Calcul Numeric

Laboratorul 6.

Metode Bayesiene

1 Inferența

Inferența Bayesiană este o aplicație a teoriei probabilităților ce permite modelarea riguroasă a incertitudinilor în legătură cu un anumit fenomen. Presupunem că dorim să hotărâm acțiunea cea mai potrivită într-o anumită situație. Decizia de a selecta o acțiune în detrimentul alteia poate fi luată în funcție de utilitatea fiecăreia, adică de diferența dintre beneficiile acțiunii și costurile sale. Însă de regulă și beneficiile și costurile presupun o serie de incertitudinii ce complică luarea deciziei. Scopul inferenței este de a oferi informații despre influența acestor incertitudini.

Concret, inferența presupune selecția, din mai multe modele posibile ce pot descrie fenomenul în cauză, a aceluia care se potrivește cel mai bine cu datele observate. Există două perspective asupra problemei modelării. Prima presupune că există un model adevărat care generează observațiile pe care le facem, iar scopul inferenței este acela de a-l determina. Cea de-a doua nu admite existența acestui model real. În această viziune, există doar moduri de a aproxima realitatea, scopul inferenței fiind de această dată de a găsi modelul cel mai potrivit.

Inferența Bayesiană include două tipuri de informații: una provine de la măsurători, iar alta de la cunoștințele prealabile în legătură cu domeniul studiat. În practică nu avem acces decât la un set de date de dimensiuni restrânse, obținute din măsurători limitate în timp/spațiu, însă ne dorim să caracterizăm în mod general procesul care stă în spatele acestor valori măsurate.

Modelele statistice sunt definite de anumiți parametrii. În cazul unei distribuții normale, spre exemplu, aceștia sunt media (μ) și deviația standard (σ). Dacă avem motive să considerăm că datele experimentale de care dispunem provin dintr-o distribuție normală, inferența ne va ajuta să găsim valorile mediei și deviației standard ale acestei distribuții, astfel că vom putea să caracterizăm nu doar datele disponibile, ci întregul fenomen. Nu va fi, așadar, vorba de a găsi media și deviația standard a datelor experimentale, ci a întregii distribuții din care ele fac parte.

Formula lui Bayes spune că

$$P(\Theta|x) = \frac{P(x|\Theta)P(\Theta)}{P(x)} \tag{1}$$

Modelul Θ, iar $P(\Theta|x)$ indică probabilitatea de a găsi parametrii acestuia (media și dispersia, în cazul unei distribuții normale) după ce am observat datele. Pentru alegerea modelului (din mai multe disponibile), ne vom folosi de această probabilitate: cu cât ea este mai mare, cu atât modelul explică mai bine realitatea. Ca atare, problema devine găsirea setului de parametrii μ și σ pentru care această probabilitate, pe care o numim *a posteriori*, este maximă.

Primul termen din numărătorul expresiei de mai sus, probabilitatea $P(x|\Theta)$, se numește *verosilimitate* și reprezintă probabilitatea de a obține datele (varibila x), având un model Θ (cu parametrii asociați). Cel de-al doilea termen, $P(\Theta)$, numit probabilitatea *a priori* descrie cunoștințele despre domeniu, care sunt independente de datele observate. Iar termenul de la numitor reprezintă probabilitatea datelor și este independentă de model.

Pentru o distribuție normală, verosimilitatea are o formă cunoscută

$$P(x|\Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

În ce privește probabilitatea a priori, aceasta depinde, în mod evident, de contextul fiecărei aplicații, însă în cazul unui fenomen ce respectă o distribuție normală, aceasta poate conține, spre exemplu, informații despre valorile posibile pentru medie și dispersie. Dacă vrem să modelăm înălțimea unei populații, în mod cert valorile vor fi pozitive, iar media nu va putea avea o valoare mai mare decât cea mai mare înălțime măsurată vreodată. Aceste informații sunt de asemenea exprimate în termeni de probabilități, dar nu este obligatoriu ca aceasta să fie de același tip cu cea exprimată de model. Mai mult, inferența Bayesiană permite utilizarea unor probabilități neinformative, deci este utilă și în cazurile în care nu există foarte multe informații despre fenomen.

Termenul $P(x)$ este cel mai greu de calculat, însă datorită independenței sale de model, se poate neglija: când selectăm un model din mai multe posibile, valoarea lui $P(x)$ va fi constantă, deci nu influențează decizia. Prin urmare, putem aproxima formula (1) cu

$$P(\Theta|x) = P(x|\Theta)P(\Theta) \tag{3}$$

Inferența Bayes este un proces iterativ: după ce am calculat probabilitatea a posteriori pentru un set de date, această informație va deveni probabilitate a priori pentru următorul set de date.

2 Ghid Python

Pentru a rezolva următoarele exerciții aveți nevoie de pachetele `numpy` și `scipy.stats`.

Pentru a calcula probabiliatea unei valori cunoscute fiind media și dispersia distribuției normale cu care se modelează procesul, utilizați `scipy.stats.norm.pdf(x, mu, sigma)`. În mod similar, dacă este vorba de o distribuție uniformă, înlocuiți `norm` cu `uniform`.

3 Exerciții

Presupunem că ne dorim să modelăm prețul apartamentelor dintr-un anumit oraș. Nu avem acces la toate prețurile, ci doar la un subset de 10 astfel de valori, exprimate în mii de euro:
`x = [82, 106, 120, 68, 83, 89, 130, 92, 99, 89]`.

1. Pentru început considerăm valori cunoscute pentru medie și dispersie. Mai târziu, vom căuta valori optime pentru aceste mărimi. Setați $\mu = 90$ și $\sigma = 10$. Afișați distribuția normală ce corespunde acestor valori.

2. Calculați verosimiltatea de a obține valoarea de a obține $x_1 = 82$ din distribuția normală de mai sus. Întâi utilizați formula (2), apoi funcția `scipy.stats.norm.pdf(x, mu, sigma)`. Verificați că obțineți același rezultat.

3. Utilizați funcția de mai sus pentru a calcula verosimilitatea tuturor datelor x.

4. Presupuneți o distribuție normală pentru probabilitatea de a cunoște media a priori: în întreaga țară prețurile sunt distribuite normal și au o valoare medie de 100 și o dispersie de 50. De asemenea, presupuneți o probabilitate a priori pentru dispersie astfel încât aceasta să fie uniform distribuită în intervalul [1,70].

5. Calculați probabilitatea a posteriori pentru datele experimentale utilizând probilitatea a priori definită mai sus și verosimilitatea calculată anterior.

6. Până acum am testat un singur model, respectiv acela având parametrii $\mu = 90$ și $\sigma = 10$. Pentru a găsi, însă, modelul cel mai bun, va trebui să calculăm probabilitatea a posteriori pentru o serie de candidați și să-l selectăm pe acela cu probabilitatea ce mai mare. Utilizați următorul set de valori posibile ale parametrilor modelului $\mu = [70, 75, 80, 85, 90, 95, 100]$, respectiv $\sigma = [5, 10, 15, 20]$. Care din aceste modele este optim?