Màxima versemblança

Aleix Torres i Camps

1 Pròleg

En aquest document s'explica mitjançant un entrada escalonada què és i per a què serveix la funció de màxima versemblança (*maximum likelihood*) i la seva acompanyant més freqüent, la funció de màxima logversemblança (*maximum loglikelihood*). L'objectiu del mateix és descriure el context en que es situa a cada moment, per a que així fer més intuitius els passos que segueixen a la deducció de la funció de màxima versemblança. També s'inclouen, en tot moment, **exemples** amb la finalitat de fer el document més pràctic i allunyar-nos puntualment de l'abstractis-me i la generalització.

2 Què és la funció de màxima versemblança?

La funció de densitat (i de probabilitat):

Sigui Y una variable aleatòria que pot dependre del paràmetre θ, aquest poguent ser un vector $\theta = (\theta_1, \cdots, \theta_k)$, direm que la seva funció de densitat (o de probabilitat si és discreta) és:

$$\boxed{f_Y(y;\theta) \qquad \text{on } y \text{ és la variable de la funció}}$$

En el cas d'una mostra aleatòria de grandària N, suposant rèpliques independents $\{Y_1, Y_2, \cdots, Y_N\}$ la funció de densitat de la mostra és

$$\boxed{f_{\{Y_1,Y_2,\cdots,Y_N\}}(y_1, y_2, \cdots, y_N;\theta) = \prod_{i=1}^{N} f_{Y_i}(y_i;\theta)}$$

és a dir, el producte de les funcions de densitat de les Y_i.

Recordatori

La funció de densitat dona la probabilitat que la variable aleatória caigui en un conjunt de valors. Però també es pot servir per comparar probabilitats de dos punts concrets, ja que un és més probable que l'altre si el valor de la funció de densitat en aquell punt és més gran que el valor per l'altre punt.

Per les variables aleatories discretes, les funcions de probabilitat ja donen la probabilitat de valors concrets i, per tant, també es pot fer servir per comparar les probabilitats de dos punts concrets.

Exemple de funció de densitat (*distribució de Poisson*)

Suposant un succés aleatori que cada variable de la mostra és independent identicament distribuida (i.i.d.) i segueixen un *distribució de Poisson*, és a dir, $Y_i \sim Poisson(\lambda)$, cada una té per funció de probabilitat:

$$\boxed{f_{Y_i}(y_i;\lambda) = e^{-\lambda}\frac{\lambda^{y_i}}{y_i!}}$$

I, per tant, la funció de probabilitat d'una mostra de grandària N és

$$\boxed{f_Y(y;\lambda) = \prod_{i=1}^{N} e^{-\lambda}\frac{\lambda^{y_i}}{y_i} = e^{-N\lambda}\frac{\lambda^{\sum_{i=1}^{N} y_i}}{\prod_{i=1}^{N} y_i!}}$$

La funció de densitat (i de probabilitat) d'un model:

El següent pas és considerar què passaria si les variables $Y = \{Y_1, Y_2, \cdots, Y_N\}$ depenguessin dels valors d'unes variables explicatives X, quan $\{X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N\}$ (on cada X_i pot pendre un vector de valors). Aleshores la funció de densitat de la variable $Y_i|X_i$ és

$$\boxed{f_{Y_i|X_i}(y_i; x_i, \theta)}$$

I la del model és:

$$\boxed{f_{Y|X}(y; x, \theta) = \prod_{i=1}^{N} f_{Y_i|X_i}(y_i; x_i, \theta)}$$

que depèn de la variable $y = (y_1, y_2, \cdots, y_N)$.

Exemple de funció de densitat d'un model (*recta de regressió*)

Siguin $Y_i|X_i$ unes variables aleatòries que segueixen una $N(\beta_0+\beta_1 x_i, \sigma^2)$. Donats els valors $\{X_1 = x_1, \cdots, X_N = x_N\}$, la funció de densitat de cada Y_i és

$$\boxed{f_{Y_i|X_i}(y_i; x_i, \beta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right)^2}}$$

per tant, la funció de densitat del model és

$$\boxed{f_{Y|X}(y; x, \beta, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right)^2}}$$

Continuació de l'exemple, ara amb valors numérics (*recta de regressió*)

Suposem que, per estudiants de sexe masculí i amb edat entre els 18 i els 26 anys, la funció de densitat del seu pes (en Kg) coneixent la seva alçada (en cm) segueix una $N(-33 + 0.62 \cdot Alçada, 8^2)$. Per un estudiant que medeixi 170 cm tindrem:

$$\boxed{f_{y_1}(y_1) = \frac{1}{\sqrt{2\pi 8^2}} e^{-\frac{1}{2}\left(\frac{y_1-(-33+0.62\cdot 170)}{8}\right)^2} = 0.0499 e^{-\frac{1}{128}(y_1-72.4)^2}}$$

I, per un conjunt d'estudiants amb alçades $X = \{170, 188, 178, 192, 184\}$ tindrem:

$$\boxed{f_Y(y) = (0.0499)^5 e^{-\frac{1}{128}\sum_{i=1}^{5}(y_i-(-33+0.62 x_i))^2}}$$

Canvi de paper entre les Y_i i dels paràmetres θ.

Fins ara, suposavem que de les variables aleatòries Y_i en coneixíem els paràmetres i podiem utilitzar-la per calcular probabilitats. Ara, podem pensar-ho al revés, seguim suposant que les variables Y_i segueix una distribució que depen d'uns paràmetres, però aquests paràmetres els desconeixem. I el que tenim és una sèrie de valors experimentals, és a dir, per cada Y_i coneixem y_i. Aleshores, per una mostra tindrem:

$$\boxed{\mathcal{L}(\theta; y) = \prod_{i=1}^{N} f_{Y_i}(y_i; \theta) \qquad \text{On } \theta \text{ és la variable de la funció.}}$$

A la funció $\mathcal{L}(\theta; y)$ l'anomenarem. I, al seu logaritme l'anomenarem funció de logversemblança:

$$\boxed{\ell(\theta; y) = \log \mathcal{L}(\theta; y) = \sum_{i=1}^{N} \log f_{Y_i}(y_i; \theta)}$$

Exemple de funció de funció de versemblança I (*distribució de Poisson*)

Una variable que segueix una distribució de Poisson té per funció de densitat:

$$\boxed{f_{Y_i}(y_i;\lambda) = e^{-\lambda}\frac{\lambda^{y_i}}{y_i!}}$$

Les seves funcions de versemblança i de logversemblança són, respectivament:

$$\boxed{\mathcal{L}(\lambda;y) = e^{-N\lambda}\frac{\lambda^{\sum_{i=1}^N y_i}}{\prod_{i=1}^N y_i!}}$$

$$\boxed{\mathscr{L}(\lambda;y) = -N\lambda + \log(\lambda)\sum_{i=1}^N y_i - \sum_{i=1}^N \log(y_i!)}$$

Continuació de l'exemple, ara amb valors numérics (*distribució de Poisson*)

L'exemple clàssic d'una distribució de Poisson és el de comptar el nombre de cotxes que passen per una carretera poc transitada. Desconeixem el valor de λ però obtenim 5 recomptes experimentals $y = \{11, 6, 8, 5, 8\}$. Aleshores, per aquest cas, les funcions de versemblança i logveresemblança són, respectivament:

$$\boxed{\mathcal{L}(\lambda) = e^{-5\lambda}1.78\cdot 10^{-22}\lambda^{38}}$$

$$\boxed{\mathscr{L}(\lambda) = -5\lambda + 38\log(\lambda) - 50.078}$$

Canvi de paper entre les Y_i i dels paràmetres θ en un model.

Ara fem el mateix canvi, però per les variables aleatòries $Y_i|X_i$. A més, ara coneixem el valor experimental de y_i i x_i, però també desconeixem els parametres. Tambe anomenarem funció de versemblança a:

$$\boxed{\mathcal{L}(\theta;y,x) = \prod_{i=1}^N f_{Y_i|X_i}(y_i;x_i,\theta) \qquad \text{On } \theta \text{ és la variable de la funció.}}$$

I funció de logversemblança a:

$$\boxed{\mathscr{L}(\theta;y,x) = \log\mathcal{L}(\theta;y,x) = \sum_{i=1}^N \log f_{Y_i|X_i}(y_i;x_i,\theta)}$$

Exemple de funció de versemblança II (*recta de regressió*)

Una mostra, les variables aleatòries de la qual segueixen un distribució $N(\beta_0+\beta_1 x_i,\sigma^2)$, té per funció de versemblança i per funció de logversemblança:

$$\boxed{\mathcal{L}(\beta,\sigma;y,x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2}\sum_{i=1}^N\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right)^2}}$$

$$\boxed{\mathscr{L}(\beta,\sigma;y,x) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - (\beta_0+\beta_1 x_i))^2}$$

Continuació de l'exemple, ara amb valors numérics (*recta de regressió*)

Si possem per exemple una recta de regressió que prediu el pes (y_i) d'uns estudiants en funció de la seva alçada (x_i), és a dir, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Agafem una mostra de 5 estudiants d'alçades $x = \{170, 188, 178, 192, 184\}$ i veiem que pesen $y = \{74, 90, 75, 86, 60\}$ però desconeixem els paràmetres β i σ, fent els càlculs corresponents obtenim:

$$\boxed{\mathcal{L}(\beta, \sigma) = 0.0101\sigma^{-5} e^{-\frac{32997-2(\beta_0 405+\beta_1 74082)+5\beta_0^2+2\beta_0\beta_1 912+\beta_1^2 166648}{2\sigma^2}}}$$

$$\boxed{\mathscr{L}(\beta, \sigma) = -4.595 - 5\log(\sigma^2) - \frac{1}{2\sigma^2}(32997 - 2(\beta_0 405 + \beta_1 74082) + 5\beta_0^2 + 2\beta_0\beta_1 912 + \beta_1^2 166648)}$$

Significat de la funció de versemblança

Podem interpretar la sortida de la funció $\mathcal{L}(\theta; y)$ com quant de creïble és que amb la θ d'entrada, una mostra obtingui valors y. Per variables continues no és ben bé una probabilitat, però dues θ diferents es poden comparar amb la funció de versemblança. És a dir, si $\mathcal{L}(\theta_1; y) > \mathcal{L}(\theta_2; y)$, aleshores és més probable que obtinguem els valor de y amb θ_1 que amb θ_2. Direm que θ_1 és més versemblant que θ_2.

Mètode de màxima versemblança

El mètode de màxima veremblança consisteix en estimar el paràmetre θ amb els valor més versemblant, és a dir: $\hat{\theta} = \max_\theta \mathcal{L}(\theta; y)$. En general, aquest màxim existeix i és únic, però hi ha situacions que podria no existir o no ser únic.

Càlcul de l'estimador màxim veremblant

Per trobar $max\theta = \mathcal{L}(\theta; y)$, com normalment es tractaran de funcions diferenciables, s'haurà de resoldre les equacions:

$$\boxed{\frac{\partial}{\partial \theta_i}\mathcal{L}(\theta; y) = 0}$$

Però, utilitzant que el logaritme és una funció creixent, sabem que obtindrem el mateix valor si es resolen les equacions (sovint més simples):

$$\boxed{\frac{\partial}{\partial \theta_i}\mathscr{L}(\theta; y) = 0}$$

Ara, és convenient comprovar que realment es tracte d'un màxim mirant que la hessiana és definida negativa.

Propietats importants dels estimadors màxim versemblants

- Invariància: Si podem parametritzar la variable aleatòria amb el canvi $\theta = h(\alpha)$, podem obtenir $\hat{\theta}$ o $\hat{\alpha}$ i compleixen que $\hat{\theta} = h(\hat{\alpha})$.
- $\hat{\theta}$ és asimptòticament no esbiaxat, és a dir, $\mathbb{E}[\hat{\theta}] \xrightarrow[n\to\infty]{} \theta$. Tot i que, en molts casos és no esbiaxat, és a dir, $\mathbb{E}[\hat{\theta}] = \theta$.
- La distribució asimptòtica de $\hat{\theta}$ és $N(\theta, \sigma^2_{\hat{\theta}})$ amb $-\left(\frac{\partial^2}{\partial\theta}\mathscr{L}(\hat{\theta}; y)\right)^{-1} \xrightarrow[n\to\infty]{} \sigma^2_{\hat{\theta}}$.

Exemple de càlcul de l'estimador màxim versemblant (*distribució de Poisson*)

En un punt d'una carretera amb poc tràfic comptem el nombre de vehicles que passen durant 10 minuts. Suposem que aquest recompte segueix una distribució de Poisson(λ). Es fan 5 recomptes i s'ha obtingut una mostra $y = \{11, 6, 8, 5, 8\}$. Ens interessa l'estimador màxim versemblant $\hat{\lambda}$ i la seva distribució asimptòtica. Primer calculem la funció de logversemblança.

$$\boxed{\mathscr{L}(\lambda) = -5\lambda + 38\log(\lambda) - 50.078}$$

Llavors volem solucionar l'equació:

$$\frac{\partial}{\partial\lambda}\mathscr{L}(\lambda) = -5\lambda + \frac{38}{\lambda} = 0$$

Que ens dona $\hat{\lambda} = \frac{38}{5} = 7.6$. A més:

$$\frac{\partial^2}{\partial\lambda^2}\mathscr{L}(\lambda) = -\frac{38}{\lambda^2} < 0$$

Amb el que podem assegurar que $\hat{\lambda}$ és un màxim. Per calcular la seva distribució asimptòtica necessitem el valor:

$$-\left(\frac{\partial^2}{\partial\theta}\mathscr{L}(\hat{\lambda})\right)^{-1} = \frac{\hat{\lambda}^2}{38} = \frac{7.6^2}{38} = 1.52$$

Llavors la distribució asimptòtica de $\hat{\lambda}$ és $N(\lambda, 1.233^2)$.