Modellezés Gauss-folyamatokkal

Csató Lehel

2005–2006 első félév, Október

Kivonat

Számítógépen tárolt adatok feldolgozásához matematikai modellek használata szükségszerű. A modellezés a vizsgált adatok tulajdonságait és az adatgyüjtés módját – a zaj típusát – fogalmazza meg egy számítógépen kódolható módon. A gyors feldolgozás érdekében illetve az adatgyűjtés során elkerülhetetlen zaj szűréséért egy számítógépes program feltételez egy „igazi", valódi adatot, amit nem tudunk megfigyelni, ellenben amely zajos és transzformált változata a rendelkezésre álló adat. Ezen modellek gyűjtőneve „rejtett változós modellek". A rejtett változók a megfigyelt adatok függvényei. Keressük tehát egy $f(\theta, \mathbf{x})$ függvény θ paraméterét, mely egy adott $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ adathalmazra optimális:

$$\hat{\theta} = \underset{\theta \in \Omega}{\operatorname{argmin}} \, L\left(\mathcal{D}, f(\cdot, \theta)\right)$$

ahol az Ω a függvény paramétereinek a definiciós tartománya és az $L(\mathcal{D}, f(\cdot, \theta))$ hibafüggvény méri a θ paraméter által választott függvény illeszkedését az adatokhoz.

A fenti modellezés hátránya, hogy a függvény értékeivel párhuzamosan nem becsüljük az egyes értékekhez tartozó *bizonytalanságot*, azaz a függvény értékeinek a szórását adott pontokban. A bizonytalanság becslésének egy módszere a valószínűségi modellek használata: itt a függvény paraméterének *egy* értéke – tehát deteminisztikus becslés – helyett a lehetséges értékek eloszlását szeretnénk megtalálni, tehát valószínűségi becslést szeretnénk végezni. Az értékek eloszlásából megállapítható az egyes értékekhez tartozó bizonytalanság.

Valószínűségi modellezésben tehát a kérdés szintén egy $f(\cdot, \theta)$ függvény θ paraméterének a becslése, azonban itt a paraméter *eloszlását* – illetve annak a változását – szeretnénk meghatározni a $\mathcal{D}$ adathalmaz függvényeként. Ebben az esetben a modell része a θ paraméter $p_0(\theta)$ a–priori eloszlása, durván az általunk feltételezett lehetséges paraméter értékeknek az eloszlása. Szintén a modell része a fentebb is említett zajmodell, tehát a „valódi" $f(\mathbf{x}_n, \theta)$ értékek és a megfigyelt zajos y_n „adatok" közötti kapcsolat, a $P(y_n|f(\mathbf{x}_n, \theta))$ eloszlás, a $\mathcal{D}$ adathalmazra meg a $P(\mathcal{D}|\theta) = \prod_n P(y_n|f(\mathbf{x}_n, \theta))$ feltételes valószínűség. A $p_{\text{post}}(\theta|\mathcal{D})$ *a–posteriori* eloszlást Bayes tétele adja meg:

$$p_{\text{post}}(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)p_0(\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta)p_0(\theta)}{\int_\Omega d\theta \; P(\mathcal{D}|\theta)}.$$

A Bayes-becslés hátránya az alkalmazásának a nehézsége: az a–posteriori eloszlás nem – illetve nagyon ritkán – írható fel paraméteres modellként. Ebben az esetben még egy optimálási feladatot meg kell oldani: $\hat{p}(\theta) = \operatorname{argmin}_p d(p_{\text{post}}(\theta)|p(\theta))$, vagyis egy d-optimálisan illeszkedő paraméterezett eloszlás keresését. Egy másik hátrány az $f(\mathbf{x}_n, \theta)$ függvény paraméteres jellege: nagyon gyakran nem tudjuk, hogy mekkora legyen a komplexitása a függvényosztálynak: az adatok megkétszerezése esetén több, kevesebb adat esetén pedig kevesebb függvény közül szeretnénk választani. A Gaussz-folyamatokkal történő modellezésben az $f(\mathbf{x}, \theta)$ függvény és a $p_0(\theta)$ specifikációját egy lépésben oldjuk meg: keressük az $f(\mathbf{x})$ *véletlen–függvény* eloszlásának paramétereit. A függvények *a–priori* eloszlása Gaussz folyamat, azaz minden lehetséges $\mathcal{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]$ halmazra a függvény $\mathbf{f}(\mathcal{X}) = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_K)]$ értékei Gaussz eloszlásúak és az *a–posteriori* folyamatot is Gaussz–folyamattal közelítjük. Az előadás során a Gaussz-folyamatokkal történő becslések specifikumairól, a becslések során adódó nehézségekről valamint azok megoldásairól fogok beszélni.