46114: Estadística Multivariada y Datos Categóricos
Sesión 1: Paradigma bayesiano

Juan Carlos Martínez-Ovando

juan.martinez.ovando@itam.mx

Enero 14, 2016

Introducción

La metodología estadística tradicional se ve forzada a evolucionar debido a:

- La complejidad de muchos problemas prácticos actuales.
- La disponibilidad de una vasta cantidad de información.
- El desarrollo de nuevos productos tecnológicos.

Reto particular en regresión

Problemas donde el número de observaciones es significativamente menor al número de regresores (i.e., $p >> n$).

Introducción

Lo anterior ha dado origen al desarrollo de métodos estadísticos en alta dimensión:

- ¿Alta dimensionalidad en variables?
- ¿Alta dimensionalidad en parámetros?

Implicaciones

La complejidad de los modelos, donde el número de observaciones no está 'balanceado' con el número de parámetros, propicia que algunos parámetros sean estimados 'deficientemente' (se dice que algunos parámetros toman prestado potencia de otros).

Introducción

Estimación por máxima verosimilitud

- Falta de identificabiliad y consistencia asintótica.
- ¡Inadmisibilidad! (Paradoja de Stein, en teoría de decisión).
- ¡Complejidad computacional!

Paradoja de Stein (Stein, 1956; James & Stein, 1961)

Si $X \sim N_p(\theta, \lambda)$, con $p > 2$, se tienen dos estimadores para θ:

$$\begin{aligned} \hat{\theta}_{MV} &= \bar{x} \\ \hat{\theta}_{JS} &= \left(1 - \frac{p-2}{\lambda \|x\|^2}\right) \bar{x} \end{aligned}$$

tienen las mismas propiedades...

Introducción

Paradoja de Stein (Stein, 1956; James & Stein, 1961)

La suma de cuadrados de los datos ajustados:

- Invariante para MV en términos de la dispersión implícita a θ.
- Variante para JS en términos de la dispersión implícita a θ (la cota superior es la de MV).

¿Por qué el paradigma bayesiano?

- $\hat{\theta}_{JS}$ surge como un estimador admisible bayesiano de θ.
- El estimador 'reduce' ciertos componentes del estimador de θ con una *prior* centrada en cero (*shrinkage*).
- Es una forma natural de 'lidiar' con el problema $p >> n$.

Paradigma bayesiano: Fundamentos

Probabilidad subjetiva

- Incertidumbre es una 'interpretación personal' originada por nuestra falta de conocimiento.
- Podemos 'manifestar' una opinión (más allá del modelo de probabilidad) antes de observar datos (*a priori* o creencia inicial).
- Actualizamos nustra creencia inicial a la luz de datos (*a posteriori* o creencia posterior).
- Argumentos probabilistas (distribuciones condicionales, teorema de Bayes).

Paradigma bayesiano: Fundamentos

Modelo bayesiano sobre 'observables' y 'no observables':

$$\begin{aligned} p(X|\theta) & \qquad \text{(verosimilitud)}, \\ \pi(\theta) & \qquad \text{(prior)}. \end{aligned}$$

Teorema de Bayes (creencia posterior):

$$\pi(\theta|x) \;=\; \frac{p(x|\theta)\pi(\theta)}{p(x)}.$$

Predicción:

$$p(x^f|x) \;=\; \int p(x^f|\theta)\pi(\theta|x)d\theta.$$

Paradigma bayesiano: Fundamentos

Teorema de representación (una justificación)

Supongamos que $X_1, X_2, \ldots$ es una sucesión de variables aleatorias intercambiables (i.e. invariante ante permutaciones de orden), entonces 'existe' un ente estocástico θ y una medida de probabilidad, π, tal que

$$p(X_1, \ldots, X_n) = \int \prod_{i=1}^{n} p(X_i|\theta)\pi(\theta)d\theta.$$

Naturalmente,

$$p(X_{n+1}|x_1, \ldots, x_n) = \int p(X_{n+1}|\theta)\pi(\theta|x_1, \ldots, x_n)d\theta.$$

Paradigma bayesiano: Fundamentos

Estimación (problema de decisión)

- θ componente de incertidumbre.
- $\hat{\theta}$ componente de decisión.
- $l(\theta, \hat{\theta})$ componente de preferencia (función de pérdida).
- $p(\theta|x_1, \ldots, x_n)$ cuantificación de incertidumbre.

$$\hat{\theta} \;=\; \arg\min_{\theta} \int l(\theta, \hat{\theta}) \pi(\theta|x_1, \ldots, x_n) d\theta.$$

Dos casos:

- Pérdida cuadrática $\Rightarrow$ $\hat{\theta} =$ media de $\pi(\theta|x_1, \ldots, x_n)$.
- Pérdida 0-1 $\Rightarrow$ $\hat{\theta} =$ moda de $\pi(\theta|x_1, \ldots, x_n)$.

Paradigma bayesiano: Ejemplo 1

Distribución normal

$X_1, \ldots, X_n, \ldots$ son v.a. intercambiables $N(\theta, \lambda)$, con λ conocida y θ desconocido.

$$\begin{aligned} p(x|\theta) &= (\lambda/2\pi)^{1/2} \exp\{-\lambda/2(x-\theta)^2\} \\ &\propto \exp\{-\lambda/2(x-\theta)^2\}. \end{aligned}$$

Creencia inicial sobre θ, $N(\theta_0, \lambda_0)$ (con θ_0 y λ_0 dadas),

$$\begin{aligned} \pi(\theta) &= (\lambda_0/2\pi)^{1/2} \exp\{-\lambda_0/2(\theta-\theta_0)^2\} \\ &\propto \exp\{-\lambda_0/2(\theta-\theta_0)^2\}. \end{aligned}$$

Paradigma bayesiano: Ejemplo 1

Creencia posterior

$$
\begin{aligned}
\pi(\theta|x_1,\ldots,x_n) &\propto p(x_1,\ldots,x_n|\theta)\pi(\theta) \\
&\propto \exp\{-\lambda_n/2(\theta-\theta_n)^2\},
\end{aligned}
$$

donde

$$
\begin{aligned}
\lambda_n &= \lambda_0 + n\lambda, \\
\theta_n &= \lambda_n^{-1}(\lambda_0\theta_0 + n\lambda\bar{x}).
\end{aligned}
$$

Predicción

$$
p(X^f|x_1,\ldots,x_n) = N(X^f|\theta_n, \lambda\lambda_n(\lambda+\lambda_n)^{-1}).
$$

Paradigma bayesiano: Ejemplo 1

160114_Ej1.jpg

- Datos: $\{1, 1{,}3, 2, 2{,}3, 1{,}3, 3, 2\}$, con $\lambda = 0{,}3$.
- Hiperparámetros: $\theta_0 = 0$ y $\lambda_0 = 1$.
- Actualización de creencias:
  $\lambda_n = 7(0{,}3) + 1$, y
  $\theta_n = \lambda_n^{-1}(0 + 7(0{,}3)(1{,}843))$.

Paradigma bayesiano: Ejemplo 2

Distribución normal

$X_1, \ldots, X_n, \ldots$ son v.a. intercambiables $N(\theta, \lambda)$, con λ conocida y θ desconocido.

$$\begin{aligned} p(x|\theta) &= (\lambda/2\pi)^{1/2} \exp\{-\lambda/2(x-\theta)^2\} \\ &\propto \exp\{-\lambda/2(x-\theta)^2\}. \end{aligned}$$

Creencia inicial sobre θ, dada por,

$$\pi(\theta) \propto \text{constante.}$$

Paradigma bayesiano: Ejemplo 2

Creencia posterior

$$\begin{aligned} \pi(\theta|x_1,\ldots,x_n) &\propto p(x_1,\ldots,x_n|\theta)\pi(\theta) \\ &\propto \exp\{-\lambda_n/2(\theta-\theta_n)^2\}, \end{aligned}$$

donde

$$\begin{aligned} \lambda_n &= n\lambda, \\ \theta_n &= \bar{x}. \end{aligned}$$

Predicción

$$p(X^f|x_1,\ldots,x_n) = N(X^f|\bar{x},\lambda n(n+1)^{-1}).$$

Paradigma bayesiano: Ejemplo 2

160114_Ej2.jpg

- Datos: $\{1, 1{,}3, 2, 2{,}3, 1{,}3, 3, 2\}$, con $\lambda = 0{,}3$.
- Hiperparámetros: No hay...
- Actualización de creencias: $\lambda_n = 7(0{,}3)$, y $\theta_n = 1{,}843$.

Paradigma bayesiano: Ejemplo 1

Distribución normal

$X_1, \ldots, X_n, \ldots$ son v.a. intercambiables $N(\theta, \lambda)$, con λ conocida y θ desconocido.

$$
\begin{aligned}
p(x|\theta) &= (\lambda/2\pi)^{1/2} \exp\{-\lambda/2(x-\theta)^2\} \\
&\propto \exp\{-\lambda/2(x-\theta)^2\}.
\end{aligned}
$$

Creencia inicial sobre θ, entre dos opciones $N(\theta_1, \lambda_1)$, con probabilidad 0.5, y $N(\theta_2, \lambda_2)$, con probabilidad 0.5, (sup. $\theta_1 \neq \theta_2$).

$$
\pi(\theta) = (0{,}5)N(\theta_1, \lambda_1) + (0{,}5)N(\theta_2, \lambda_2).
$$

Paradigma bayesiano: Ejemplo 3

160114_Ej3.jpg

- Datos: $\{1, 1{,}3, 2, 2{,}3, 1{,}3, 3, 2\}$, con $\lambda = 0{,}3$.
- Hiperparámetros: $\theta_1 = 0$, $\theta_1 = 1$ y $\lambda_1 = \lambda_2 = 1$.
- Actualización de creencias: ???

Herramientas computacionales: Múltiples parámetros

Abstracción

Generalmente, el número de parámetros involucrados con la verosimilitud es de dimensión alta, i.e.

$$X_1, \ldots, X_n \sim p(X|\theta_1, \ldots, \theta_p),$$

con $p > 2$. La distribución inicial $\pi(\theta_1, \ldots, \theta_p)$ puede inducir una creencia actualizada

$$\pi(\theta_1, \ldots, \theta_p|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|\theta_1, \ldots, \theta_p)\pi(\theta_1, \ldots, \theta_p),$$

de forma desconocida.

Aproximación vía simulación

Podemos 'aproximar' $\pi(\theta_1, \ldots, \theta_p|x_1, \ldots, x_n)$ usando muestras datos simulados de $(\theta_1, \ldots, \theta_p)$.

- Como una especie de función de distribución empírica.
- Método de Monte Carlo.

Herramientas computacionales: Múltiples parámetros

Aproximación vía MCMC: Gibbs sampler

Se genera computacionalmente una sucesión $\{(\theta_1^{(j)}, \ldots, \theta_p^{(j)})\}_{j>1}$ de una cadena de Markov, con distribución de transición

$$q(\theta_1^{(j)}, \ldots, \theta_p^{(j)} | \theta_1^{(j-1)}, \ldots, \theta_p^{(j-1)})$$

dada por la siguiente iteración:

$$\begin{aligned}
\theta_1^{(j)}|\ldots &\sim \pi(\theta_1^{(j)}|\theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, \text{datos}), \\
&\vdots \\
\theta_k^{(j)}|\ldots &\sim \pi(\theta_k^{(j)}|\theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)} \ldots, \theta_p^{(j-1)}, \text{datos}), \\
&\vdots \\
\theta_p^{(j)}|\ldots &\sim \pi(\theta_1^{(j)}|\theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, \text{datos}).
\end{aligned}$$

Herramientas computacionales: Múltiples parámetros

Aproximación vía MCMC: Gibbs sampler

Cuando

$$\pi(\theta_k|\theta_1,\ldots,\theta_{k-1},\theta_{k+1},\ldots,\theta_p,\text{datos}) \propto p(x_1,\ldots,x_n|\theta_1,\ldots,\theta_p)\pi(\theta_k),$$

para $k = 1,\ldots,p$, entonces

- $q(\theta_1^{(j)},\ldots,\theta_p^{(j)}|\theta_1^{(j-1)},\ldots,\theta_p^{(j-1)})$ define una cadena de Markov estacionaria y ergódica.
- la distribución invariante de la cadena de Markov es $\pi(\theta_1,\ldots,\theta_p|x_1,\ldots,x_n)$.
- La distribución empírica para $\{(\theta_1^{(j)},\ldots,\theta_p^{(j)})\}_{j=1}^{J}$ es una buena aproximación de $\pi(\theta_1,\ldots,\theta_p|x_1,\ldots,x_n)$ (especialmente para $J >> 0$).

Herramientas computacionales: Ejemplo 4

Distribución normal

$X_1, \ldots, X_n, \ldots$ son v.a. intercambiables $N_p(\boldsymbol{\theta}, \boldsymbol{\lambda})$, con $\boldsymbol{\lambda}$ conocida y $\boldsymbol{\theta}$ desconocido.

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= (2\pi)^{-1/2}|\boldsymbol{\lambda}|^{p/2}\exp\{-1/2(x-\boldsymbol{\theta})\boldsymbol{\lambda}^{-1}(x-\boldsymbol{\theta})\} \\ &\propto \exp\{-1/2(x-\boldsymbol{\theta})\boldsymbol{\lambda}^{-1}(x-\boldsymbol{\theta})\}. \end{aligned}$$

Creencia inicial sobre $\boldsymbol{\theta}$, entre dos opciones $N_p(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)$.

$$\pi(\theta) \propto \exp\{-1/2(\boldsymbol{\theta}_0-\boldsymbol{\theta})\boldsymbol{\lambda}_0^{-1}(\boldsymbol{\theta}_0-\boldsymbol{\theta})\}.$$

Herramientas computacionales: Ejemplo 4

Creencia posterior

$$\begin{aligned}\pi(\boldsymbol{\theta}|x_1,\ldots,x_n) &\propto p(x_1,\ldots,x_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto \exp\{-1/2(\boldsymbol{\theta}-\boldsymbol{\theta}_n)'\boldsymbol{\lambda}_n^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_n)\},\end{aligned}$$

donde

$$\begin{aligned}\lambda_n &= \boldsymbol{\lambda}_0 + n\boldsymbol{\lambda}, \\ \theta_n &= \lambda_n^{-1}(\boldsymbol{\lambda}_0\boldsymbol{\theta}_0 + n\boldsymbol{\lambda}\bar{x}).\end{aligned}$$

Predicción

$$p(X^f|x_1,\ldots,x_n) = N(X^f|\theta_n, \lambda\lambda_n(\lambda+\lambda_n)^{-1}).$$

Herramientas computacionales: Ejemplo 4

160114_Ej4.jpg

- Datos: sitio web del curso, con $\lambda = 0{,}3\boldsymbol{I}_2$.
- Hiperparámetros: $\boldsymbol{\theta}_0 = \mathbf{0}$, $\boldsymbol{\lambda}_0 = \boldsymbol{I}_2$.
- Actualización de creencias: ???