Curso sobre Modelamiento Estadístico y Sistemas Recomendadores

Clase Presencial 2: Modelos Estadísticos y Big DATA

2.1.- MODELOS ESTADÍSTICOS

Modelos estadísticos

- Los modelos estadísticos se construyen en base a modelos de probabilidad.
- Los datos se tratan como realizaciones de variables aleatorias $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, donde $\boldsymbol{X}_i$ es en si mismo un vector de variables aleatorias recogidas para la i-ésima unidad experimental en una muestra de tamaño n desde una población de interés.
- El supuesto es que $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ son generados desde un modelo de probabilidad P.

Modelo estadístico

- Si P se conoce completamente, por ejemplo, al conocer la densidad $f(\boldsymbol{x})$, entonces no existe la necesidad de usar estadística.
- El "problema" estadístico surge cuando existe incertidumbre sobre P.
- Los modelos estadísticos surgen cuando P se trata como un miembro de una familia $\mathcal{M} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexada por un conjunto de parámetros $\boldsymbol{\theta}$, que a su vez pertenece a un conjunto $\boldsymbol{\Theta}$ (espacio paramétrico).

2.2.- Tipos de modelos estadísticos

Modelos paramétricos

- Modelos que se describen a través de un número finito de, típicamente, valores reales se denominan como **modelos paramétricos o finito dimensionales**.

- Los modelos paramétricos pueden ser descritos por la familia

$$\mathcal{M} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$$

  donde la dimensión p es un entero positivo.

Modelos paramétricos

- En el caso de modelos paramétricos, el problema se especifica de tal forma que el valor de los parámetros (o alguna función de ellos) es de importancia para los investigadores.

- El objetivo del análisis estadístico es entonces especificar un valor possible de θ, o determiner un subconjunto de $\mathbf{\Theta}$ para el cuál es posible afirmar que contiene o no a θ, en base a la muestra observada.

Modelos paramétricos

Ejemplo 2:

Suponga que $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$.

Modelos paramétricos

Ejemplo 3 (regresión lineal):

Suponga que $Y_i \mid x_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \ldots, n$.

Modelos no-paramétricos

- En muchas situaciones restringir las inferencias a formas paramétricas específicas pueden limitar los tipos de conclusiones que se pueden obtener.
- Nos gustaría relajar supuestos paramétricos para ganar mayor flexibilidad y robustez en contra de la especificación incorrecta de un modelo paramétrico.
- En estos casos, podemos considerar modelos donde la clase de distribuciones de probabilidad es tan grande que el parámetro θ es de dimensión infinita.

Modelos no-paramétricos

Ejemplo 4 (regresión):

Suponga que $Y_i \mid x_i \overset{ind.}{\sim} N(m(x_i), \sigma^2)$, $i = 1, \ldots, n$, donde $m : \mathbb{R} \longrightarrow \mathbb{R}$ es una función continua.

Modelos no-paramétricos

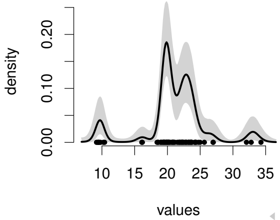
Ejemplo 5 (regresión):

Suponga que $Y_i \mid x_i \overset{ind.}{\sim} f(\cdot - m(x_i))$, $i = 1, \ldots, n$, donde $m : \mathbb{R} \longrightarrow \mathbb{R}$ es una función continua y f es una densidad continua en todas partes y definida en $\mathbb{R}$.

2.3.- BIG DATA

- El tamaño muestral y el número de variables es enorme.
- Los datos no caben en la memoria o no pueden ser almacenados en una sola máquina.

- Los datos deben ser super raros también.



Ejemplos

- Las distribuciones de datos de transacciones en Internet tienen picos en cero y en otros valores discretos (por ejemplo, 1 o \$ 99).
- Colas grandes que importan (por ejemplo, \$ 12 mil/mes de gasto de usuario de eBay).
- El espacio de características potenciales es inmanejablemente grande.
- No podemos escribir modelos simples para explicar los datos.

- Siempre tenemos que almacenar todos los datos para aprender sobre θ?

- Podemos resumir los datos y todavía tener la misma información sobre θ?

- Consideremos un ejemplo simple:

Estadísticos suficientes

Definición:

Un estadístico $T(\boldsymbol{X})$ es **suficiente para** θ si la distribución condicional de la muestra $\boldsymbol{X}$ dado el valor de $T(\boldsymbol{X})$ no depende de θ.

Estadísticos suficientes

Ejemplo:

Sean $X_1, \ldots, X_n \overset{i.i.d.}{\sim}$ Bernoulli(θ), $\theta \in (0, 1)$. $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ es un estadístico suficiente para θ.

Estadísticos suficientes

Teorema de factorización:

Sea $f(\boldsymbol{x} \mid \theta)$ la función de densidad o de probabilidad de la muestra $\boldsymbol{X}$. Un estadístico $T(\boldsymbol{X})$ es suficiente para θ si y sólo si existen unas funciones $g(t \mid \theta)$ y $h(\boldsymbol{x})$, tal que para todo punto muestral $\boldsymbol{x}$ y todo punto en el espacio paramétrico θ,

$$f(\boldsymbol{x} \mid \theta) = g\,(T(\boldsymbol{x}) \mid \theta)\, h(\boldsymbol{x}).$$

Estadísticos suficientes

Ejemplo:

Sean $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, donde $\boldsymbol{\theta} = (\mu, \sigma^2)$. $T(\boldsymbol{X}) = (\bar{X}, S^2)$ es un estadístico suficiente para $\boldsymbol{\theta}$.

Estadísticos mínimo suficientes

Definición:

Un estadístico suficiente $T(\boldsymbol{X})$ para θ se denomina **mínimo suficiente**, si para cualquier otro estadístico suficiente $T'(\boldsymbol{X})$, $T(\boldsymbol{X})$ es una función de $T'(\boldsymbol{X})$.

Comentarios finales

- Una reducción de los datos no es siempre posible y útil. Existe alguna esperanza si $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$, con p finito.
- No existe necesidad de Big DATA en esos casos!!!
- Una reducción es inutil cuando $\boldsymbol{\Theta}$ es un espacio de dimensión infinita (i.e., modelos noparamétricos).

Comentarios finales

- La promesa del Big Data no tiene que ver con conocer algunas características de una población de interés con mayor precisión.
- El interés en los grandes conjuntos de datos se debe a que ellos pueden ser “extraños” y nos permiten conocer los mecanismos complejos que los generan (modelos noparamétricos).
- En estos contextos, puede ser difícil o incluso contraproducente emplear modelos estadísticos simples para el proceso de aprendizaje, donde el número de parámetros se fija a priori (modelos paramétricos).

Comentarios finales

- Los modelos Bayesianos noparamétricos permiten esta felxibilidad (los modelos pueden crecer en tamaño y complejidad con la llegada de más datos), y por un tratamiento coherente de las incertidumbres.