A maximum entrópia elv

2018. június 15.

Legyen X egy diszkrét valószínűségi változó, amely $x_1, \ldots, x_n$ értékeket vehet fel, és ezek valószínűsége $P(X = x_i) = p_i$. Annak az információértéke, hogy X az x_i értéket veszi fel,

$$I(x_i) = -\ln p_i.$$

Így pl. egy biztos esemény ($p_i = 1$) bekövetkezésének információtartalma 0, egy valószínűtlen eseményé pedig egy nagy szám lesz. A természetes logaritmus helyett az információ-elméletben 2-es alapú logaritmust szoktak használni[1], de ennek a maximum entrópia szempontjából nincsen jelentősége.

Az entrópia a rendszerben levő rendezetlenség mértéke. Minél rendezettebb a rendszer, várhatóan annál kevésbé fogunk meglepődni a viselkedésén, vagyis annál kevesebb lesz a várható információtartalma. Következésképp az entrópiát az információ várható értékével lehet jellemezni:

$$H(X) = E[I(X)] = \sum_{i=1}^{n} p_i I(x_i) = -\sum_{i=1}^{n} p_i \ln p_i.$$

Tegyük fel most, hogy a p_i $(i = 1, \ldots, n)$ valószínűségi eloszlás nem ismert, viszont vannak rá különféle megkötéseink. A két kötelező kényszer az alábbi:

$$0 \leq p_i \leq 1 \quad \text{és} \quad \sum_{i=1}^{n} p_i = 1.$$

A maximum entrópia elve azt mondja ki, hogy egy ilyen feladat megoldásakor azt az eloszlást kell választanunk, amelyiknek az entrópiája maximális, tehát amelyik a legkevésbé rendezett a feltételeket teljesítő eloszlások közül, ezzel minimalizálva a részrehajlást (*bias*). A rendszer Lagrange-szorzókkal oldható meg. Ha csak a kötelező feltételek adottak, akkor az egyenletes eloszlást kapjuk, ahogy ez várható:

$$\mathcal{L}(p_1, \ldots, p_n, \lambda) = -\sum_{i=1}^{n} p_i \ln p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right),$$

$$\frac{\partial}{\partial p_i} \mathcal{L}(p_1, \ldots, p_n, \lambda) = -\ln p_i - p_i \frac{1}{p_i} + \lambda = 0,$$

amiből $p_i = \exp(\lambda - 1)$, tehát minden p_i azonos, így a kényszer (a λ szerinti parciális derivált) miatt $p_i = \frac{1}{n}$.

[1] Ez annak felel meg, hogy hány bit kell a változó optimális kódolásához, pl. $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$ valószínűségek esetén az információtartalom rendre 1, 2, 2, és ennek megfelelően egy optimális (várható értékben minimális hosszú) kódolás a 0, 10, 11 bitsorozatokkal történhet.

Az elv egy „bayes-i" kiterjesztését kapjuk, ha feltételezzük, hogy ismerünk egy *a priori* eloszlást $(m_1, \dots, m_n)$, és az ehhez képest vett várható információtöbbletet minimalizáljuk[2]:

$$\begin{aligned}\min E[I_m(X) - I(X)] &= \min \sum_{i=1}^{n} p_i(I_m(x_i) - I(x_i)) \\ &= \min \sum_{i=1}^{n} p_i \left(-\ln m_i + \ln p_i\right) \\ &= \min \sum_{i=1}^{n} p_i \ln \frac{p_i}{m_i} \\ &= \max - \sum_{i=1}^{n} p_i \ln \frac{p_i}{m_i}.\end{aligned}$$

Ha az *a priori* eloszlás egyenletes, akkor ez ugyanazt az eredményt adja, mint az eredeti maximum entrópia elv; és ha nincsen egyéb kényszer, akkor a maximális eloszlás megegyezik az *a priori* eloszlással. A minimalizált mennyiséget hívják relatív entrópiának ill. Kullback–Leibler divergenciának is. Ez működik folytonos valószínűségi változókra is:

$$\max - \int p(x) \ln \frac{p(x)}{m(x)}\, \mathrm{d}x.$$

Érdekes párhuzam van a baricentrikus koordináták és a diszkrét valószínűségek közt: ha extra kényszerként hozzávesszük a csúcs-rekonstrukciót, tehát

$$\sum_{i=1}^{n} p_i v_i = v,$$

akkor a maximum entrópia elvből adódó $b_i(v) = p_i$ valószínűségek baricentrikus koordináták lesznek. A Lagrange-tulajdonság $(b_i(v_j) = \delta_{ij})$ azonban általánosan csak szigorúan konvex esetben teljesül. Megfelelő *a priori* eloszlás választásával azonban ez is megoldható, ugyanis a 0-értékek mindig átöröklődnek, tehát ha az m_i-kre teljesült a Lagrange-tulajdonság, akkor az optimalizációból kapott p_i koordinátákra is teljesülni fog. Itt érdemes még megjegyezni, hogy az m_i nem muszáj, hogy valószínűségi eloszlás legyen, a módszer tetszőleges függvénnyel működik.

Hivatkozások

[1] K. Hormann, N. Sukumar, *Maximum entropy coordinates for arbitrary polytopes.* Computer Graphics Forum, Vol. 27(5), pp. 1513–1520, 2008.

[2] Információ-elméleti szemszögből nézve, amikor a p_i eloszlású X változót az m_i eloszlás alapján kódoljuk, nem optimális a kódolás, és hosszabb kódokat kapunk; a kiterjesztett maximum entrópia elv a várható többletbit-szám minimalizálásának felel meg.