14. Procjena parametara II

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, predavanja, v2.2

Prošli put krenuli smo pričati o procjeni parametara – mehanizmu učenja probabilističkih modela. Malo smo ponovili statistiku, osnovne razdiobe i objasnili što je to procjenitelj. Rekli smo da postoje tri osnovne vrste procjenitelja: **MLE, MAP i bayesovski**. Mi ćemo raditi MLE i MAP.

Procjenitelj MLE je najjednostavniji i često se koristi. Međutim, vidjet ćemo da s njim imamo jedan problem, a to je **prenaučenost**. To će nam biti motivacija za MAP procjenitelj. MAP procjenitelj će nam omogućiti da u procjenu ugradimo svoje pozadinsko znanje, i na taj način izbjegnemo ili barem ublažimo prenaučenost.

1 Funkcija izglednosti

Naša današnja priča započinje sa **funkcijom izglednosti** (engl. *likelihood function*). Ideja izglednosti jedna je od osnovnih ideja u statistici i, kao što ćemo vidjeti, temelj za procjenu parametara metodama MLE i MAP. Zapravo, mi smo funkciju izglednosti već susreli, kada smo izvodili funkcije empirijske pogreške za poopćene linearne modele, samo što je nismo tako zvali. Prisjetite se: ideja je bila da napišemo vjerojatnost oznaka na temelju predikcije modela, odnosno, iz razloga matematičke jednostavnosti, logaritam te vjerojatnosti. Zatim smo tražili parametre koji maksimiziraju logaritam vjerojatnosti oznaka ili, ekvivalentno, minimiziraju negativan logaritam vjerojatnosti oznaka, a za koji smo onda utvrdili da je jednak ili proporcionalan empirijskoj pogrešci koju želimo minimizirati. Na taj smo način izveli funkciju pogreške i funkciju gubitka nekoliko modela. Vratimo se sada opet na to i pogledajmo detaljnije o čemu se radi.

Kao i uvijek u strojnom učenju, krećemo od skupa podataka. Neka je to neki općeniti skup podataka, $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$. To mogu biti primjeri, ali mogu primjeri i oznake, ili samo oznake – nije bitno, metoda će vrijedi univerzalno, neovisno o tome što su točno podatci. Skup $\mathcal{D}$ zapravo je **slučajan uzorak** koji smo uzorkovali iz populacije, tj. skupa svih mogućih primjera $\mathcal{X}$. Pretpostavit ćemo da se primjeri $\mathbf{x} \in \mathcal{X}$ ravnaju po nekoj distribuciji. To pišemo ovako: [1]

$$\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$$

gdje je $\boldsymbol{\theta}$ vektor parametara te distribucije. Primijetite da p ovdje označava funkciju gustoće vjerojatnosti, no distribucija, naravno, može biti i diskretna. Budući da je $\mathcal{D}$ slučajan uzorak iz $\mathcal{X}$, to je on onda reprezentativan uzorak, pa možemo pretpostaviti i da se primjeri $\mathbf{x}^{(i)} \in \mathcal{D}$ također ravnaju po istoj distribuciji kao i primjeri iz populacije $\mathcal{X}$:

$$\mathbf{x}^{(i)} \sim p(\mathbf{x}|\boldsymbol{\theta})$$

Sada je ideja da napišemo **vjerojatnost uzorka** – vjerojatnost da smo iz dotične distribucije $p(\mathbf{x}|\boldsymbol{\theta})$ izvukli baš taj uzorak $\mathcal{D}$:

$$p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Primijetite da smo ovdje (kod druge jednakosti) upotrijebili još jednu pretpostavku: pretpostavku **i.i.d** (nezavisno i identično distribuirani primjeri). Naime, ako su primjeri $\mathbf{x}^{(i)}$ i.i.d., onda vrijedi $p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = p(\mathbf{x}^{(i)})p(\mathbf{x}^{(j)})$, pa vjerojatnost slučajnog vektora možemo napisati kao produkt vjerojatnosti pojedinačnih slučajnih varijabli.

Sada je bitno da shvatimo da vjerojatnost $p(\mathcal{D}|\boldsymbol{\theta})$ u stvarnosti ovisi samo o parametru $\boldsymbol{\theta}$. Naime, u stvarnosti je uzorak $\mathcal{D}$ **fiksiran** – to je skup podataka kojim raspolažemo, i imamo samo taj jedan skup i ne možemo ga više mijenjati. Dakle, ono što je ovdje varijabilno je parametar $\boldsymbol{\theta}$, dok je $\mathcal{D}$ fiksiran. Kako bismo to naglasili, napisat ćemo eksplicitno da je vjerojatnost $p(\mathcal{D}|\boldsymbol{\theta})$ zapravo **funkcija od parametra** $\boldsymbol{\theta}$, za što uvodimo novu oznaku, $\mathcal{L}$:

$$p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\theta) \equiv \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$$

Funkcija $\mathcal{L}$ zove se **funkcija izglednosti** (engl. *likelihood function*). Ona parametrima $\boldsymbol{\theta}$ pridjeljuje vjerojatnost da iz populacije s parametrima $\boldsymbol{\theta}$ izvučemo uzorak $\mathcal{D}$:

$$\mathcal{L} : \boldsymbol{\theta} \mapsto p(\mathcal{D}|\boldsymbol{\theta})$$

Dakle, kako bismo od vjerojatnosti $P(\mathcal{D}|\boldsymbol{\theta})$ ili gustoće vjerojatnosti $p(\mathcal{D}|\boldsymbol{\theta})$ dobili funkciju izglednosti parametra $\boldsymbol{\theta}$, jednostavno tu vjerojatnost odnosno gustoću vjerojatnosti promatramo kao funkciju od $\boldsymbol{\theta}$ a ne od $\mathcal{D}$. Notacijski, jednostavno zamijenimo argumente: parametar $\boldsymbol{\theta}$ postaje varijabla funkcije, a varijabla $\mathcal{D}$ postaje parametar, pa pišemo $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$.

Jako je važno da primijetimo da izglednost parametara $\boldsymbol{\theta}$ nije vjerojatnost parametara $\boldsymbol{\theta}$, nego da je to vjerojatnost skupa podataka $\mathcal{D}$ za distribuciju s parametrima $\boldsymbol{\theta}$. Također kažemo da je to "vjerojatnost podataka pod modelom". Sljedeći primjer će to malo razjasniti.

▸ **PRIMJER**

Pogledajmo kako bi izgledala **izglednost parametra Bernoullijeve distribucije**.

Neka je naš skup podataka dobiven bacanjem novčića i bilježenjem jesmo li dobili glavu ili pismo. Ishod da dobijemo glavu modeliramo **Bernoullijevom varijablom**: $x = 1$ znači da smo dobili glavu, a $x = 0$ da smo dobili pismo. Ishod Bernoullijeve varijable definira **Bernoullijeva distribucija**, koja ima parametar μ (vjerojatnost da dobijemo $x = 1$):

$$P(X = x|\mu) = \mu^x(1-\mu)^{1-x}$$

Skup $\mathcal{D}$ čini 10 bacanja novčića ($N = 10$). Glavu smo dobili 8 puta, a pismo 2 puta. Za takav skup podataka, funkcija izglednosti je:

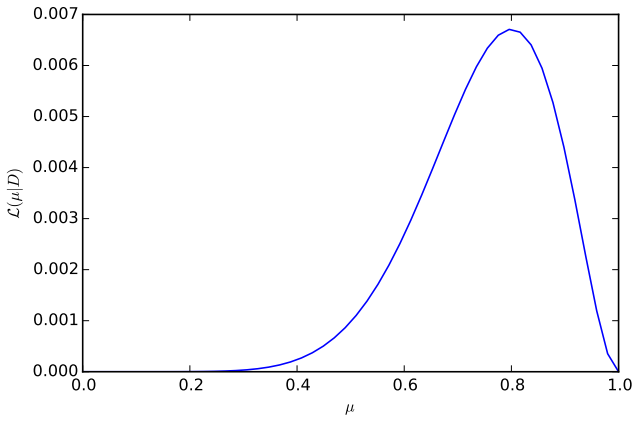
$$\begin{aligned}\mathcal{L}(\mu|\mathcal{D}) &= P(\mathcal{D}|\mu) = P(x^{(1)}, x^{(2)}, \ldots, x^{(10)}|\mu) \\ &= \prod_{i=1}^{10} P(x^{(i)}|\mu) = \mu \cdot \mu \cdot \mu \cdot \mu \cdot \mu \cdot \mu \cdot \mu \cdot \mu \cdot (1-\mu) \cdot (1-\mu) = \mu^8(1-\mu)^2\end{aligned}$$

Općenito, ako u N bacanja novčića imamo m glava, onda je funkcija izglednosti jednaka:

$$\mathcal{L}(\mu|\mathcal{D}) = \mu^m(1-\mu)^{(N-m)}$$

gdje je $m = \sum x^{(i)}$ broj glava.

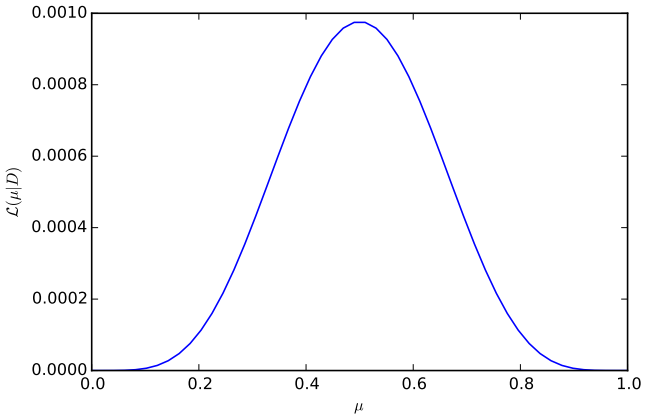
Za konkretan slučaj $m = 8$ i $n = 10$, graf funkcije izglednosti izgleda ovako:



$m = 8$, $N = 10$

Ovdje opet napominjemo da ovo nije vjerojatnost, odnosno, preciznije, funkcija $\mathcal{L}$ nije funkcija gustoće vjerojatnosti. Stoga općenito ne vrijedi $\int_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\,\mathrm{d}\boldsymbol{\theta} = 1$. Ovaj primjer to dobro pokazuje, budući da je integral ovdje sasvim sigurno manji od jedinice.

Funkcija izglednosti $\mathcal{L}$ je funkcija parametra μ, dok je skup podataka $\mathcal{D}$ fiksiran. Skup podataka ovdje smo saželi u dva broja: m i N. Da smo imali drugačiji uzorak $\mathcal{D}$, odnosno da su m i N bili drugačiji, onda bi i funkcija izglednosti bila drugačija. Na primjer, da smo u 10 bacanja dobili 5 puta glavu, onda bi graf funkcije izglednosti bio ovakav:



$m = 5$, $N = 10$

2 Procjenitelj MLE

Izglednost je, dakle, funkcija koja nam za neke parametre $\boldsymbol{\theta}$ kazuje koliko je, uz te parametre, vjerojatno da nam se dogodi skup podatka $\mathcal{D}$ kojim raspolažemo. Sada je pitanje: kako to iskoristiti da nađemo parametre modela odnosno parametre vjerojatnosne distribucije?

Ovdje pomaže da se nakratko vratimo na gornja dva primjera s bacanjem novčića. Pogledajte graf funkcije izglednosti za prvi slučaj, za skup $\mathcal{D}$ za koji $m = 8$ i $N = 10$. Na temelju tog grafa, za koju biste vrijednost parametra μ rekli da vrijedi za naš novčić, koliko je vjerojatno da dobijemo glavu? Sigurno biste odabrali $\mu = 0.8$, jer je uz tu vrijednost parametra μ vjerojatnost skupa $\mathcal{D}$, tj. vjerojatnost da od deset bacanja dobijemo osam glava, najveća. Drugim riječima, vrijednost 0.8 je *najizglednija* vrijednost parametra μ. Pogledajte sada graf funkcije izglednosti za drugi slučaj, za skup $\mathcal{D}$ za koji $m = 5$ i $N = 10$. Ovdje je najizglednija vrijednost parametra

μ jednaka 0.5, tj. vjerojatnosti glave i pisma su jednake.

Ideja je, dakle, da se postavimo vrlo pragmatično: ako smo dobili uzorak $\mathcal{D}$, onda mora da je baš on najvjerojatniji mogući, inače ga ni ne bismo dobili! Ako je tako, hajde da nađemo parametre koji naš uzorak čine najvjerojatnijim. Drugim riječima, hajde da nađemo koji parametri **maksimiziraju funkciju izglednosti**. Upravo to je **procjenitelj najveće izglednosti** (engl. *maximum likelihood estimator; MLE*). Dakle, MLE nalazi $\boldsymbol{\theta}$ koji maksimiziraju funkciju izglednosti. 2
Formalno:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$$

Kao i mnogo puta do sada, pokazat će se da je matematički često jednostavnije maksimizirati logaritam izglednosti, tzv. **log-izglednost**:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \big(\ln \mathcal{L}\big(\boldsymbol{\theta}|\mathcal{D}\big)\big)$$

Naravno, činjenica da maksimiziramo logaritam funkcije izglednosti, a ne izravno funkciju izglednosti, ništa ne mijenja na stvari jer je logaritam monotono rastuća funkcija, pa su maksimizatori u oba slučaja identični. Kad god je to moguće i kada se računalno isplati, maksimizaciju ćemo provesti analitički (tj. nalaženjem rješenja u zatvorenoj formi), a kada to nije moguće ili se ne isplati, optimizaciju ćemo provesti iterativnim metodama, npr. gradijentnim spustom.

Sažmimo ovu fantastičnu ideju. Pred sobom imam neki **uzorak podataka**. Činjenica da se baš taj uzorak podataka našao kod mene uzimam kao znak da je upravo taj uzorak bio najvjerojatniji mogući uzorak, inače bih dobio neki drugi uzorak. Ako pretpostavim da se podatci pokoravaju nekoj distribuciji, čiji su mi parametri još nepoznati, mogu napisati **funkciju izglednosti**, koja mi daje vjerojatnost uzorka u ovisnosti o parametrima distribucije. Ako je uzorak koji imam najvjerojatniji, onda je moja najbolja procjena za parametre distribucije upravo ona koja taj moj uzorak čini najvjerojatnijim, tj. ona koja **maksimizira funkciju izglednosti** (odnosno log-izglednosti).

Pokažimo sada kako izvesti procjenitelje MLE za osnovne vjerojatnosne distribucije koje smo bili ponovili prošli put: Bernoullijeva, multinulijeva, Gaussova i multivarijatna Gaussova. U svim tim slučajevima MLE ćemo moći izvesti analitički – deriviranjem i izjednačavanjem s nulom (nalaženjem stacionarne točke funkcije izglednosti).

2.1 MLE za Bernoullijevu distribuciju

Već smo rekli da je funkcija log-izglednosti za parametar μ Bernoullijeve distribucije sljedeća:

$$\begin{aligned}\ln \mathcal{L}(\mu|\mathcal{D}) &= \ln \prod_{i=1}^{N} P(x|\mu) = \ln \prod_{i=1}^{N} \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}} \\ &= \sum_{i=1}^{N} x^{(i)} \ln \mu + \Big(N - \sum_{i=1}^{N} x^{(i)}\Big) \ln(1-\mu)\end{aligned}$$

Maksimizaciju log-izglednosti provodimo nalaženjem nul-točke prve derivacije funckije:

$$\begin{aligned}\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}\mu} &= \frac{1}{\mu} \sum_{i=1}^{N} x^{(i)} - \frac{1}{1-\mu}\Big(N - \sum_{i=1}^{N} x^{(i)}\Big) = 0 \\ \Rightarrow \quad \hat{\mu}_{\text{MLE}} &= \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = \frac{m}{N}\end{aligned}$$

gdje u indeksu pišemo "MLE" kako bismo naznačili da je procjenitelj izveden metodom najveće izglednosti.

Vidimo da je procjena za μ zapravo broj realizacija ishoda 1 podijeljen s veličinom uzorka. Statistički termin za to je **relativna frekvencija**: broj realizacija ishoda 1 podijeljen s veličinom uzorka. Relativna frekvencija je i vrlo intuitivan procjenitelj za parametar μ: naime, ako smo 10 puta bacili novčić i od toga 8 puta dobili glavu, onda bismo rekli da je vjerojatnost glave jednaka $8/10 = 0.8$. Upravo to je MLE procjena za μ. Zapravo, ispada da svakodnevno radimo MLE. Vrijedi $\mathbb{E}(\mu_{\text{MLE}}) = \mathbb{E}[X] = \mu$, pa je ovo je nepristran procjenitelj.

2.2 MLE za kategoričku (multinulijevu) distribuciju

Prisjetite se (od prošlog puta) da vjerojatnost kategoričke varijable sa K mogućih vrijednosti definiramo ovako:

$$P(X = \mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

gdje je vektor $\boldsymbol{\mu}$ parametar ove distribucije. Funkcija log-izglednosti za taj parametar je:

$$\ln \mathcal{L}(\boldsymbol{\mu}|\mathcal{D}) = \ln \prod_{i=1}^{N} P(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = \ln \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_k^{(i)}} = \sum_{k=1}^{K} \sum_{i=1}^{N} x_k^{(i)} \ln \mu_k$$

Kako bismo izrazili MLE, trebamo maksimizirati ovu funkciju. Međutim, za razliku od gornjeg slučaja s Bernoullijevom varijablom, ovdje imamo ograničenje $\sum_{k=1}^{K} \mu_k = 1$, što znači da ovdje govorimo o problemu **optimizacije uz ograničenje**. Tu optimizaciju možemo provesti (nama već poznatom) metodom **Lagrangeovih multiplikatora**. Ovdje ćemo se toga poštedjeti te dati samo konačan razultat. MLE za k-tu komponentu vektora $\boldsymbol{\mu}$ je: 3

$$\hat{\mu}_{k,\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} x_k^{(i)} = \frac{N_k}{N}$$

gdje je N_k je broj nastupanja k-te vrijednosti. Vidimo da je ovo posve analogno procjenitelju MLE za Bernullijevu varijablu, jedino što ovdje imamo posebnu procjenu za svaku varijablu k (koje odgovaraju pojedinačnim vrijednostima kategorijske varijable). Ako je $K = 2$, procjenitelj očekivano degenerira na procjenitelj MLE za Bernoullijevu varijablu.

2.3 MLE za Gaussovu distribuciju

Gaussova distribucija je distribucija kontinuirane varijable, dakle ovdje ćemo raditi s funkcijom gustoće vjerojatnosti p, a ne s vjerojatnošću P. Prisjetimo se najprije (od prošlog puta) funkcije Gaussove gustoće vjerojatnosti:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Ovdje sada imamo dva parametra, μ i σ^2, pa će funkcija log-izglednost biti funkcija ta dva parametra:

$$\begin{aligned} \ln \mathcal{L}(\mu, \sigma^2|\mathcal{D}) &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x^{(i)}-\mu)^2}{2\sigma^2} \right\} \\ &= -\frac{N}{2} \ln(2\pi) - N \ln \sigma - \frac{\sum_i (x^{(i)}-\mu)^2}{2\sigma^2} \end{aligned}$$

Također, budući da imamo dva parametra, maksimizaciju trebamo provesti po oba ta parametra. Dobivamo (izvod preskačemo):

$$\begin{aligned}
&\nabla \ln \mathcal{L}(\mu, \sigma^2|\mathcal{D}) = 0 \\
&\quad \Rightarrow \quad \hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)} \\
&\quad \Rightarrow \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} (x^{(i)} - \hat{\mu}_{\text{MLE}})^2
\end{aligned}$$

Primijetite da je procjenitelj $\hat{\sigma}^2_{\text{MLE}}$ izražen pomoću procjenitelja $\hat{\mu}_{\text{MLE}}$, zato jer nam je prava vrijednost parametra μ nepoznata.

Za procjenitelj srednje vrijednosti već smo prošli puta ustanovili da je nepristran. Međutim, ustanovili smo da je ovakav procjenitelj varijance pristran, i da ga možemo korigirati ako podijelimo sa $N-1$ umjesto s N. Vidimo, dakle, da procjenitelj najveće izglednosti ne mora nužno biti nepristran. Najveća izglednost nije isto što i nepristranost!

2.4 MLE za multivarijatnu Gaussovu distribuciju

Pogledajmo sada multivarijatnu Gaussovu distribuciju. Prisjetimo se, funkcija gustoće multivarijante Gaussove distribucije definirana je ovako:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big\}$$

gdje su $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ srednja vrijednost odnosno kovarijacijska matrica. Funkcija log-izglednosti onda je jednaka:

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) &= \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= -\frac{nN}{2}\ln(2\pi) - \frac{N}{2}|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}^{(i)}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu})
\end{aligned}$$

Maksimizacija funkcije log-izglednosti (izvod preskačemo) daje:

$$\begin{aligned}
&\nabla \ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) = 0 \\
&\quad \Rightarrow \quad \hat{\boldsymbol{\mu}}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \\
&\quad \Rightarrow \quad \hat{\boldsymbol{\Sigma}}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\text{MLE}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\text{MLE}})^{\mathrm{T}}
\end{aligned}$$

Procjenitelji $\hat{\boldsymbol{\mu}}_{\text{MLE}}$ i $\hat{\boldsymbol{\Sigma}}_{\text{MLE}}$ analogni su procjeniteljima $\hat{\mu}_{\text{MLE}}$ odnosno $\hat{\sigma}^2_{\text{MLE}}$ univarijatne Gaussove razdiobe. Također vrijede ista zapažanja što se tiče (ne)pristranosti procjenitelja.

2.5 MLE za parametre linearne regresije

Pokazali smo kako izvesti MLE za osnovne distribucije. Vidimo da je rezultat poprilično intuitivan. Međutim, istu tehniku možemo primijeniti kako bismo izveli bilo kakve druge parametre neke distribucije. Zapravo, mi smo to već radili. Prisjetimo se kako smo, na primjer, izveli

kvadratnu pogrešku linearne regresije. Krenuli smo od logaritma vjerojatnosti oznaka $\mathbf{y}$:

$$\begin{aligned}
\ln p(\mathbf{y}|\mathbf{X}) &= \ln \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}) \\
&= \ln \prod_{i=1}^{N} \mathcal{N}\big(h(\mathbf{x}^{(i)};\mathbf{w}), \sigma^2\big) \\
&= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{\big(y^{(i)} - h(\mathbf{x}^{(i)};\mathbf{w})\big)^2}{2\sigma^2}\Big) \\
&= \underbrace{-N \ln(\sqrt{2\pi}\sigma)}_{\text{konst.}} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \big(y^{(i)} - h(\mathbf{x}^{(i)};\mathbf{w})\big)^2 \\
&\propto -\frac{1}{2} \sum_{i=1}^{N} \big(y^{(i)} - h(\mathbf{x}^{(i)};\mathbf{w})\big)^2
\end{aligned}$$

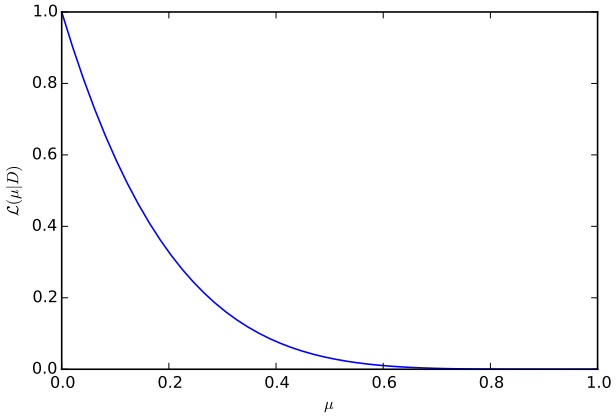
Na osnovu ovoga smo zaključili da, uz pretpostavku Gaussovog šuma u oznakama $\mathbf{y}$, maksimizacija (logaritma) vjerojatnosti oznaka odgovara minimizaciji pogreške kvadratnog gubitka. No, primijetimo sada da logaritam vjerojatnost oznaka, $\ln p(\mathbf{y}|\mathbf{X})$ nije ništa drugo nego **funkcija log-izglednosti težina**, $\ln \mathcal{L}(\mathbf{w}|\mathcal{D})$. To znači da smo naš zaključak mogli formulirati i ovako: uz pretpostavku da je $p(y|\mathbf{x})$ Gaussova gustoća vjerojatnosti, minimizacija kvadratne pogreške jednaka je MLE procjeni za $\mathbf{w}$. Tako vidimo da je učenje modela zapravo isto što i procjena parametara neke pretpostavljene distribucije. 4

2.6 MLE i prenaučenost

MLE je najjednostavniji i najintuitivniji procjenitelj, međutim ima jedan velik problem – sklon je **prenaučenosti**. Pokažimo prenaučenost MLE-a na konkretnom primjeru.

▸ **PRIMJER**

Zamislimo da radimo procjenu parametra μ Bernoullijeve distribucije. Nadalje, zamislimo da se radi o običnom (nemodificiranom) novčiću, za koji očekujemo da je pravedan, tj. vjerojatnost da dobijemo glavu jednaka vjerojatnosti da dobijemo pismo, tj. $\mu = 0.5$. Recimo da bacamo novčić 5 puta. Zamislimo da smo u svih 5 bacanja dobili pismo. Kolika je vjerojatnost glave, tj. kolika je vrijednost parametra μ prema MLE procjenitelju? Da bismo odgovorili na to pitanje, trebamo iskazati funkciju izglednosti, i zatim naći μ koji ju maksimizira. U ovom slučaju, funkcija izglednosti izgleda ovako: 5



$m = 0$, $N = 5$

Vidimo da funkcija izglednosti doseže svoj maksimum (∞) za $\mu = 0$. To znači da je vjerojatnost da dobijemo glavu jednaka je nuli. No, hoćemo li doista vjerovati takvoj procjeni? Možda je novčić ipak pravedan, ali smo samo imali peh. Na kraju krajeva, uzorak je vrlo malen, i, premda je malo vjerojatno, ipak nije nemoguće da dobijemo pet puta pismo (vjerojatnost za to je $0.5^5 = 0.03125 \approx 3\%$).

Potpuno isti problem bismo imali da smo pet puta dobili glavu. Onda bi MLE dao $\mu = 1$.

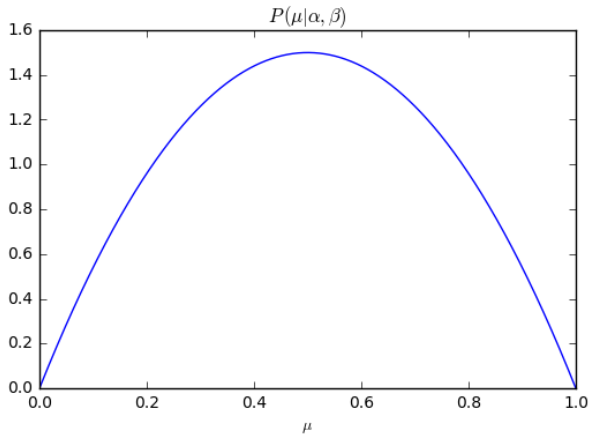
Ovdje je, dakle, problem u tome što, ako u uzorku imamo sve nule odnosno sve jedinice, onda će maksimum funkcije izglednosti biti na $\mu = 0$ odnosno $\mu = 1$.

Uzrok ovog problema jest u tome što se MLE previše oslanja na podatke. To je u redu ako imamo mnogo podataka (ako je uzorak velik). Na primjer, zamislimo da smo 1000 puta bacali novčić. Ako u 1000 bacanja novčića stvarno niti jednom ne dobijemo glavu, onda se čini sasvim razumnim procijeniti da je $\mu = 0$. Međutim, ako nemamo puno podataka, npr., ako imamo samo pet bacanja novčića, onda nije dobro previše se osloniti na takve podatke. Tada se, osim na podatke, trebamo osloniti i na naše znanje (bilo ekspertizu ili zdrav razum). Kada pričamo o procjeniteljima, "znanje" se svodi na **apriornu distribuciju** kojom opisujemo kakve vrijednosti parametara očekujemo. Konkretno, u slučaju novčića, očekujemo da je za normalan (nemodificirani) novčić μ negdje oko 0.5. Ne očekujemo baš $\mu = 0$ ili $\mu = 1$ – kakav bi to novčić bio koji uvijek pada na istu stranu?

Nažalost, MLE nam tu ne može pomoći – ne postoji način da u procjenitelj ugradimo svoje apriorno znanje o parametrima. Za to nam treba neki procjenitelj koji bi bio u stanju kombinirati informacije iz podataka s našim apriornim znanjem. To je **procjenitelj MAP**.

3 Procjenitelj MAP

Procjenitelj **maksimum aposteriori (MAP)** kombinira informacije koje dolaze iz podataka s našim pozadinskim znanjem o mogućim vrijednostima parametra. To naše znanje definiramo kroz **apriornu distribuciju parametra**, koju ćemo označiti sa $p(\boldsymbol{\theta})$ (formalno, to je gustoća vjerojatnosti, jer općenito $\boldsymbol{\theta}$ je kontinuirana varijabla). Apriorna distribucija nam kazuje koje su vrijednosti parametra $\boldsymbol{\theta}$ više, a koje manje vjerojatne. Npr., za novčić je više vjerojatno da je $\mu = 0.5$ nego da $\mu = 0.1$, pa bismo apriornu distribuciju mogli definirati ovako:



Primijetite da smo sada parametar $\boldsymbol{\theta}$ odjednom počeli tretirati kao slučajnu varijablu koja ima svoju distribuciju. To je veliki napredak u odnosu na ono što smo radili kod procjenitelja MLE, gdje $\boldsymbol{\theta}$ nismo eksplicitno tretirali kao slučajnu varijablu (premda on to jest, kad malo razmislite, jer je to procjena na temelju uzorka, međutim kod MLE to potpuno ignoriramo).

Sada nekako trebamo kombinirati izglednost parametra $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$, za koju znamo da je zapravo jednaka vjerojatnosti $p(\mathcal{D}|\boldsymbol{\theta})$, i apriornu distribuciju parametra $p(\boldsymbol{\theta})$. Mogli bismo to napraviti na razne načine, ali umjesto da nešto petljamo, bolje je da se zapitamo što zapravo želimo dobiti. Ako već parametar $\boldsymbol{\theta}$ tretiramo kao slučajnu varijablu, onda ima smisla da pokušamo dobiti distribuciju za tu varijablu. Konkretno, ako je $p(\boldsymbol{\theta})$ apriorna vjerojatnost za tu varijablu, onda je ono što mi želimo dobiti zapravo **aposteriorna vjerojatnost parametra** $p(\boldsymbol{\theta}|\mathcal{D})$. To je vjerojatnost da parametar $\boldsymbol{\theta}$ poprimi neku vrijednost, nakon što nam je predočen skup primjera $\mathcal{D}$. Sada je lako vidjeti da te aposteriorne vjerojatnosti možemo jednostavno doći primjenom uvijek nam dragog **Bayesovog pravila**:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{p(\mathcal{D})}$$

Aposteriorna vjerojatnost parametara (zapravo: aposteriorna gustoća vjerojatnosti parametra) kombinira apriorno znanje i informaciju iz podataka. MAP radi maksimizaciju te vjerojatnosti. Budući da je $\mathcal{D}$ fiksan, to je nazivnik konstanta, pa ga pri maksimizaciji možemo zanemariti. Prema tome, MAP procjenitelj je:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ p(\boldsymbol{\theta}|\mathcal{D}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

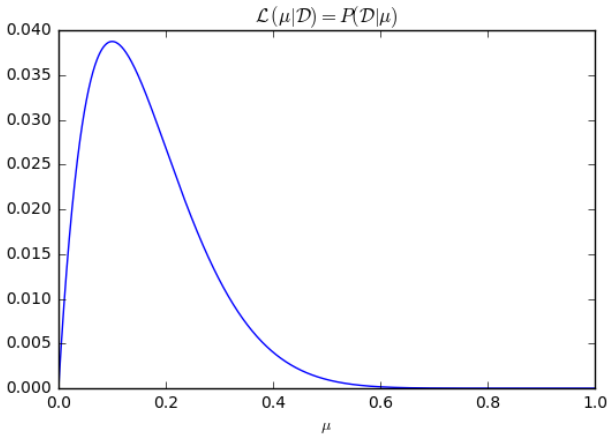
Usporedimo to s procjeniteljem MLE:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

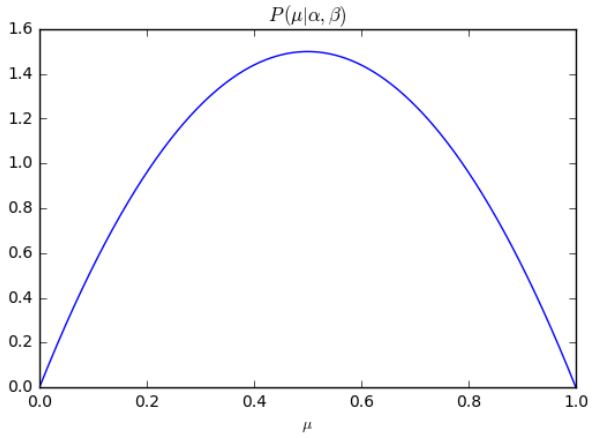
Vidimo da je razlika u tome što MAP zapravo kombinira izglednost parametara $\boldsymbol{\theta}$ (koja je vjerojatnost uzorka $\mathcal{D}$) s apriornom vjerojatnošću parametara $\boldsymbol{\theta}$. Kombinacija se ostvaruje jednostavnim množenjem tih dviju vjerojatnosti. Ako neke vrijednosti za parametar $\boldsymbol{\theta}$ imaju malu apriornu vjerojatnost, onda će i njihova aposterirorna vjerojatnost biti smanjena! Pogledajmo primjer.

▸ **PRIMJER**

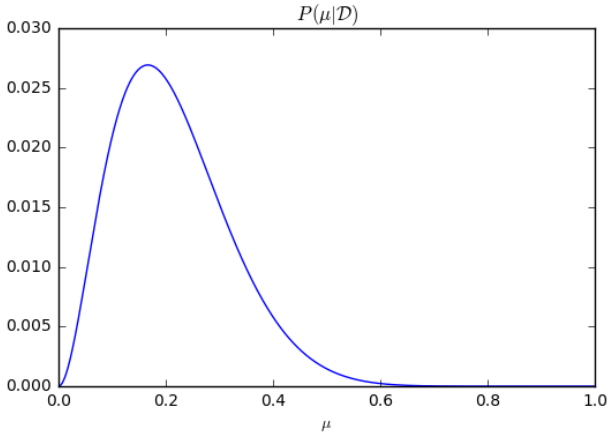
Pogledajmo opet bacanje novčića, dakle procjenu parametra μ Bernoullijeve razdiobe. Recimo da smo od 10 bacanja dobili samo 1 puta glavu. Funkcija izglednosti $\mathcal{L}(\mu|\mathcal{D})$ izgleda ovako:



Najvjerojatnija vrijednost za μ (tj. mod funkcije izglednosti) biti će $\mu = 0.8$. Ali, recimo da mi ipak vjerujemo da je novčić pravedan. To znači da ćemo apriornu distribuciju $p(\mu)$ modelirati ovako:



Mod ove funkcije je $\mu = 0.5$. Kada pomnožimo ove dvije vjerojatnosti, preciznije, kada pomnožimo izglednost $\mathcal{L}(\mu|\mathcal{D})$ i gustoću vjerojatnosti $p(\mu)$, dobivamo aposteriornu gustoću vjerojatnosti parametara, $p(\mu|\mathcal{D})$, koja izgleda ovako:



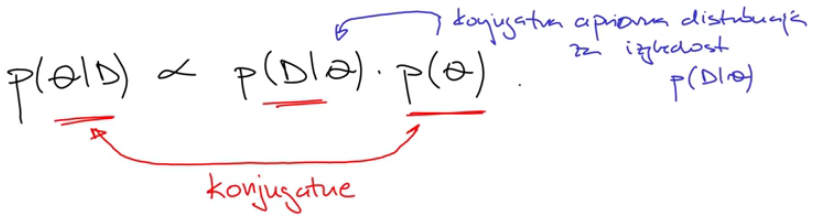
Mod ove funkcije je negdje između 0.1 i 0.2. Dakle, vidimo da se efektivno događa to da naše apriorno znanje pomiče aposteriornu vjerojatnost u smjeru koji smatramo više vjerojatnim. Možemo reći i obrnuto: informacije iz podataka utječu na naše apriorno znanje u smjeru vrijednosti parametara koje su izglednije na temelju podataka. Dakle, to je kao nekakav susret teorije i empirije (teorija je naše apriorno znanje, a empirija su konkretni podatci koje imamo).

Toliko o ideji. Pogledajmo sada kako to funkcionira matematički. Mi bismo željeli da se maksimizacija aposteriorne vjerojatnosti $p(\boldsymbol{\theta}|\mathcal{D})$ može provesti analitički, tj. da makimizator možemo izraziti u zatvorenoj formi. Međutim, problem je što umnožak $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ općenito može poprimiti različite oblike koje nije moguće analitički maksimizirati. Srećom, ovaj problem možemo vrlo pragmatično riješiti. Naime, ništa nas ne sprječava da pokušamo odabrati takve distribucije $p(\mathcal{D}|\boldsymbol{\theta})$ i $p(\boldsymbol{\theta})$ da njihov umnožak daje neku nama poznatu teorijsku distribuciju, s kojom znamo raditi. Istini za volju, distribuciju $p(\mathcal{D}|\boldsymbol{\theta})$ nemamo što birati, jer je ona već definirana vrstom podataka s kojima radimo: to će biti izglednost Bernoullijeve varijable, ili multinulijeve, ili Gaussove. Prema tome, ono što nam preostaje jest da pametno odaberemo distribuciju $p(\boldsymbol{\theta})$, tako da umnožak $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ bude neka poznata distribucija.

Štoviše, kad već biramo, bilo bi jako dobro da odaberemo tako da aposteriorna distribucija $p(\boldsymbol{\theta}|\mathcal{D})$ bude ista vrsta distribucije kao i apriorna distribucija $p(\boldsymbol{\theta})$. Zašto? Zato što bi nam to onda omogućilo da radimo **"online" (pojedinačno) učenje**: da učimo postepeno kako dolaze

novi podatci. Naime, ako je aposteriorna distribucija istoga tipa kao apriorna distribucija, onda, kada izračunamo aposteriornu distribuciju, možemo je u idućoj iteraciji, kada dođu novi podatci, koristiti kao novu apriornu distribuciju. I taj postupak onda možemo ponavljati u svakoj iteraciji "online" učenja: novo znanje koje smo upravo naučili (aposteriorna distribucija) već se u idućoj iteraciji koristi kao staro znanje (apriorna distribucija).

Ako su $p(\boldsymbol{\theta}|\mathcal{D})$ i $p(\boldsymbol{\theta})$ odabrane tako da su to iste vrste distribucija, onda ih nazivamo **konjugatnim distribucijama**. Kako odabrati apriornu distribuciju $p(\boldsymbol{\theta})$, a da $p(\boldsymbol{\theta}|\mathcal{D})$ i $p(\boldsymbol{\theta})$ budu konjugatne? Pokazuje se da, barem za standardne funkcije izglednosti, to i nije neki problem: naime, za svaku izglednost $p(\mathcal{D}|\boldsymbol{\theta})$ koja je iz **eksponencijalne familije**, postoji apriorna distribucija $p(\boldsymbol{\theta})$ takva da su apriorna i aposteriorna distribucija konjugatne. Takva apriorna distribucija naziva se **konjugatna apriorna distribucija** za funkciju izglednosti. Sljedeća skica pokazuje te odnose između distribucija: 6



Za nas to onda znači sljedeće: ako za našu funkciju izglednosti odaberemo njoj konjugatnu apriornu distribuciju, onda će aposteriorna i apriorna distribucija biti konjugatne, tj. aposteriorna će distribucija biti ista teorijska razdioba kao i apriorna distribucija. To će nam omogućiti da analitički pronađemo mod (maksimizator) aposteriorne distribucije.

Sada još ostaje da odgovorimo na pitanje koje su apriorne distribucije konjugatne za izglednosti s kojima u strojnom učenju često baratamo. Konjugatna apriorna distribucija za funkciju izglednosti Bernoullijeve varijable je **Beta distribucija**. Za funkciju izglednosti multinulijeve (kategoričke) distribucije to je **Dirichelova distribucija**. Za funkciju izglednosti Gaussove (normalne) varijable to je opet **Gaussova (normalna) distribucija**. Prikažimo to tablično:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

Aposteriorna $p(\boldsymbol{\theta}|\mathcal{D})$	Izglednost $p(\mathcal{D}|\boldsymbol{\theta})$	Apriorna $p(\boldsymbol{\theta})$
Beta	Bernoulli	Beta
Dirichlet	Multinuli	Dirichlet
Normal	Normal	Normal
Multivariate normal	Multivariate normal	Multivariate normal

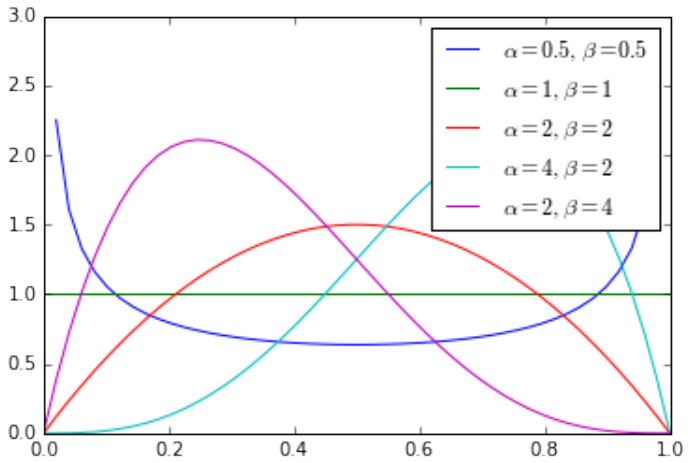
U nastavku ćemo konkretnije pogledati prva dva slučaja, odnosno kako konkretno izračunati MAP procjenitelj za Bernoullijevu varijablu i za kategoričku (multinulijevu) varijablu.

4 Beta-Bernoullijev model

Pogledajmo prvo najjednostavniji slučaj: kada je varijabla Bernoullijeva, a apriorna vjerojatnost parametra μ je beta-distribucija. To je tzv. **Beta-Bernoullijev model**, jer se za modeliranje apriorne distribucije $p(\mu)$ koristi Beta-distribucija, koja je konjugatna Bernoullijevoj izglednosti. Pogledajmo kako je definirana beta-distribucija, točnije funkcija gustoće beta-distribucije:

$$p(\mu|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

Ovdje je B tzv. **beta-funkcija** – normalizirajuća konstanta koja osigurava da je integral funkcije gustoće jednak 1. Parametri distribucije su α i β, za koje mora vrijediti $\alpha, \beta > 0$. Ti parametri određuju oblik distribucije. To prikazuje sljedeći grafikon:



Najčešće želimo postaviti $\alpha > 1$ i $\beta > 1$, kako bismo modelirali veću apriornu gustoću vjerojatnosti za neku vrijednost $\boldsymbol{\theta}$. Što su α i β veći, to je više gustoće vjerojatnosti smješteno oko 0.5. Ako je $\alpha < \beta$, onda je maksimizator manji od 0.5, inače je veći od 0.5. Ako $\alpha = \beta = 1$, onda dobivamo **uniformnu apriornu distribuciju** (engl. *uniform prior*), kojim efektivno modeliramo da nemamo nikakvoga apriornog znanja o vrijednosti parametra, pa takvu apriornu razdiobu zovemo **neinformativna apriorna distribucija** (engl. *uninformative prior*). 7

Važno je da ovdje primijetimo da je μ parametar koji želimo procijeniti, dok su α i β parametri koje trebamo definirati unaprijed i koji onda definiraju izgled gustoće $p(\mu)$. Drugim riječima, budući da parametre α i β ne procjenjujemo već ih moramo definirati unaprijed, iz perspektive strojnog učenja to su zapravo **hiperparametri**. 8

Vidimo da je beta-distribucija vrlo prikladna jer njome možemo lako modelirati različita apriorna vjerovanja o vrijednosti parametra μ. Drugo lijepo svojstvo beta-distribucije je to što se njezin maksimizator (mod) može jednostavno analitički izraziti na temelju parametara α i β:

$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

pod uvjetom da $\alpha > 1$ i $\beta > 1$.

Pogledajmo sada kako množenjem izglednosti Bernoullijeve varijable i apriorne beta-distribucije dobivamo **aposteriornu beta-distribuciju**. Neka je naš uzorak veličine N, i neka se u tom uzorku Bernoullijeva varijabla realizirala kao 1 u ukupno m puta (npr., bacamo novčić N puta, i od toga smo m puta dobili glavu). Već znamo da je funkcija izglednosti Bernoullijeve varijable sljedeća:

$$p(\mathcal{D}|\mu) = \mu^m (1-\mu)^{N-m}$$

Apriorna vjerojatnost definirat ćemo beta-distribucijom:

$$p(\mu|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

Prema Bayesovom pravilu, aposteriorna vjerojatnost je:

$$p(\mu|\mathcal{D}, \alpha, \beta) = \mu^m (1-\mu)^{N-m} \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \frac{1}{p(\mathcal{D})}$$

Nakon što malo presložimo, dobivamo:

$$p(\mu|\mathcal{D}, \alpha, \beta) = \mu^{m+\alpha-1} (1-\mu)^{N-m+\beta-1} \frac{1}{B(\alpha, \beta) p(\mathcal{D})}$$

Znamo da ovo što smo dobili mora biti funkcija gustoće vjerojatnosti, koja se integrira u 1. Budući da izraz na desnoj strani oblikom liči opet na beta-distribuciju, ideja je da izraz pokušamo napisati kao beta-distibuciju. Doista, izraz možemo napisati kao beta-distribuciju sa sljedećim parametrima:

$$p(\mu|\mathcal{D}, \alpha', \beta') = \mu^{\overbrace{m+\alpha}^{\alpha'}-1}(1-\mu)^{\overbrace{N-m+\beta}^{\beta'}-1}\underbrace{\frac{1}{B(\alpha,\beta)p(\mathcal{D})}}_{B(\alpha',\beta')}$$

Vidimo, dakle, da smo dobili novu beta-distribuciju. Parametri te distribucije su α' i β', koje računamo na temelju parametara α i β iz apriorne distribucije, te parametara N i m funkcije izglednosti. Konkretno:

$$\alpha' = m + \alpha$$
$$\beta' = N - m + \beta$$

Naravno, ovdje nas ne treba čuditi da smo kao aposteriornu distribuciju dobili beta-distribuciju. Naime, to smo i očekivali, budući da smo već rekli da je beta-distribucija konjugatna distribucija za Bernoullijevu izglednost, pa smo zato dobili istu vrstu distribucije kao što je i apriorna distribucija.

Vratimo se na MAP procjenitelj. Prisjetimo se:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ p(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

Dakle, zanima nas **maksimizator aposteriorne distribucije**. Budući da se radi o beta-distribuciji, zna se da je njezin maksimizator jednak:

$$\hat{\mu}_{\text{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{m + \alpha - 1}{m + \alpha + N - m + \beta - 2} = \frac{m + \alpha - 1}{\alpha + N + \beta - 2}$$

Ovisno o odabiru vrijednosti hiperparametara α i β, i naravno ovisno o vrijednostima m i N koje dobivamo iz podataka, dobivat ćemo različite vrijednosti procjenitelja. Primijetite, očekivano da, MAP procjenitelj degenerira na procjenitelj MLE $\hat{\mu}_{\text{MLE}} = m/N$ za $\alpha = \beta = 1$, tj. kada je apriorna distribucija uniformna. Ako nemamo apriornog znanja, MAP nije ništa bolji od MLE.

▸ **PRIMJER**

Recimo da smo novčić bacali 10 puta, a da smo samo 1 dobili glavu. Pretpostavljamo da je novčić ipak pravedan, što smo odlučili modelirati sa $\alpha = \beta = 2$. Onda za MAP procjenitelj dobivamo:

$$\hat{\mu}_{\text{MAP}} = \frac{1 + 2 - 1}{2 + 10 + 2 - 2} = \frac{2}{12} = 0.167$$

Što bismo dobili procjeniteljem MLE? Dobili bismo $\mu = 1/10 = 0.1$. Vidimo dakle da smo s MAP procjeniteljem dobili procjenu koja kombinira naše apriorno znanje ($\mu = 0.5$ je najvjerojatnije) i podatke (dobili smo samo 1 glavu u 10 bacanja).

Fantastična stvar sa MAP procjeniteljem je da "automatski" balansira između podataka (empirije) i apriornog znanja (teorije). To se lijepo vidi kod MAP procjenitelja za Bernoullijevu distribuciju. Naime, iz razlomka MAP procjenitelja vidi se da, što je podataka više (N je veći), to više vjerujemo podatcima. Suprotno, što je podataka manje (N je manji), to više vjerujemo apriornom znanju. Drugim riječima, ako je skup za učenje velik, onda N dominira, inače dominiraju α i β.

U strojnom učenju najčešće se koristi MAP procjenitelj s apriornom razdiobom $\alpha = \beta = 2$. To konkretno daje:

$$\hat{\mu}_{\text{MAP}} = \frac{m+1}{N+2}$$

Ovaj se procjenitelj još naziva i **Laplaceov procjenitelj**. Primijetite da ovaj procjenitelj efikasno rješava problem **prenaučenosti** do koje dolazi kada je $m = 0$ ili $m = N$. Naime, ono što se događa je da, premda je $m = 0$, ipak nećemo reći da je $\mu = 0$, nego će μ biti malo veći (analogno za $m = N$). To je kao da smo ukupnu masu vjerojatnosti malo preraspodijelili, tako da vjerojatnost pozitivnog ishoda Bernoullijeve varijable ipak nije nula, nego ima neku malo vrijednost. Taj efekt raspodjele dijela mase vjerojatnosti s vrlo vjerojatnih događaja na manje vjerojatne događaje općenito se u strojnom učenju naziva **zaglađivanje** (engl. *smoothing*). MAP procjenitelj je jedan (principijelan) način da **zagladimo procjene parametara** i tako smanjimo mogućnost prenaučenosti probabilističkih modela.

Sada znamo kako zaglađivati, odnosno kako izračunati MAP procjenitelj kada je varijabla binarna. No, što ako je varijabla multinomijalna, tj. kategorijska? To nas vodi do Dirichlet-kategoričkog modela.

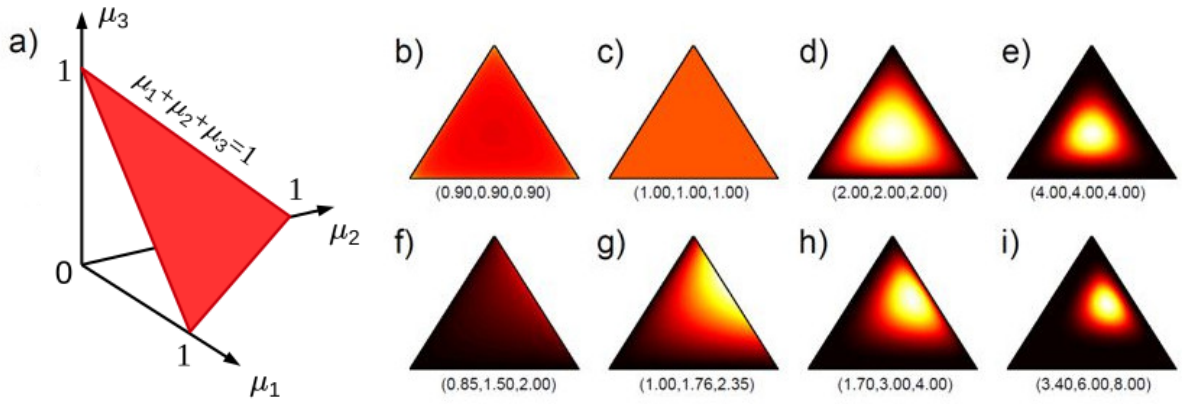
5 Dirichlet-kategorički model

Nećemo to izvoditi, ali slično kao što smo izveli MAP procjenitelj za Bernullijevu distribuciju može se izvesti procjenitelj za **multinulijevu (kategoričku) distribuciju**, koristeći **Dirichletovu distribuciju** kao konjugatnu apriornu distribuciju multinulijevoj izglednosti. To daje tzv. **Dirichlet-kategorički model**.

Dirichletova distribucija definirana je ovako:

$$P(\boldsymbol{\mu}|\boldsymbol{\alpha}) = P(\mu_1, \ldots, \mu_K|\alpha_1, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

Dirichletova distribucija poopćenje je beta distribucije na više μ_k. Hiperparametri $\alpha_1, \ldots, \alpha_K$ određuju vjerojatnosti parametara $\mu_1, \ldots, \mu_K$. Međutim, ovdje lako dolazi do zabune: nije slučaj da pojedinačni α_k direktno određuje vjerojatnost pojedinačnog parametra μ_k, nego je ta veza indirektna, na način da cijeli vektor $\boldsymbol{\alpha}$ određuje vjerojatnosti parametara $\boldsymbol{\mu}$. Nadalje, budući da mora vrijediti $\sum_{k=1}^{K} \mu_k = 1$, to znači da se parametri μ_k nalaze na tzv. $(K-1)$-dimezijskom **standardnom simpleksu**. Npr. , za $K = 3$, to je trokut u trodimenzijskom prostoru:



Svaka točka na ovom trokutu odgovara jednoj kombinaciji (μ_1, μ_2, μ_3) te vrijedi $\mu_1 + \mu_2 + \mu_3 = 1$. Nadalje, svakoj takvoj točki pridružena je određena vrijednost funkcije gustoće vjerojatnosti, a integral po gustoći vjerojatnosti po cijelom trokutu je jednak 1. Vektor $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$

određuje kako je vjerojatnost distribuirana između (μ_1, μ_2, μ_3). Primjeri za konkretne vrijednosti vektora $\boldsymbol{\alpha}$ prikazani su na slikama b–f ispod trokuta. Ako $\alpha_1 = \alpha_2 = \alpha_3 = 1$, onda je svaka kombinacija $\mu_1, \mu_2, \mu_3)$ jednako vjerojatna (slika c). Ako $\alpha_1 = \alpha_2 = \alpha_3 > 1$, onda su vjerojatnije kombinacije koje daju jednaku vjerojatnosti za μ_1, μ_2 i μ_3, tj. one koje su u sredini trokuta (slike c–e). Asimetrične vrijednosti za α_1, α_2 i α_3 davat će veću masu vjerojatnosti kombinacijama kod kojih μ_k nisu jednaki (npr. slike f–i).

Lako bismo mogli izvesti Dirichlet-kategorički model, ali nećemo. Kada bismo to napravili, za procjenitelj za μ_k dobili bismo opet da je jednak modu aposteriorne (Dirichletove) distribucije:

$$\hat{\mu}_{k,\text{MAP}} = \frac{\alpha'_k - 1}{\sum_{k=1}^{K} \alpha'_k - K}$$

gdje

$$\alpha'_k = N_k + \alpha_k$$

gdje je N_k broj nastupanja k-te vrijednosti.

Ako želimo modelirati da je najvjerojatnije da su sve pojedinačne vrijednosti kategorijske varijable jednako vjerojatne, tj. $\mu_k = 1/K$, onda možemo staviti $\alpha_k = 2$. Za MAP procjenitelj ćemo onda dobiti:

$$\hat{\mu}_{k,\text{MAP}} = \frac{N_k + 2 - 1}{\sum_k (N_k + 2) - K} = \frac{N_k + 1}{N + K}$$

gdje je $N = \sum_k N_k$ ukupan broj primjera.

▸ **PRIMJER**

Npr., ako kategorijska varijabla ima $K = 3$ moguće vrijednosti u skupu od $N = 10$ primjera, ali prva vrijednost se nikada nije pojavila, $N_1 = 0$, MLE procjena za tu vrijednost bi bila $\hat{\mu}_1 = 0$, dok je MAP procjena $\hat{\mu}_1 = \frac{0+1}{10+3} = 0.077$.

Sažetak

- **Procjenitelj najveće izglednosti (MLE)** odabire parametre koji maksimiziraju vjerojatnost realizacije uzorka (tj. izglednost)
- Kod poopćenih linearnih modela, MLE je istovjetan minimizaciji empirijske pogreške
- MLE je jednostavan, ali lako daje **prenaučene modele**
- **MAP-procjenitelj** dodatno koristi apriornu razdiobu parametara i maksimizira aposteriornu vjerojatnost parametara, efektivno provodeći **zaglađivanje**
- MAP-procjenitelj za binarnu varijablu koristi **Beta-Bernoullijev model**
- MAP-procjenitelj za kategorijsku varijablu koristi **Dirichlet-kategorički model**

Bilješke

[1] Već smo to bili napomenuli, ali napomenimo za svaki slučaj ponovo, da je $\mathbf{x}$ u izrazu $p(\mathbf{x}|\boldsymbol{\theta})$ slučajna varijabla, dok $\boldsymbol{\theta}$ to nije, već je to samo parametar distribucije. Bilo bi, stoga, korektnije pisati $p(\mathbf{x}; \boldsymbol{\theta})$, jer ako pišemo $p(\mathbf{x}|\boldsymbol{\theta})$ to izgleda kao uvjetna vjerojatnost gdje su i $\mathbf{x}$ i $\boldsymbol{\theta}$ varijable. Međutim, u literaturi se uvriježilo pisati $p(\mathbf{x}|\boldsymbol{\theta})$, pa ćemo i mi ostati na tome. Imajte, međutim, na umu da je

$\boldsymbol{\theta}$ parametar, a ne slučajna varijabla. Ova napomena ne vrijedi za bayesovsku statistiku, kod koje su i $\mathbf{x}$ i $\boldsymbol{\theta}$ slučajne varijable. Više o tome pri kraju današnjeg predavanja.

2 Jasnoće radi, u nastavku ćemo često koristiti naziv **procjentielj MLE**, premda je to pleonazam, budući da "E" već stoji za "procjenitelj".

3 Primjenimo **metodu Lagrangeovih multiplikatora** na maksimizaciju funkcije log-izglednosti kategoričke (multinulijeve) varijable. Želimo maksimizirati sljedeću funkciju log-izglednosti:

$$\ln \mathcal{L}(\boldsymbol{\mu}|\mathcal{D}) = \ln \prod_{i=1}^{N} P(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = \ln \prod_{i=1}^{N} \prod_{k=1}^{K} \mu_k^{x_k^{(i)}} = \sum_{k=1}^{K} \sum_{i=1}^{N} x_k^{(i)} \ln \mu_k$$

uz ograničenje $\sum_{k=1}^{K} \mu_k = 1$. Prisjetite se (iz skriptice 8) kako se definira Lagrangeova funkcija. U ovom slučaju, Lagrangeova je funkcija sljedeća:

$$\sum_{i=1}^{K} \sum_{i=1}^{N} x_k^{(i)} \ln \mu_k + \lambda \Big(\sum_{k=1}^{K} \mu_k - 1 \Big)$$

gdje je λ Lagrangeov multiplikator. Deriviranjem po μ_k i izjednačavanjem s nulom dobivamo:

$$\mu_k = -\frac{1}{\lambda} \sum_{i=1}^{N} x_k^{(i)}$$

Kako bismo izračunali vrijednost multiplikatora λ, dobiveni izraz uvrštavamo u ograničenje $\sum_k \mu_k = 1$ i tako dobivamo:

$$\sum_{k=1}^{K} \mu_k = -\frac{1}{\lambda} \underbrace{\sum_{k=1}^{K} \sum_{i=1}^{N} x_k^{(i)}}_{=N} = 1$$

Budući da svaka multinomijalna varijabla $\mathbf{x}$ ima jedinicu postavljenu na samo jednoj komponenti, zbroj komponenata svih varijabli jednak je broju primjera N. Vrijedi dakle $\lambda = -N$, pa uvrštavanjem u gornju jednadžbu dobivamo:

$$\hat{\mu}_{k,\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_k^{(i)} = \frac{N_k}{N}$$

4 Preciznije, $\ln p(\mathbf{y}|\mathbf{X})$ nije logaritam vjerojatnosti oznaka, nego logaritam funkcije gustoće vjerojatnosti. Osim toga, ta je funkcija implicitno i funkcija težina $\mathbf{w}$, dakle trebali bismo pisati $\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})$. Iz toga je onda jasnije da je odgovarajuća funkcija log-izglednosti $\ln \mathcal{L}(\mathbf{w}|\mathbf{X}, \mathbf{y})$. Par $(\mathbf{X}, \mathbf{y})$ čini skup označenih primjera $\mathcal{D}$, pa je, dakle, funkcija log-izglednosti $\ln \mathcal{L}(\mathbf{w}|\mathcal{D})$, tj. to je funkcija težina $\mathbf{w}$ uz fiksirani skup primjera $\mathcal{D}$.

5 Za pravedan novčić očekujemo $\mu = 0.5$. Međutim, Jaynes smatra: "... anyone familiar with the law of conservation of angular momentum can, after some practice, cheat at the usual coin-toss game and call his shots with 100 per cent accuracy. You can obtain any frequency of heads you want; and the bias of the coin has no influence at all on the results!" (Jaynes, 2003). Ipak, nisam uvjeren da se *znanje* o očuvanje kutne količine gibanja doista, pa čak i nakon "nešto vježbe", može pretočiti u *vještinu* bacanja novčića tako da se uvijek dobije željeni ishod, jednako kao što nisam uvjeren da bi netko znanje o kinematici krutog tijela mogao iskoristiti da uvijek obrani jedanaesterac, koliko god marljivo vježbao. Ali, u duhu epistemičke skromnosti, ostajem otvoren za takvu mogućnost (ovo se ne računa: https://www.youtube.com/watch?v=MZc3VD2140U).

6 Ovdje postoji mala terminološka zbrka, jer se pridjev **konjugatan** koristi na dva različita načina: za opis toga da su apriorna i aposteriodna distribucije iste vrste i za opis toga da je apriorna distribucija distribucija takva da, kada se pomnoži s izglednošću, daje distribuciju koja je iste vrste kao i aposteriorna distribucija. Upamtite: ako je $p(\boldsymbol{\theta})$ **konjugatna distribucija za izglednost** $p(\mathcal{D}|\boldsymbol{\theta})$, onda su $p(\boldsymbol{\theta})$ i $p(\boldsymbol{\theta}|\mathcal{D})$ **konugatne distribucije**, i obrnuto.

7 Također, za $\alpha = \beta = 1$ to je ujedno i tzv. **nepravilna apriorna distribucija** (engl. *improper prior*), jer to onda nije prava funkcija gustoće vjerojatnosti budući da je funkcija veća od nule na intervalu od

minus do plus beskonačno, pa njezin integral nije jednak. Međutim, u praksi nas to ne smeta, sve dok je rezultirajuća aposteriorna funkcija dobro definirana distribucija (a bit će, jer ju normaliziramo u nazivniku Bayesovog pravila).

[8] Također, u tom smislu možemo reći da je gustoća $p(\mu)$ zapravo **distribucija distribucije** ili **distribucija drugog reda**. Zašto? Zašto što za svaku vrijednost od μ imamo jednu Bernoullijevu distribuciju, pa onda distribucija $p(\mu)$ opisuje kako su distribuirane te Bernoullijeve distribucije. Npr., za $\alpha = \beta = 2$, najverojatnije Bernoullijeve distribucije su one za koje je $\mu = 0.5$.

Literatura

E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.