Bayesi dobókocka

Az induktív következtetés axiomatikus tárgyalására több formális keretet is javasoltak, ismeretanyagában azonban egyikük sem mérhető a filozófus Thomas Bayes és a matematikus Pierre-Simon Laplace által megalapazott valószínűségelmélethez. Az újabb és újabb tudományos áttörésekkel, a „tudományos módszer" terjedésével párhuzamosan a XX. századra a bayesi episztemológia, bayesi konfirmációelmélet és tanuláselmélet, vagyis összefoglalva a „bayesiánus" filozófia vezető irányzattá vált. Mára a gépi tanulás átfogó nyelvének szerepét is ez az elmélet tölti be.

A bayesi megközelítés alapfogalmainak megértéséhez tekintsünk egy egyszerű példát. Vegyünk egy hatoldalú dobókockát, amely vagy cinkelt, vagy nem – ezt előre persze nem tudjuk. Jelölje θ annak valószínűségét, hogy a d dobás eredménye mondjuk hármas lesz, azaz

$$p\left(d = 3 \mid \theta\right) = \theta.$$

Célunk, hogy θ értékét megismerjük. Ehhez Bernoulli-kísérleteket végzünk, majd megszámláljuk, hogy hány esetben kaptunk hármast, és hány esetben mást. Annak valószínűségét, hogy éppen adott számú hármas van a dobásaink között, a binomiális eloszlás mondja meg:

$$\ell(\theta) = p\left(D \mid \theta\right) = \binom{t+o}{t} \cdot p\left(d = 3 \mid \theta\right)^t \cdot p\left(d \neq 3 \mid \theta\right)^o = \binom{t+o}{t} \cdot \theta^t \cdot (1-\theta)^o,$$

ahol t a hármasok száma, o a más dobásoké, D pedig az összes dobásunkat jelöli. Ezt a mennyiséget az adatok *likelihood*-jának (ℓ) nevezzük.

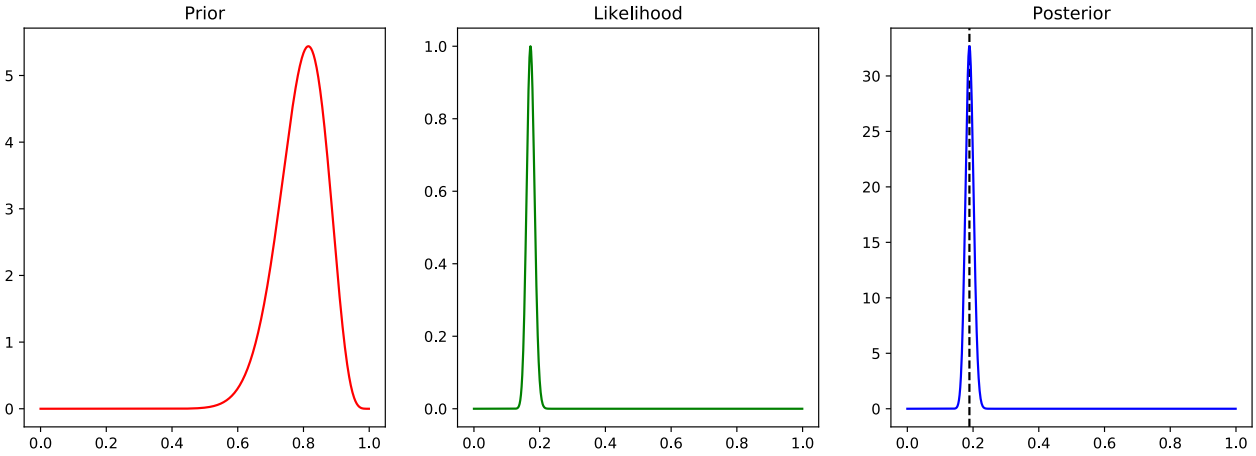
Egy adott dobássorozat eredményére feltehetjük a kérdést: vajon milyen θ érték mellett volt ez a kimenetel a legvalószínűbb? Nincs más dolgunk, mint az előbbi ℓ függvény maximumának megkeresése, ami – emelt szintű matematika érettségi birtokában legalábbis – gyerekjáték:

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta} = \binom{t+o}{t} \cdot \left[t \cdot \theta^{t-1} \cdot (1-\theta)^o - o \cdot \theta^t \cdot (1-\theta)^{o-1}\right] = 0, \\
t \cdot (1-\theta) - o \cdot \theta = 0, \\
\theta = \frac{t}{t+o}.
\end{aligned}$$

Logikus eredményre jutottunk: θ a hármas dobások aránya az összes dobáshoz viszonyítva. Ezt a megoldást nevezik *Maximum Likelihood (ML)* megoldásnak, mivel a likelihood-függvényt maximalizáltuk.

A bayesi megközelítésben a valószínűség szubjektivista értelmezését szoktuk követni, azaz a valószínűségre egy adott eseménybe vetett hit mértékeként gondolunk. Ilyen hiedelmeink előzetesen is lehetnek, mielőtt még bármilyen kísérletet végeznénk. Jelölje a θ értékét illető előzetes hiedelmünket $p\left(\theta\right)$, amelyet *prior*-nak is nevezünk. Ekkor felírhatjuk Bayes tételét:

$$p\left(\theta \mid D\right) = \frac{p\left(D \mid \theta\right) p\left(\theta\right)}{p\left(D\right)},$$



1. ábra. A prior, likelihood és posterior függvény $\beta(23, 6)$ eloszlású prior és 1000 megfigyelés esetén. Látható, hogy a teljesen téves előzetes hiedelmeinket korrigáltuk a megfigyelések felhasználásával.

amelyet egyébként magunk is könnyen bizonyíthatunk a $p(D, \theta) = p(\theta \mid D)\, p(D) = p(D \mid \theta)\, p(\theta)$ összefüggésből. A fenti formula bal oldalát *posterior*-nak nevezzük. Mivel θ értékére vagyunk kíváncsiak, sőt, valójában csak maximumot keresünk, a nevezőt akár el is hagyhatjuk, azaz

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}.$$

Itt ütközünk az első problémába. Bár a jobb oldalt könnyen kiszámolhatjuk, ennek maximalizálása rendszerint bonyolult feladat – általában olyan egyenletre vezet, amelyet nem tudunk megoldani (emlékezzünk például arra, hogy megoldóképlet már ötödfokú egyenletre sem létezik). Például Gauss-eloszlást feltételezve a posterior

$$p(\theta \mid D, \mu, \sigma) = \binom{t+o}{t} \cdot \theta^t \cdot (1-\theta)^o \cdot \mathcal{N}\left(\theta \mid \mu, \sigma^2\right) \propto \theta^t \cdot (1-\theta)^o \cdot \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\},$$

amelynek maximalizálása csak numerikus módszerekkel lehetséges (pl. Newton–Raphson).

Hiedelmeinket tehát célszerű kifejezetten olyan formában megfogalmazni, hogy a posterior analitikusan számolható maradjon. Próbálkozzunk tehát újra, de ezúttal priorként béta-eloszlást használva:

$$p(\theta \mid a, b) = \beta(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1},$$

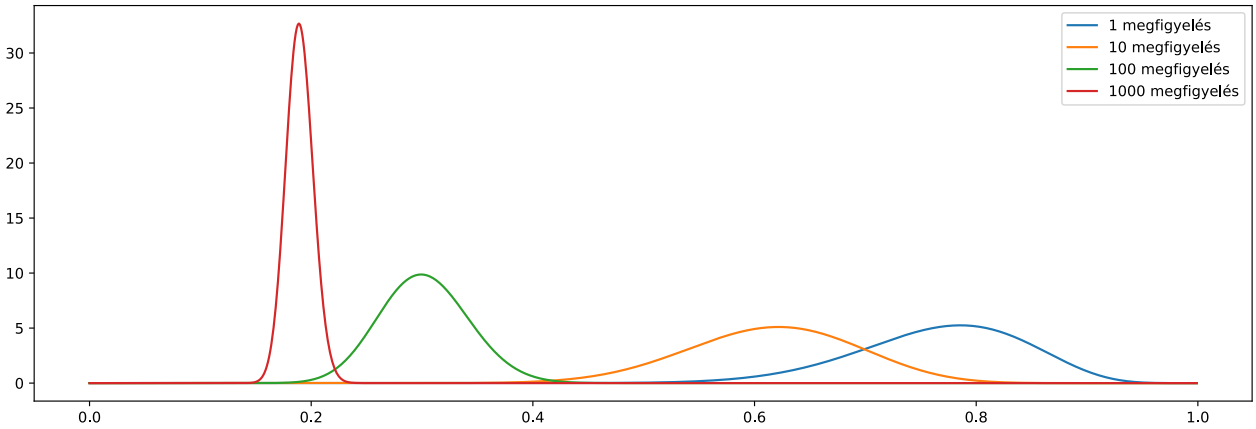
és a posterior kiszámításához ismét felhasználjuk Bayes tételét:

$$p(\theta \mid D, a, b) \propto p(D \mid \theta) \cdot p(\theta \mid a, b) \propto \theta^{t+a-1} \cdot (1-\theta)^{o+b-1}.$$

Vegyük észre, hogy ez egy arányossági tényezőtől eltekintve nem más, mint egy béta-eloszlás! Sőt, mivel tudjuk, hogy a posterior egy valószínűségi eloszlás, az aranyossági tényező automatikusan adódik (hiszen a görbe alatti terület 1 kell, hogy legyen), tehát a posterior pontosan egy béta eloszlás:

$$p(\theta \mid D, a, b) = \beta(\theta \mid a + t, b + o).$$

Az olyan priorokat, amelyek a likelihood-dal kombinálva ugyanolyan formájú posteriorokat adnak, *konjugált* prioroknak nevezzük. Látjuk tehát, hogy a béta-eloszlás konjugált prior a binomiális eloszlásra nézve.



2. ábra. A posterior alakulása a megfigyelések számának növekedésével. Látjuk, hogy a priorról elindulva a posterior maximuma lassan felveszi a helyes értéket.

Itt egy másik érdekes megfigyelést is tehetünk. Az előzetes hiedelmünket a prior a és b paraméterei kódolták; a posterior alakját figyelembe véve azt vesszük észre, hogy ezek afféle „virtuális megfigyelések”: a hármasok számát a kísérletek előtt a-nak, az egyéb dobásokét b-nek tettük fel, majd ehhez jöttek hozzá a kísérletek eredményei, így egy mindkettőt magába foglaló poszteriorhoz jutottunk. A θ paraméter értéke a korábbi számításunkkal analóg módon

$$\theta = \frac{a + t - 1}{a + t + b + o - 2}.$$

Ezt a megoldást, amely immár az előzetes hiedelmeinket magába foglalja, *Maximum A Posteriori (MAP)* megoldásnak nevezzük.

Ez a megoldás még mindig nem használja ki a bayesi valószínűségelmélet teljes erejét. Gondoljunk bele, hogy valójában minket nem is igazán a θ paraméter legvalószínűbb értéke érdekel (hiszen az, hogy legvalószínűbb, egyáltalán nem jelent bizonyosságot), hanem az, hogy milyen eséllyel fogunk hármast dobni. Ehhez pedig figyelembe kell vennünk – bármily valószínűtlen – θ *összes* lehetséges értékét, ezeket az eseteket a valószínűségükkel súlyozva. Matematikailag ezt integrálszámítással érjük el:

$$\begin{aligned} p\left(d = 3 \mid D, a, b\right) &= \int p\left(d = 3 \mid \theta\right) p\left(\theta \mid D, a, b\right) d\theta \\ &= \int \theta \cdot \beta\left(\theta \mid a + t, b + o\right) d\theta \\ &= \mathbb{E}\left[\beta\left(\theta \mid a + t, b + o\right)\right] \\ &= \frac{a + t}{a + t + b + o}, \end{aligned}$$

ahol a béta-eloszlás várható értékére vonatkozó összefüggést használtuk fel. Ezt a módszert nevezzük *bayesi modellátlagolás*-nak.