Bayesian Statistics and Machine Learning Workshop 2023

Parametric Priors and Hyperparameters

Martín Onetto

1. Questions

1. Describir la relación estadística entre los datos, parámetros e hiperparámetros.
2. Describir el rol de los hiperparámetros en el modelo estadístico.

1.1. Problems

1. **Prostate cancer**

   a) Cargar la base de datos que describe las variables relacionadas al cancer de prostata. La variable respuesta es *lpsa*.

   b) Describir las columnas como numéricas y categóricas.

   c) Ajustar un modelo lineal (con *intercept*) $y = \beta^T X$. Para eso determinar el máximun likelihood β_{MLE} y su varianza Σ_β. Tomar como σ del ajuste gaussiano al desvío estandard de los datos respuesta.

   d) Calcular el *Z-score* de cada β definido como:

   $$z_j = \frac{\beta_j}{\sqrt{\Sigma_{jj}}}$$

   e) Repetir el ajuste y el cálculo del *Z-score* removiendo de a uno a la vez el parámetro con menor valor asboluto de z. Y ver:

      1) Cómo evolucionan los valores de los parámetros
      2) Cómo cambia su incerteza
      3) Cómo cambia el mean squared error del ajuste.

   f) Proponer modelos *Ridge* y *Lasso* para ver la misma evolución del punto *e)* a medida que aumentamos el hyper parámetro. Interpretar

2. **Hierarchical models**

   Rat tumor problem

   a) Expreimentos previos sobre efectividad de tratamiento contra cancer en ratones de experimentos generaron los siguientes datos:

```
data = [0/20,0/20,0/19,0/18,1/18,1/18,2/20,1/10,3/20,2/13,
         4/20,10/48,6/23,5/19,0/20,0/18,2/25,5/49,9/48,4/19,
         6/22,0/20,0/17,2/24,2/19,10/50,4/19,6/20,0/20,1/20,
         2/23,5/46,4/20,4/19,6/20,0/20,1/20,2/20,3/27,4/20,5/22,
         6/20,0/20,1/20,2/20,2/17,4/20,11/46,16/52,0/19,1/20,2/20,
         7/49,4/20,12/49,15/47,0/19,1/19,2/20,7/47,4/20,5/20,15/46,
         0/19,1/19,2/20,3/20,4/20,5/20,9/24]
```

   donde el numerador es número de ratones con tumores después del tratamiento y el denominador es el número total de ratones en el ensayo.

   b) Usar un modelo binomial donde cada resultado se toma como independiente para estimar la taza de efectividad del tratamiento θ. Calcular la likelihood de que un nuevo experimento resulte en $y_{new} = 4/14$ dado los experimentos anteriores. Interpretar

   c) Proponer un modelo jerárquico de la forma:

$$P(\alpha, \beta) \propto (\alpha + \beta)^{-5/2} \tag{1}$$
$$P(\theta_i|\alpha, \beta) \sim Beta(\alpha, \beta) \tag{2}$$
$$P(n_i|N_i, \theta_i) \sim Binom(\theta_i, N_i) \tag{3}$$

   d) La posterior marginal de los hierparámetros tiene la forma:

$$P(\alpha, \beta|\{n_i, N_i\}) \propto P(\alpha, \beta) \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+n_j)\Gamma(\beta+N_j-n_j)}{\Gamma(\alpha+\beta+n_j)} \tag{4}$$

   Encontrar los α y β que hacen máxima a la posterior. Comparar el valor de $\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}$ con el máximo de la posteior de θ en el inciso *b)*

   e) Simular θ de la distribución $P(\theta|D, \alpha, \beta) \approx P(\theta|D, \hat{\alpha}, \hat{\beta})$ y hacer un histograma con ellos. (Al método de quedarnos con la máxima posterior de los hiperárametros y no con toda su distribución se lo conoce como EM algorithm.)

   f) Para cada parámetro simulado θ^s simular un n^s para $N = 14$ (correspondiente al nuevo experimento). Hacer un histograma de los n^s resultantes y evaluar en que percentil cae el nuevo experimento. Comparar con el resultado anterior.

Eight schools problem

Nos interesa evalaluar como cambia el resultado de los alumnos en los SAT después de un pograma de coaching. El programa se implementó en 8 escuelas y sus resultados se resumen en la siguiente tabla:

School	Estimated effect y_j	Standard Error of effect estimate σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

a) Calcular los intervalos de 95 %, $(y_j \pm 2\sigma_j)$ de cada escuela independientemente y ver que todos se solapan sustancialmente. Concluir que pensar que cada escuela está aislada de la otra en los efectos del coaching no es una buena hipótesis. Por qué?

b) Considerar que los resultados observados son datos independientes de un mismo fenómeno e independientes que vienen de una distribución normal con parámetros (μ, τ). Calcula la posterior de μ y dar su intervalo de credivilidad de 95 %. Usando el máximo de la posterior de μ calcular la likelihood de que un efect haya sido 28. Es verosimil ?

c) Proponer un modelo jerárquico de la forma:

$$P(\mu, \tau^2) \propto 1$$
$$P(\theta_j|\mu, \tau^2) \sim N(\mu, \tau^2)$$
$$P(y_j|\theta_j, \sigma_j) \sim N(\theta_j, \sigma_j)$$

donde vamos a tomar a los σ_j como los valores std effects estimados de los datos y no haremos inferencias sobre ellos.

En este modelo la posterior marginal de $\mu|\tau, y$ es:

$$P(\mu|\tau, y) \sim N(\hat{\mu}, V_\mu) \tag{5}$$

donde:

$$\hat{\mu} = \frac{\sum_{i=1}^{8} \frac{1}{\sigma_j^2+\tau^2} \bar{y_j}}{\frac{1}{\sigma_j^2+\tau^2}} \tag{6}$$

$$V_\mu^{-1} = \sum_{i=1}^{8} \frac{1}{\sigma_j^2 + \tau^2} \tag{7}$$

La posterior de τ resulta una función intrincada de la forma:

$$P(\tau^2|y) \propto V_\mu^{-1/2} \prod_{i=1}^{8} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(\bar{y_j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right) \tag{8}$$

Una manera de ahcer essto es usar el siguiente código:

```
import numpy as np
# Define the unnormalized distribution function
def unnormalized_distribution(tau2, y, sigma2,mu_hat,V_mu):
    product_term = np.sqrt(V_mu)*np.prod(1 / np.sqrt(sigma2 + tau2))
    exponent_term = np.exp(-np.sum((y - mu_hat)**2 / (2 * (sigma2 + tau2))))
    return product_term * exponent_term

# Create an array of tau values
tau_min = 0
tau_max = 30
num_tau_samples = 30000
tau_samples = np.linspace(tau_min, tau_max, num_tau_samples)

# Evaluate the unnormalized distribution at tau values
unnormalized_values = unnormalized_distribution(tau_samples, y, sigma2,mu_hat,V_mu)

# Normalize the distribution
normalized_values = unnormalized_values / np.sum(unnormalized_values)

# Sample from the normalized distribution using numpy's choice function
num_samples = 1000
samples = np.random.choice(tau_samples, size=num_samples, p=normalized_values)
```

3. Simular nuevos efectos para cada escuela j y graficar cómo se distribuyen. Para esto tomar 1000 simulaciones de la posterior de $\tau^2|y$ luego samplear de la posterior $P(\mu|\tau, y)$, esto nos da 1000 muestras de la distribución cojunta $P(\mu, \tau^2|y)$. Para cada una de las muestras simuladas simular un θ_j de cada escuela y con él simular un dato y_j, eso resulta en $(1000, 8)$ simulaciones de datos.

4. Hacer un histograma con las simulaciones obtenidas para cada escuela, y calcular numericamente el intervalo de confianza del 95 % para cada una. Comparar con los intervalos calculados en los incisos anteriores.

5. Ver el como es el comportamiento de $E[\theta_j|\hat{mu}, \tau^2$ como función de $\tau \in [0, 30]$.

6. Ver el como es el comportamiento de $Var[\theta_j|\hat{mu}, \tau^2$ como función de $\tau \in [0, 30]$.