Aprendizaje automatizado

Estimación de parámetros

Gibran Fuentes-Pineda

Febrero 2023

Interpretaciones de la probabilidad

- ¿Qué significan las probabilidades?
- ¿Cómo las obtengo?
- Ejemplo: lanzamiento de una moneda
  - ¿Qué valores asigno a águila y a sol?
  - ¿Qué representan esos valores?

Interpretación clásica

- Basado en principio de indiferencia: todas las posibilidades tienen la misma probabilidad

Interpretación clásica

- Basado en principio de indiferencia: todas las posibilidades tienen la misma probabilidad
- Ejemplo
  - Lanzamiento de una moneda

$$P(S) = \frac{1}{2}, P(A) = \frac{1}{2}$$

Interpretación frecuentista

- Probabilidades representan aspectos reales del universo (**perspectiva objetivista**)
- Límite de las frecuencias en un gran número de experimentos
- Ejemplo
  - Lanzamiento de una moneda: $A, A, S, A, S, A, S, S, A, A$

$$P(A) = \frac{6}{10}$$

$$P(S) = \frac{4}{10}$$

Interpretación bayesiana

- Probabilidades son grados de creencia de un observador (**perspectiva subjetivista**)
- Probabilidades se actualizan con nueva evidencia
- Ejemplo
  - Lanzamiento de una moneda. E = A, A, S, A, S, A, S, S, A, A

$$P(A|E) = \frac{P(E|A)P(A)}{P(E)}$$

$$P(S|E) = \frac{P(E|S)P(S)}{P(E)}$$

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 12 veces (datos i.i.d) y obtenemos

$$\mathcal{X} = \{A, S, A, A, A, A, S, S, A, A, A, S\}$$

El problema de la estimación de parámetros

- Se asumen ciertas distribuciones en modelo, es decir,

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

- Lanzamiento de una moneda 12 veces (datos i.i.d) y obtenemos

$$\mathcal{X} = \{A, S, A, A, A, A, S, S, A, A, A, S\}$$

- Si asumimos una distribución de Bernoulli

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

El problema de la estimación de parámetros

- Se asumen ciertas distribuciones en modelo, es decir,

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

- Lanzamiento de una moneda 12 veces (datos i.i.d) y obtenemos

$$\mathcal{X} = \{A, S, A, A, A, A, S, S, A, A, A, S\}$$

- Si asumimos una distribución de Bernoulli

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

- ¿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 i.i.d)

$$\mathcal{L}(q|\mathcal{X}) = q^{x^{(1)}}(1-q)^{1-x^{(1)}} \times q^{x^{(2)}}(1-q)^{1-x^{(2)}} \times \cdots \times q^{x^{(n)}}(1-q)^{1-x^{(n)}}$$

- 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 i.i.d)

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

- Aplicando el logaritmo

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

- 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

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

EMV para otras distribuciones

Nombre	Definición	EMV
Poisson	$f(x;\lambda) = \dfrac{\lambda^x e^{-\lambda}}{x}$	$\hat{\lambda} = \dfrac{1}{n}\sum_{i=1}^{n} x^{(i)}$
Categórica	$f(x;\mathbf{q}) = \prod_{k=1}^{K} q_k^{[x=k]}$	$\hat{q}_k = \dfrac{1}{n} c_k$
Multinomial	$f(\mathbf{c}; n, \mathbf{q}) = \dfrac{n!}{\prod_{k=1}^{K} c_k!} \prod_{k=1}^{K} q_k^{c_k}$	$\hat{q}_k = \dfrac{1}{n} c_k$

$$\mathbf{c} = [c_1, \ldots, c_K] = \left[\sum_{i=1}^{n}[x=1], \ldots, \sum_{i=1}^{n}[x=K]\right]$$

$[x = k]$ son los corchetes de Iverson

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[1]

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

[1] Puedes encontrar una lista de distribuciones a priori conjugadas en `https://en.wikipedia.org/wiki/Conjugate_prior`.

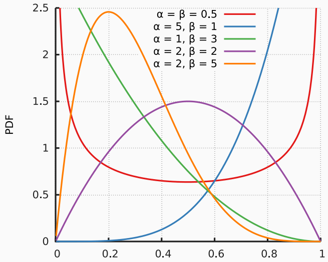
A priori conjugado de Bernoulli: Beta

- 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

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



$$Moda = \frac{\alpha-1}{\alpha+\beta-2}$$

$$Media = \frac{\alpha}{\alpha+\beta}$$

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

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

MAP para otras distribuciones

Nombre	MAP
Poisson	$\hat{\lambda} = \frac{\sum_{i=1}^{n} x^{(i)} + \alpha - 1}{n + \beta}$
Categórica	$\hat{q}_k = \frac{c_k + \alpha_k - 1}{n + \sum_{k=1}^{K} \alpha_k - K}$
Multinomial	$\hat{q}_k = \frac{c_k + \alpha_k - 1}{n + \sum_{k=1}^{K} \alpha_k - K}$
Normal (σ^2 conocido)	$\hat{\mu} = \frac{\sigma_0^2 \left(\sum_{i=1}^{n} x^{(i)}\right) + \sigma^2 \mu_0}{\sigma_0^2 n + \sigma^2}$

$$\mathbf{c} = [c_1, \ldots, c_K] = \left[\sum_{i=1}^{n} [x = 1], \ldots, \sum_{i=1}^{n} [x = K]\right]$$

$[x = k]$ son los corchetes de Iverson

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

$$P(\widetilde{x}|\mathcal{X}) = \int_\theta P(\widetilde{x}|\theta, \mathcal{X})P(\theta|\mathcal{X})d\theta$$

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

$$\begin{aligned} P(\widetilde{x}|\mathcal{X}) &= \int_{\theta} P(\widetilde{x}|\theta, \mathcal{X}) P(\theta|\mathcal{X}) d\theta \\ &= \int_{q} q \cdot Beta(q|\alpha + \sum_{i=1}^{n} x^{(i)}, \beta + (n - \sum_{i=1}^{n} x^{(i)}))dq \\ &= \mathbb{E}[P(q|\mathcal{X})] = \frac{\alpha + \sum_{i=1}^{n} x^{(i)}}{\alpha + \beta + n} \end{aligned}$$

Estimador bayesiano para distribución normal

- Suponiendo σ^2 conocida, la distribución a priori conjugada sobre μ es una normal:

$$P(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$$

- La distribución a posteriori es también normal:

$$P(\mu|\mathcal{X}) = \mathcal{N}\left(\underbrace{\frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left[\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x^{(i)}}{\sigma^2}\right]}_{\mu_p}, \underbrace{\left[\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right]^{-1}}_{\sigma_p^2}\right)$$

- La distribución predictiva a posteriori está dada por:

$$P(\widetilde{x}|\mathcal{X}) = \mathcal{N}(\mu_p, \sigma_p^2 + \sigma^2)$$

Estimador bayesiano para otras distribuciones

Nombre	A posteriori	Predictiva
Poisson	$\Gamma(\underbrace{\alpha + \sum_{i=1}^{n} x^{(i)}}_{\alpha'}, \underbrace{\beta + n}_{\beta'})$	$P(\widetilde{x}) = NB(\alpha', \beta')$
Cat	$Dir(\boldsymbol{\alpha} + \mathbf{c})$	$P(\widetilde{x} = k) = \dfrac{\alpha_k + c_k}{n + \sum_{k=1}^{K} \alpha_k}$
Mult.	$Dir(\boldsymbol{\alpha} + \mathbf{c})$	$P(\widetilde{x} = k) = DirMult(\widetilde{x} \mid \boldsymbol{\alpha} + \mathbf{c})$

$$\mathbf{c} = [c_1, \ldots, c_K] = \left[\sum_{i=1}^{n}[x = 1], \ldots, \sum_{i=1}^{n}[x = K]\right]$$

$[x = k]$ son los corchetes de Iverson