Aprendizaje automatizado

Estimación de parámetros

Gibran Fuentes-Pineda

Febrero 2019

El problema de la estimación de parámetros

- Se asumen ciertas distribuciones en modelo, es decir,

$$\mathcal{X} \sim f(\theta)$$

El problema de la estimación de parámetros

- Se asumen ciertas distribuciones en modelo, es decir,

$$\mathcal{X} \sim f(\theta)$$

- Lanzamiento de una moneda 50 veces (datos)
    - Águila: 15 veces
    - Sol: 35 veces

El problema de la estimación de parámetros

- Se asumen ciertas distribuciones en modelo, es decir,

$$\mathcal{X} \sim f(\theta)$$

- Lanzamiento de una moneda 50 veces (datos)
    - Águila: 15 veces
    - Sol: 35 veces

- Si asumimos una distribución de Bernoulli

$$Ber(k; q) = q^k (1-q)^{1-k},$$

El problema de la estimación de parámetros

- Se asumen ciertas distribuciones en modelo, es decir,

$$\mathcal{X} \sim f(\theta)$$

- Lanzamiento de una moneda 50 veces (datos)
  - Águila: 15 veces
  - Sol: 35 veces

- Si asumimos una distribución de Bernoulli

$$Ber(k; q) = q^k (1 - q)^{1-k},$$

- ¿Qué parámetro q produjo los datos?

Estrategias generales de estimación de parámetros

1. Estimador de máxima verosimilitud (puntual)

$$\hat{\theta}_{EMV} = \arg\max_{\theta} P(\mathcal{X}|\theta)$$

2. Estimador de máximo a posteriori (puntual)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

3. Estimador bayesiano (distribución completa)

$$P(\theta|\mathcal{X}) = \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

Estimador de máxima verosimilitud (EMV)

- Busca los valores de los parámetros que mejor se ajusten a los datos

- Función de verosimilitud

$$\mathcal{L}(\theta|\mathcal{X}) = P(\mathcal{X}|\theta)$$

- Se aproxima al valor real del parámetro cuando $|\mathcal{X}| \to \infty$

EMV para distribución de Bernoulli

- Función de verosimilitud (dadas n muestras)

$$\mathcal{L}(q|\mathcal{X}) = q^{x_1}(1-q)^{1-x_1} \times q^{x_2}(1-q)^{1-x_2} \times \cdots \times q^{x_n}(1-q)^{1-x_n}$$

- Simplificando

$$\mathcal{L}(q|\mathcal{X}) = q^{\sum_{i=1}^{n} x^{(i)}}(1-q)^{n-\sum_{i=1}^{n} x^{(i)}}$$

- Aplicando el logaritmo

$$\log \mathcal{L}(q|\mathcal{X}) = \left(\sum_{i=1}^{n} x^{(i)}\right) \log q + \left(n - \sum_{i=1}^{n} x^{(i)}\right) \log(1-q)$$

- Derivando respecto a q, igualando a cero y despejando

$$\hat{q}_{EMV} = \frac{\sum_{i=1}^{n} x^{(i)}}{n}$$

EMV para distribución de normal

- Función de verosimilitud (dadas n muestras)

$$\mathcal{L}(\mu, \sigma^2|\mathcal{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{-(x^{(i)} - \mu)}{2\sigma^2}$$

- Aplicando el logaritmo

$$\ell(\mu, \sigma^2|\mathcal{X}) = -\frac{1}{2} n \log 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{(x_n - \mu)^2}{2\sigma^2}$$

- Derivando respecto a μ y σ^2 e igualando a cero

$$\hat{\mu}_{EMV} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$$

- Para la varianza

$$\hat{\sigma^2}_{EMV} = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \hat{\mu})^2$$

Estimador de máximo a posteriori (MAP)

- MAP: valor de θ con la probabilidad a posteriori más grande

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathcal{X}) = \arg\max_{\theta} \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

Estimador de máximo a posteriori (MAP)

- MAP: valor de θ con la probabilidad a posteriori más grande

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathcal{X}) = \arg\max_{\theta} \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

- Incorpora información a priori sobre los parámetros
- ¿Qué distribución a priori usamos?

Distribuciones a priori conjugadas

- $P(\theta)$ es una distribucion **a priori conjugada** para $P(\mathcal{X}|\theta)$ si la distribución a posteriori es de la misma familia

Verosimilitud	Parám.	Conjugada	Hiperparám.
Bernoulli	q	Beta	α, β
Binomial	q	Beta	α, β
Multinomial	$\mathbf{q}$	Dirichlet	α
Normal (σ^2 conocida)	μ	Normal	μ_0, σ_0^2
Normal multivar. ($\boldsymbol{\Sigma}$ conocida)	$\boldsymbol{\mu}$	Normal multivar.	$\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$

- `https://en.wikipedia.org/wiki/Conjugate_prior`

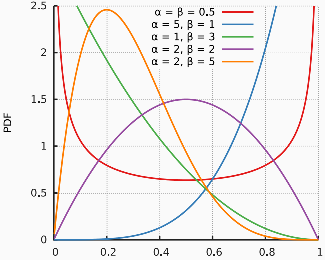
A priori conjugado de Bernoulli: Beta

- Dada la función de verosimilitud de la distribución Bernoulli y n muestras

$$\mathcal{L} = q^{x_1}(1-q)^{1-x_1} \times q^{x_2}(1-q)^{1-x_2} \times \cdots \times q^{x_n}(1-q)^{1-x_n}$$

- Su a priori conjugada es la distribución Beta data por

$$P(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$



MAP para distribución de Bernoulli (1)

- Valor del parámetro que maximice la distribución a posteriori

$$\hat{q}_{MAP} = \arg\max_{q} P(q|\mathcal{X}) = \arg\max_{q} \frac{P(\mathcal{X}|q)P(q)}{P(\mathcal{X})}$$

- Como buscamos el máximo no es necesario calcular la probabilidad marginal, por lo tanto

$$\hat{q}_{MAP} = \arg\max_{q} P(\mathcal{X}|q)P(q)$$

$$\hat{q}_{MAP} = \arg\max_{q} \left(\prod_{i=1}^{|\mathcal{X}|} P(x^{(i)}|q) \right) P(q)$$

$$P(q|\mathcal{X}) \propto \left(\prod_{i=1}^{|\mathcal{X}|} Ber(x^{(i)}|q) \right) Beta(q|\alpha, \beta)$$

MAP para distribución de Bernoulli (2)

- Dada la función de verosimilitud de la distribución Bernoulli y n muestras

- ¿Por qué la distribución Beta?

$$P(q|\mathcal{X}) \propto q^{\sum_{i=1}^{n} x^{(i)}}(1-q)^{n-\sum_{i=1}^{n} x^{(i)}} q^{\alpha-1}(1-q)^{\beta-1}$$

$$P(q|\mathcal{X}) = Beta(q|\alpha + \sum_{i=1}^{n} x^{(i)}, \beta + (n - \sum_{i=1}^{n} x^{(i)}))$$

MAP para distribución de Bernoulli (3)

- Aplicando el logaritmo a $P(q|\mathcal{X})$

$$\hat{q}_{MAP} = \arg\max_q \left(\sum_{i=1}^{n} \log Ber(x^{(i)}|q) \right) + \log Beta(q|\alpha, \beta)$$

- Derivando respecto a q y encontrando el máximo

$$\hat{q}_{MAP} = \frac{\sum_{i=1}^{n} x^{(i)} + \alpha - 1}{n + \beta + \alpha - 2}$$

Estimador bayesiano

- No sólo obtiene el valor de θ del máximo a posteriori, estima la distribución a posteriori completa

$$P(\theta|\mathcal{X}) = \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

- Dado un nuevo dato $\widetilde{x}$, la distribución predictiva a posteriori está dada por

$$P(\widetilde{x}|\mathcal{X},\boldsymbol{\alpha}) = \int_\theta P(\widetilde{x}|\theta,\mathcal{X},\boldsymbol{\alpha})P(\theta|\mathcal{X},\boldsymbol{\alpha})d\theta = E_{\theta|\mathcal{X},\boldsymbol{\alpha}}[P(\widetilde{x}|\theta)]$$

donde $\boldsymbol{\alpha}$ son los hiperparámetros de la distribución a priori

Estimador bayesiano para distribución de Bernoulli

- Usando la Beta como distribución a priori conjugada, tenemos

$$P(q|\mathcal{X}) = Beta\left(q|\alpha + \sum_{i=1}^{n} x^{(i)}, \beta + (n - \sum_{i=1}^{n} x^{(i)})\right)$$

- Dado un nuevo dato $\widetilde{x}$, la distribución predictiva a posteriori está dada por

$$\begin{aligned} P(\widetilde{x}|\mathcal{X}, \alpha, \beta) &= E_{q|\mathcal{X},\alpha,\beta}[P(\widetilde{x}|q)] \\ &= \frac{\alpha + \sum_{i=1}^{n} x^{(i)}}{\alpha + \sum_{i=1}^{n} x^{(i)} + \beta + (n - \sum_{i=1}^{n} x^{(i)})} \end{aligned}$$

Clasificador bayesiano ingenuo

- El clasificador bayesiano ingenuo modela la distribución conjunta de los atributos y las clases, esto es,

$$P(a_1, \dots, a_d, c|\theta_1, \dots, \theta_d, \theta_c) = \left(\prod_{j=1}^{d} P(a_j|c, \theta_j) \right) P(c|\theta_c)$$

  donde θ_c es el parámetros de la distribución a priori sobre las clases y $\theta_j, j = 1, \dots, d$ son los parámetros de las distribuciones de los atributos dada la clase
- Para clasificar usamos teorema de bayes

$$P(c|a_1, \dots, a_d, c, \theta_1, \dots, \theta_d, \theta_c) \propto \left(\prod_{j=1}^{d} P(a_j|c, \theta_j) \right) P(c|\theta_c)$$

Clasificador bayesiano ingenuo para detección de spam

- N documentos representados como bolsas de palabras
  - Vectores $x^{(i)} = [a_1(i), \ldots a_d^{(i)}], i = 1, \ldots, N$
  - $a_j^{(i)} = 1$ si la palabra j ocurre en el documento i, 0 si no ocurre
  - d es el tamaño del vocabulario
- Presuposición

$$a_j \sim Ber(k; q_j), j = 1, \ldots, d$$

- Para clasificar un nuevo documento $x = [a_1, \ldots a_d]$ se aplica el teorema de bayes para cada clase

$$P(c|a_1, \ldots, a_d, q_1, \ldots, q_d, q_c) \propto \left(\prod_{j=1}^{d} P(a_j|c, q_j) \right) P(c|q_c)$$