y ádemas que los errores de y tienen una distribución gaussiana por lo que:

$$L = \frac{1}{\sqrt{2\pi}\sigma} \prod_{i=1}^{n} e^{-\frac{(y_i - y(x_i,\theta))^2}{2\sigma^2}} \tag{2}$$

Ahora en este caso $\theta = a, b$, es decir que tenemos 2 parámetros para maximizar, por tanto tenemos que calcular

$$\frac{\partial l}{\partial a} = 0 \quad \frac{\partial l}{\partial b} = 0$$

Además que en caso de obtener mas de un valor crítico hay que tomar la segunda derivada para asegurar que dicho valor sea el máximo y no un mínimo o un punto silla.

De esto se obtiene que los parámetros son:

$$a = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

De estas también podemos establecer

Referencias

Hastie T., 2009, The elements of statistical learning: data mining, inference, and prediction