Bayesi talumatu kergus

JÜRI LEMBER
Tartu Ülikool

Bernoulli protsess. Olgu $Y_1, Y_2, \ldots$ sõltumatud Bernoulli jaotusega juhuslikud suurused, jaotuse parameeter olgu $p \in (0,1)$. Sellist juhuslike suuruste jada nimetatakse teinekord **Bernoulli protsessiks**. Bernoulli protses tähendab järgmist: iga juhuslik suurus Y_i võib võtta väärtusi 0 või 1 (kull/kiri, kass/koer, emane/isane), kusjuures väärtuse 1 tõenäosus on p ja väärtuse 0 tõenäosus on $1-p$. Matemaatiliselt seega $P(Y_i = 1) = p$, $P(Y_i = 0) = 1-p$, ehk $Y_i \sim B(1,p)$ iga $i = 1,2,\ldots$ korral. Seega on kõik juhuslikud suurused **sama (Bernoulli) jaotusega**. Teine oluline eeldus on **sõltumatus**. Intuitiivselt tähendab sõltumatus seda, et kui juhuslike suuruste $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$ väärtused on teada, siis see ei mõjuta kuidagi juhusliku suuruse Y_i jaotust (ehk Y_i tõenäosusi võtta väärtusi 0 või 1) ja seda iga i ja n korral. Matemaatiliselt defineeritakse Bernoulli jaotusega juhuslike suuruste sõltumatus järgmiselt: iga $n \geq 2$ ja $i_1, \ldots, i_n \in \{0,1\}$ korral

$$P(Y_1 = i_1, Y_2 = i_2, \ldots, Y_n = i_n) = P(Y_1 = i_1) \cdots P(Y_n = i_n). \tag{1}$$

Näiteks, kui $i_1 = 1, i_2 = 0, i_3 = 1, i_4 = 1$, siis definitsioonist (1) järeldub

$$P(Y_1 = 1, Y_2 = 0, Y_3 = 1, Y_4 = 1) =$$
$$= P(Y_1 = 1)P(Y_2 = 0)P(Y_3 = 1)P(Y_4 = 1)\,.$$

Samuti esitub sõltumatuse korral ühisjaotus korrutisena ka siis, kui juhulikud suurused pole järjestatud, näiteks

$$P(Y_5 = 1, Y_7 = 0, Y_{13} = 0) = P(Y_5 = 1)P(Y_7 = 0)P(Y_{13} = 0).$$

Seega definitsiooni järgi tähendab juhuslike suuruste jada sõltumatus lõpmatu paljude võrduste samaaegset kehtimist; piisab sellest, et üks võrdus ei kehti ja juhuslikud suurused pole enam sõltumatud. Siit näeme, et sõltumatus, kuigi tihti üsna loomulik (nt mündivisked), on

tegelikult väga tugev eeldus. Sellele vaatamata on sõltumatud mudelid stohhastikas enamlevinud ning statistikas selgelt üleekspluateeritud, enamus üldkasutatavatest statistikameetodidest põhineb sõltumatusel ja praktikas eeldatakse seda pahatihti automaatselt. Osaliselt on see igati arusaadav, sest sõltumatus on (nii matemaatiliselt kui intuitiivselt) lihtne ja üheselt mõistetav ning sõltumatute ja sama jaotusega juhuslike suuruste jadal on palju häid ning hästituntud omadusi. Nii teame, et Bernoulli protsessi korral kehtib nn **suurte arvude seadus**, mille kohaselt n juhusliku suuruse aritmeetiline keskmine koondub n kasvades (peaaegu kindlasti) arvuks p (juhusliku suuruse Y_i keskväärtus):

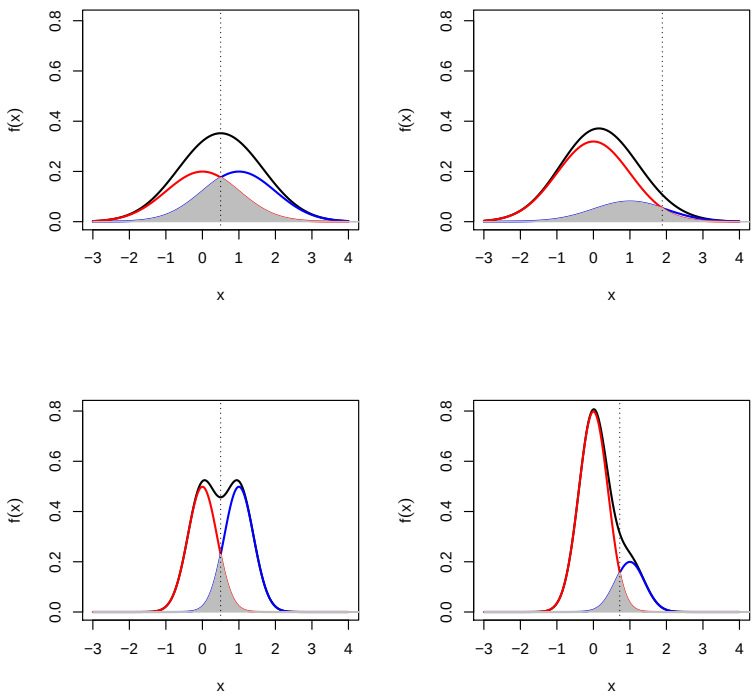
$$\frac{S_n}{n} \to p, \quad \text{p.k, kus } S_n = \sum_{i=1}^{n} Y_i.$$

See tähendab, et suure n korral on $Y_1, \dots, Y_n$ seas ligikaudu pn ühte. Arusaadavalt on ühtede arv S_n juhuslik suurus ja meie lihtsa mudeli korral on selle juhusliku suuruse jaotus täpselt teada: $S_n \sim B(n,p)$ (binoomjaotus parameetritega n ja p). Loobudes sõltumatusest, peaksime täpselt defineerima sõltuvusmudeli (juhusliku protsessi) ja selleks on lõpmata palju võimalusi, igaühe valik neist on samuti tugev eeldus. Samuti ei pruugi enam kehtida ülaltoodud omadused.

Segumudel. Niisiis, meil on Bernoulli protsess $Y_1, Y_2, \dots$, kusjuures $P(Y_i = 1) = p$ iga i korral. Ilmselt maailma kõige lihtsam ja vanem stohhastiline mudel. Kujutagem nüüd, et nende juhuslike suuruste väärtuste lugemine toimub vigadega, st kui $Y_i = 0$ (või $Y_i = 1$), siis vaatlusena registreeritakse normaaljaotusega juhusliku suuruse väärtus, kusjuures selle normaaljaotuse keskväärtus on 0 (või 1). Olgu nende normaaljaotuste dispersioon alati σ^2 ning olgu vead samuti sõltumatud. Edaspidi tähistame vaatlusi $X_1, X_2, \dots$ ja nende jaotus on järgmine: kui $Y_i = 0$, siis on $X_i \sim \mathcal{N}(0, \sigma^2)$, kui aga $Y_i = 0$, siis on $X_i \sim \mathcal{N}(1, \sigma^2)$. Seega tõenäosus, et juhuslik suurus X_i võtab väärtusi hulgas B avaldub täistõenäosuse valemi kohaselt

$$\begin{aligned} P(X_i \in B) &= P(X_i \in B | Y_i = 1) P(Y_i = 1) + P(X_i \in B | Y_i = 0) P(Y_i = 0) \\ &= P(Z + 1 \in B) p + P(Z \in B)(1 - p), \end{aligned}$$

kus $Z \sim \mathcal{N}(0,1)$. Siit järeldub, et X_i tihedus on Gaussi kõverate kaalutud keskmine: $f(x) = pf_{1,\sigma^2}(x) + (1-p)f_{0,\sigma^2}(x), \quad x \in \mathbb{R}$. Siin f_{1,σ^2} (f_{0,σ^2}) on keskväärtusega 1 (keskväärtusega 0) ja dispersiooniga σ^2 normaaljaotuse tihedus. Kui kaalud on võrdsed ($p = 0,5$) ja σ^2 piisavalt väike, on f ilus kahe küüruga kaamel, ebavõrdsete kaalude korral on üks küür domineeriv ja väga ebavõrdsete kaalude korral ei pruugi teist küüru nähagi olla. Sellist jaotust nimetatakse **normaaljaotuste seguks**. Järgmisel joonisel on toodud segujaotuse tihedus erinevate p ja σ korral.



Joonis 1: Segujaotuse (must) ja kaalutud komponentide (sinine, punane) tihedus. Punktiirjoon: klassifitseerimisreegel $t(p,\sigma)$. Ülemine vasak: $\sigma = 1$ ja $p = 0,5$; ülemine parem: $\sigma = 1$ ja $p = 0,2$; alumine vasak: $\sigma = 0,4$ ja $p = 0,5$; alumine parem: $\sigma = 0,4$ ja $p = 0,2$. Hall ala: klassifitseerimisvea tõenäosus. Väheneb koos parameetriga σ.

Meie eelduse kohaselt on vead sõltumatud ja see tähendab, et juhuslikud suurused (vaatlused) $X_1, X_2, \ldots$ on ka sõltumatud. Pidevate juhuslike suuruste korral on (1) trivaalselt kehtiv ja sõltumatus defineeritakse järgmiselt: iga $n \geq 2$ ja iga (mõõtuvate) hulkade komplekti $B_1, \ldots, B_n$ korral

$$P(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n). \qquad (2)$$

Seega meil on ühest küljest väga lihtne mudel – sõltumatud ja sama jaotusega (segujaotusega) juhuslikud suurused, teisest küljest on meil aga **varjatud (latentse) tunnusega mudel**. Sellistes mudelites on lisaks mõõdetavatale tunnusel X varjatud tunnus Y, mida otseselt mõõta ei saa. Meie mudelis on $Y \sim B(1,p)$ varjatud tunnus ja mõõdetav tunnus X avaldub kujul $X = Y + Z$, kus $Z \sim \mathcal{N}(0,\sigma^2)$ on sõltumatu viga. Arusaadavalt võib varjatud tunnusena vaadelda ka mõõtmisviga Z. Nii või teisiti, just sellisest varjatud tunnusega mudelist pärinevadki meie (sõltumatud) vaatlused $X_1, X_2, \ldots$. See tähendab ka seda, et n esimest juhuslikku suurust $X_1, \ldots, X_n$ võib genereerida järgmiselt: kõigepealt genereeritakse Bernoulli protsessist n juhuslikku suurust $Y_1, \ldots, Y_n$ ja siis n sõltumatut juhuslikku suurust (mõõtmisviga) $Z_1, \ldots, Z_n$, $Z_i \sim \mathcal{N}(0,\sigma^2)$. Lõpuks need liidatakse: $X_i = Y_i + Z_i$, $i = 1, \ldots, n$.

Segmenteerimine. Olgu $x_1, \ldots, x_n$ ülalkirjeldatud segumudelist pärinevad vaatlused ehk juhuslike suuruste $X_1, \ldots, X_n$ realisatsioonid. Tuletame meelde, et arve $x_1, \ldots, x_n$ võib käsitleda juhuslike suuruste $Y_1, \ldots, Y_n$ mingite realisatsioonide - olgu need $y_1, \ldots, y_n$ - vigadega mõõtmised. (Matemaatikas tähistatakse suurte tähtedega juhuslikke suurusi ja väikeste tähtedega nende väärtusi ehk realisatsioone ja seetõttu kasutame nüüd väikesi tähti). Olgu nüüd meie eesmärk varjatud tunnuste (mündivisete) y_i teadasaamine või äraarvamine. See ülesanne on tuntud kui **klassifitseerimis-** või **segmenteerimisprobleem** ja see probleem on nn tehisõppe (*statistical learning*) prototüüpprobleem. Sellest terminoloogiast lähtuvalt nimetamegi arve y_i edaspidi **klassideks**. Olgu ette öeldud, et üldjuhul on iga y_i täpne teadasaamine võimatu, küll aga on võimalikud mingid hinnangud $\hat{y}_i$. See tähendab, et iga i korral valime $\hat{y}_i \in \{0,1\}$ nii, et vektor $(\hat{y}_1, \ldots, \hat{y}_n)$ oleks mingis mõttes parim.

Segumudeli korral on mõistlik valida $\hat{y}_i$ järgmiselt:

$$\hat{y}_i = \arg\max_{y\in\{0,1\}} P(Y_i = y | X_i = x_i) = \begin{cases} 1, & \text{kui } p f_{1,\sigma^2}(x_i) > (1-p) f_{0,\sigma^2}; \\ 0, & \text{mujal.} \end{cases} \tag{3}$$

Seega iga i korral on $\hat{y}_i$ suurima tõepäraga klass. On lihtne näha, et $\hat{y}_i = 1$ parajasti siis, kui $x_i > t$, kus $t(p,\sigma)$ (sõltub p-st ja σ-st) on „küürude ristumise punkt“:

$$t = \frac{1}{2} + \sigma^2 \ln \frac{p}{1-p}. \tag{4}$$

Sellisel valikul on mitu head omadust:

S1 Vektor $(\hat{y}_1, \ldots, \hat{y}_n)$ garanteerib keskmiselt minimaalse vigade arvu. Näiteks, kui

$$(y_1, \ldots, y_n) = (1, 1, 1, \ldots, 1)$$

ja hinnang on $(0, 0, 1 \ldots, 1)$, on tehtud kaks viga. Ülaltoodud hinnang garanteerib väikseima keskmise vigade arvu (keskmine on võetud üle kõikvõimalike $(y_1, \ldots, y_n)$ realisatsioonide).

S2 Vektor $(\hat{y}_1, \ldots, \hat{y}_n)$ on suurima tõepäraga vektor, st

$$(\hat{y}_1, \ldots, \hat{y}_n) = \arg\max_{(y_1,\ldots,y_n)\in\{0,1\}^n} P(Y_1 = y_1, \ldots, Y_n = y_n \mid X_1 = x_1, \ldots, X_n = x_n).$$

Üldiselt need kaks omadust ei käi koos, aga sõltumatuse tõttu meie lihtsa mudeli korral on see nii. Hinnangud $\hat{y}_1, \ldots, \hat{y}_n$ on väga lihtsasti leitavad – iga vaatlust x_i tuleb lihtsalt võrrelda punktiga t. Seega on hinnang $\hat{y}_i$ vaatluse x_i funktsioon, st $y_i = f(x_i)$. Nüüd on kerge leida ühe vaatluse klassifitseerimisvea tõenäosust: olgu X segujaotusega juhuslik suurus (indeks pole praegu oluline), Y tegelik klass (st $X = Y + Z$) ja $\hat{Y} = f(X)$. Klassifitseerimisvea tõenäosus avaldub:

$$\begin{aligned} P(\hat{Y} \neq Y) &= P(\hat{Y} = 1 \mid Y = 0) P(Y = 0) + P(\hat{Y} = 0 \mid Y = 1) P(Y = 1) \\ &= P(X > t \mid Y = 0)(1-p) + P(X \leq t \mid Y = 1) p \\ &= (1-p)(1 - \Phi_{1,\sigma}(t)) + p\Phi_{0,\sigma}(t), \end{aligned}$$

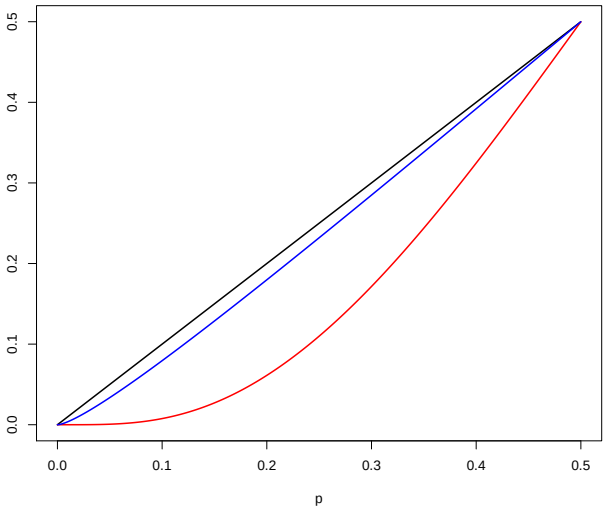
kus $\Phi_{i,\sigma}$ on $\mathcal{N}(i,\sigma^2)$ jaotusfunktioon ja t on defineeritud seosega (4). On selge, et klassifitseerimisvea tõenäosus sõltub standardhälbest σ – mida väiksem on σ, seda kergem on klassifitseerimine ja seda väiksem viga. Kui $\sigma \to 0$, siis $t \to 0,5$, $\Phi_{0,\sigma}(t) \to 0$, $\Phi_{1,\sigma}(t) \to 1$ ja seega $P(\hat{Y} \neq Y) \to 0$ (vt ka joonis 1).

Et vaatlused $X_1, \ldots, X_n$ on segujaotusega sõltumatud juhuslikud suurused, siis vastavad hinnangud $\hat{Y}_1, \ldots, \hat{Y}_n$ on sõltumatud Bernoulli jaotusega juhuslikud suurused, kuid nende juhuslike suuruste parameeter ei ole üldiselt enam p. Tõepoolest, tõenäosus, et hinnang $\hat{Y}$ on 0, avaldub

$$P(\hat{Y} = 0) = P(X \leq t) = p\Phi_{1,\sigma}(t) + (1-p)\Phi_{0,\sigma}(t).$$

Joonis 2 näitab, et hinnang $\hat{Y}$ ülehindab suurema tõenäosusega klassi, st kui klassi 1 tõenäosus p kasvab vahemikus $[0,0,5]$, siis hinnangu tõenäosus $P(\hat{Y} = 1)$ (mis sõltub nii p-st kui σ-st) kasvab aeglasemalt, kuigi $p = 0,5$ korral on need tõenäosused võrdsed (mis muidugi ei tähenda, et hinnang on veatu). Mida väiksem on σ, seda väiksem on hinnangu viga ja seda lähedasemad on ka p ja $P(\hat{Y} = 1)$. Seega, kui $\sigma = 1$ ja $p = 0,2$, siis $P(\hat{Y} = 1) \approx 0,06$. Seda illustreerib alljärgnev pilt (joonis 3), millel näeme tegelikke klasse $y_1, \ldots, y_n$ koos hinnangutega $\hat{y}_1, \ldots, \hat{y}_n$ juhul, kui $\sigma = 1$ ja $p = 0,2$. Näeme, et tegelikest klassidest on ligikaudu 20% ühed, täpselt nii nagu see suurte arvude seaduse tõttu olema peabki, samas hinnangud on märksa konservatiivsemad, st seal on vähem ühtesid (umbes 6%).

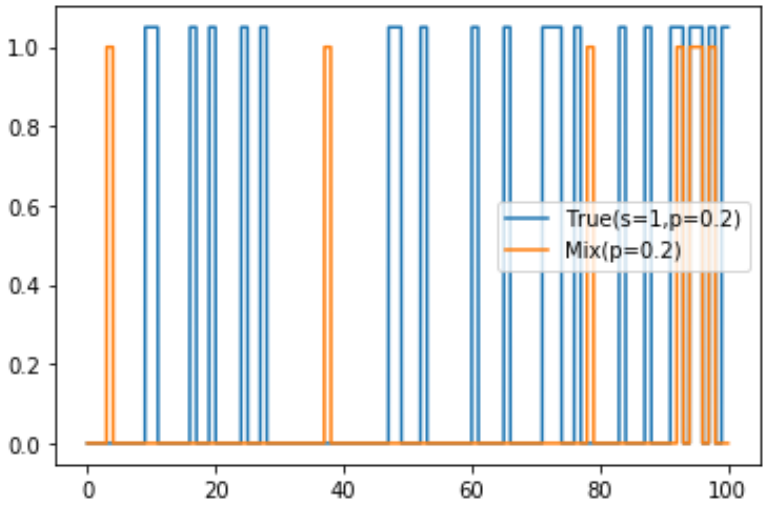
Bayesi lähenemine. Tuletame meelde klassifitseerimisülesannet: antud vaatluste $x_1, \ldots, x_n$ korral vastavate klasside hinnangute $\hat{y}_1, \ldots, \hat{y}_n$ leidmine. Eespool nägime, et segumudeli korral on mitmes mõttes hea hinnangu leidmine suhteliselt lihtne. Küll aga on selleks vaja teada segujaotuse parameetreid p ja σ, sest arv t sõltub ju mõlemast arvust. Praktikas tuleb aga tihti ette, et p ja σ pole teada. Sellisel juhul tuleb parameetrid enne klassifitseerimist hinnata (tehisõppes nimetatakse seda *plug-in* lähenemiseks) või kasutatada Bayesi lähenemisviisi. Parameetrite hindamise korral leitakse vaatluste põhjal hinnangud $\hat{\sigma}$ ja $\hat{p}$ (mõlemad sõltuvad vaatlustest) ja asetatakse need arvud valemisse (4).



Joonis 2: x-telg: p, must joon $y = p$; sinine joon $P(\widehat{Y} = 1)$ juhul, kui $\sigma = 0,4$; punane joon $P(\widehat{Y} = 1)$ juhul, kui $\sigma = 1$. Näeme, et mida suurem σ, seda vähem on hinnangus väiksema tõenäosusega klasse (seda konservatiivsem hinnang)

Varjatud tunnustega mudeli korral kasutatakse $\hat{\sigma}$ ja $\hat{p}$ leidmiseks enamasti nn EM-algoritmi. Need hinnangus ei pruugi olla täpsed. Samuti on metodoloogiliselt problemaatiline see, et parameetrite hindamiseks kasutatakse neid samu klassifitseerimisel kasutatavaid vaatlusi.

Bayesi lähenemisviisil vaadeldakse parameetreid teatava jaotusega juhuslike suurustena, mille jaotust nimetatakse **eeljaotuseks** (*prior distribution*), teoreetiliselt peaks eeljaotus postuleerima eelteadmisi parameetritest, praktikas valitakse eeljaotus enamasti muudest kriteeriumitest lähtuvalt. Kui parameetrite jaotuse kohta mingeid eelteadmisi pole (ja enamasti pole), siis (kui võimalik) võetakse eeljaotuseks tihti ühtlane jaotus. Ühtlast eeljaotust nimetatakse ka **mitteinformatiivseks eelmõõduks** (*non-informative prior*). Aga ükskõik milline eelmõõt ka pole, on Bayesi lähenemisviisi mudel keerukam, sest lisaks mõõdetud



Joonis 3: Sinine joon: tegelikud klassid, kollane joon: hinnangud. Vaatlusi pildil pole. Vasakul $\sigma = 1$ ja $p = 0,2$.

ja varjatud tunnusele on seal ka kolmas juhuslik suurus – parameeter.

Käesolevas artiklis teeme lihtsustava eelduse, et dispersioon σ on teada, eelmõõt on vaid parameetril p. Olgu see eelmõõt mitteinformatiivne, st ühtlane $U(0,1)$. See tähendab, et meil puuduvad eelteadmised p kohta, iga väärtus on võrdtõenäoline (st tihedus on konstantne). Parameetri kui juhuliku suuruse keskväärtus on 0,5. Seega on meie mudel järgmine:

B1 parameeter p on $U(0,1)$ jaotusega juhusliku suuruse realisatsioon (ehk p genereeritakse jaotusest $U(0,1)$);

B2 klassid $y_1,\dots,y_n$ on $B(1,p)$ jaotusega sõltumatute juhuslike suuruste realisatioon (ehk $y_1,\dots,y_n$ genereeritakse Bernoulli protsessist parameetriga p);

B3 Vaatlus x_i on $\mathcal{N}(y_i,\sigma^2)$ jaotusega juhusliku suuruse realisatsioon, vaatlused on sõltumatud (ehk x_i genereeritakse teistest vaatlustest sõltumatult jaotusest $\mathcal{N}(y_i,\sigma^2)$).

Segumudeli korral on samm **B1** puudub ja p on fikseeritud, kõik muu on sama. Muidugi võib sammud **B2** ja **B3** kokku võtta ja öelda, et $x_1, \ldots, x_n$ on sõltumatute segujaotustega juhuslike suuruste realisatsioonid. Aga järgnevas näeme, et segmenteerimise kontekstis on kasulikum sammud **B1** ja **B2** kokku võtta, sest nii saame uurida klasside $(Y_1, \ldots, Y_n)$ jaotust.

Beta-Bernoulli protsess. Uurime nüüd, millised on klasside tõenäosused eelmõõdu korral. Selleks vaatleme juhuslikke suurusi $Y_1, Y_2, \ldots$, mis iga p korral on Bernoulli protsess, kuid parameeter p on nüüd ühtlase jaotusega juhuslik suurus. Et ühtlane jaotus on Beta-jaotuse erijuht, nimetatakse juhuslikke suurusi $Y_1, Y_2, \ldots$ **Beta-Bernoulli protsessiks**. Alljärgnevas veendume, et Beta-Bernoulli protsess erineb Bernoulli protsessist. Selleks leiame iga n korral juhusliku vektori $(Y_1, \ldots, Y_n)$ (klassid) jaotuse. Olgu $(y_1, \ldots, y_n) \in \{0,1\}^n$ selline nullide ja ühtede vektor, milles ühtede arv on k (st $\sum_{i=1}^n y_i = k$). Selle vektori tõenäosus on täistõenäosuse valemi kohaselt järgmine:

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \int_0^1 p^k (1-p)^{n-k} \, dp = \frac{k!(n-k)!}{(n+1)!} = \frac{1}{(n+1)\binom{n}{k}}. \tag{5}$$

Paneme tähele, et tõenäosus (5) sõltub vaid sellest, kui palju on vektoris $(y_1, \ldots, y_n)$ ühtesid, mitte sellest, mis järjekorras ühed ja nullid on. Sellest järeldub, et juhuslike suuruste järjekorra vahetamine ei muuda jaotust (näiteks kui $n = 3$, siis vektorid

$$(Y_1, Y_2, Y_3), (Y_1, Y_3, Y_2), (Y_2, Y_1, Y_3), (Y_3, Y_1, Y_2), (Y_2, Y_3, Y_1), (Y_3, Y_2, Y_1)$$

on kõik sama jaotusega) ja sellest omakorda järeldub, et kõik juhuslikud suurused on sama (B(1 , 0,5)) jaotusega. Võttes valemis (5) $n = 1$ ja $k = 1$, saame jaotuse parameetriks 0,5. See on igati loogiline, sest ühtlase jaotuse keskväärtus on 0,5. Seega moodustavad Beta-Bernoulli protsessi sama jaotusega juhuslikud suurused. Kui nad oleksid sõltumatud, oleks Beta-Bernoulli protsess sama, mis Bernoulli protsess. On aga üsna selge, et need juhuslikud suurused pole sõltumatud. Näiteks valemist (5) saame $P(Y_1 = 1, Y_2 = 1) = 1/3 \neq P(Y_1 = 1)P(Y_2 = 1) = 1/4$. Veel enam

– selgub, et Beta-Bernoulli protsessi korral on juhuslikud suurused positiivselt korreleeritud. Tõepoolest, leiame

$$\operatorname{cov}(Y_1, Y_2) = E(Y_1 Y_2) - \frac{1}{4} = P(Y_1 = 1, Y_2 = 1) - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} > 0.$$

Kust see positiivne korreleeritus tuleb – on ju meie mudel sellline, et iga konkreetse parameetri väärtuse korral on juhuslikud suurused sõltumatud? Mingit positiivset korrelatsiooni pole keegi kusagil eeldanud ja nüüd on see äkki olemas!

Valemist (5) saame kergesti leida ka summa S_n jaotuse: tõepoolest

$$P(S_n = k) = \binom{n}{k} P\Big(Y_1 = \cdots = Y_k = 1, Y_{k+1} = \cdots = Y_n = 0\Big) =$$
$$\binom{n}{k} \frac{1}{(n+1)\binom{n}{k}} = \frac{1}{n+1}$$

ehk ühtede arvu jaotus on ühtlane. See on väga kaugel binoomjaotusest ja sellest järeldub, et ühtede proportsiooni S_n/n jaotus on ka sisuliselt ühtlane. See on igati loogiline, sest Beta-Bernoulli protsessi korral ei kehti suurte arvude seadus, sest ühtede proportsioon koondub ju ühtlase jaotusega juhuslikuks suuruseks p. Seega, erinevalt Bernoulli protsessist on Beta-Bernoulli protsessi korral täiesti vale eeldada, et suure n korral on ühtede proportsioon ligikaudu Y_i keskväärtus ehk 0,5.

Positiivne korreleeritus tähendab pikemat mälu. Mälu efekt tuleb eriti esile protsessi realisatioonide blokistruktuuris. Iga binaarset (0/1) jada võib esitada blokkidena. Näiteks

$$\underbrace{0000}_{\text{esimene blokk}} \quad \underbrace{11111111}_{\text{teine blokk}} \quad \underbrace{00}_{\text{kolmas blokk}} \quad \underbrace{111}_{\text{neljas blokk}} \quad \underbrace{0}_{\text{viies blokk}} \quad \underbrace{11111}_{\text{kuues blokk}} \cdots$$

Ehk teades blokkide pikkusi ja esimese bloki värvi võime üheselt taastada terve jada. Blokid jagunevad 0- ja 1-blokkideks. Bernoulli protsessi korral on blokid sõltumatud, iga 0-bloki jaotus on $G(p)$ (geomeetriline jaotus parameetriga p) ja iga 1-bloki jaotus on $G(1-p)$. Seega iga 0-bloki keskmine pikkus on $1/p$ ja iga 1-bloki keskmine pikkus

on $1/(1-p)$. Mida suurem p, seda pikemad 1-blokid ja seda lühemad 0-blokid. Igati loogiline. Beta-Bernoulli protsessi korral pole blokid enam sõltumatud, küll aga on kõik blokid ühe ja sama jaotusega. Leiame selle jaotuse keskväärtuse ehk keskmise blokipikkuse. Olgu T esimese bloki pikkus, tema jaotus ja keskväärtus avalduvad valemi (5) kohaselt

$$\begin{aligned} P(T=k) &= P(Y_1=\cdots=Y_k=1, Y_{k+1}=0) + P(Y_1=\cdots=Y_k=0, Y_{k+1}=1) \\ &= \frac{2}{(k+1)(k+2)}; \\ ET &= \sum_{k=1}^{\infty} kP(T=k) = \sum_{k=1}^{\infty} \frac{2k}{(k+1)(k+2)} = \infty. \end{aligned}$$

Seega keskmine blokipikkus on lõpmatus! See ei tähenda, et selle protsessi realisatsioonid oleksid vaid konstantsed jadad (st jadad $000000\cdots$ ja $1111\cdots$), seda mitte. Juhusliku suuruse keskväärtus võib olla lõpmatu ka siis, kui see juhuslik suurus ise on alati lõplik. Beta-Bernoulli protsessi realisatsioon võib olla näiteks selline, et keskmine 0-bloki pikkus on $1/0,4$ ja keskmine 1-bloki pikkus on $1/0,6$, see juhtub siis, kui juhuslik parameeter p juhtub olema 0,4. Et aga $\int_0^1 (p)^{-1}\,dp = \infty$, saamegi keskmiseks blokipikkuseks lõpmatuse.

Lõpetuseks anname Beta-Bernoulli protsessile veel ühe interpretatsiooni. Kujutleme urni, kus on üks must pall ja üks valge pall. Juhuslikult võetakse üks pall (musta palli valimise tõenäosus on 0,5) ja pannakse urni tagasi koos veel ühe sama värvi kuuliga. Nüüd on urnis kolm kuuli. Kui eelnevalt võeti must kuul, siis kolmest kuulist kaks on nüüd mustad. Seejärel võetakse juhuslikult järgmine kuul (kui esimesel katsel võeti must, siis teisel katsel on musta võtmise tõenäosus $2/3$) ja pannakse see tagasi koos veel ühe sama värvi kuuliga jne. Seega peale iga juhuslikku valikut kuulide arv urnis kasvab ühe võrra. Seda mudelit nimeatakse **Polya urniks**. Olgu $Y_1, Y_2, \ldots$ väljavõetud kuulide värvid (kodeeritud kui 0 ja 1). Selgub, et need juhuslikud suurused moodustavad Beta-Bernoulli protsessi. Seega vaatleja, kes ei tea kuulide välja võtmise protseduuri ja näeb ainult väljavõetud kuulide jada $y_1, y_2, \ldots$, võib teha (täiesti õige) järelduse, et väljavõetud kuulide jada on Bernoulli protsessi realisatsioon parameetriga, näiteks, 0,8. Aga järgmine järeldus, et vaatlusi genereе-

riv protseduur on järelikult Bernoulli mudel (st juhuslikud suurused $Y_1, Y_2, \ldots$ moodustavad Bernoulli protsessi) parameetriga 0,8 ja ühel värvil on seega suurem tõenäosus, on aga täielikult vale. On ju mudel täiesti sümmeetriline, ühelgi värvil pole eelist ja Polya urni protsesi korrates (st kahest kuulist uuesti alustades) näeme hoopis uut Bernoulli protsessi realisatsiooni, kuid sedapuhku hoopis teise parameetriga, näiteks 0,1. Praktikas aga tehakse järeldus genereeriva protsessi kohta enamasti ikka vaid ühe vaatluste jada põhjal ja, nagu nägime, võib see viia väga valele järeldusele.

Segmenteerimine Bayesi lähenemisviisil. Tuleme nüüd tagasi segmenteerimise ehk klassifitseerimise juurde. Probleem on ikka sama – meil on Bayesi mudelist **B1, B2, B3** pärinevad vaatlused $x_1, \ldots, x_n$ ja eesmärk on vaatlusi genereerivate klasside $y_1, \ldots, y_n$ hindamine. Ainus erinevus eelpoolkirjeldatud segumudeli klassifitseerimisest on see, et juhuslikud suurused $Y_1, \ldots, Y_n$ pärinevad Beta-Bernoulli, mitte Bernoulli protsessist. See paraku muudab väga palju. Et vaatlused pole enam sõltumatud, siis segumudeli klassifitseerimisreegel (3) ei garanteeri enam hinnanguid, mis rahuldavad omadusi **S1** ja **S2**. Veel enam, sõltuvate vaatluste korral ei leidu üldiselt hinnanguid, mis üheaegselt rahuldaksid mõlemaid omadusi – uurija peab valima, mis mõttes parimat hinnangut ta tahab. Selles artiklis valime **S2**, st otsime suurima tõepäraga vektorit. Edaspidi nimetame seda **MAP** (*maximum a posteriori*) hinnanguks. Siin aga põrkame järgmisele raskusele – sõltuvate vaatluste korral pole selliseid hinnanguid enam üldse nii kerge leida – hinnang $\hat{y}_i$ sõltub kõikidest vaatlustest, mitte ainult vaatlusest x_i. Seega Bayesi lähenemisviisil pole MAP hinnangut (ning ka omadust **S1** rahuldavat hinnangut) üldse kerge leida ja üldist eeskirja selleks pole.

Tähistame

$$q(y_1, \ldots, y_n) = P(Y_1 = y_1, \ldots, Y_n = y_n)$$

ning

$$p(x_1, \ldots, x_n | y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{y_i, \sigma^2}(x_i).$$

Seega $p(x_1,\dots,x_n|y_1,\dots,y_n)$ on vaatluste tinglik tihedus (tihedus, mitte tõenäosus, sest vaatlused on pidevatest juhuslikest suurustest) tingimusel, et klassid on $(y_1,\dots,y_n)$. Seega korrutis

$$p(x_1,\dots,x_n|y_1,\dots,y_n)q(y_1,\dots,y_n)$$

on vaatluste ja klasside ühistihedus ja vaatluste tihedus $p(x_1,\dots,x_n)$ avaldub

$$p(x_1,\dots,x_n) = \sum_{(y_1,\dots,y_n)\in\{0,1\}^n} p(x_1,\dots,x_n|y_1,\dots,y_n)q(y_1,\dots,y_n).$$

Nüüd paneme tähele, et

$$P(Y_1 = y_1,\dots,Y_n = y_n \mid X_1 = x_1,\dots,X_n = x_n) = \frac{p(x_1,\dots,x_n|y_1,\dots,y_n)q(y_1,\dots,y_n)}{p(x_1,\dots,x_n)}$$

ja see tähendab seda, et MAP hinnangu leidmine on ekvivalentne vaatluste ja klasside ühistiheduse maksimisserimisega üle klasside vektori. Võttes logaritmi (monotoone funktsioon, mis ei muuda arg max väärtust), saame, et MAP vektor maksimiseerib summat

$$\ln p(x_1,\dots,x_n|y_1,\dots,y_n) + \ln q(y_1,\dots,y_n).$$

Et

$$\ln p(x_1,\dots,x_n|y_1,\dots,y_n) = \sum_{i=1}^{n} \ln f_{y_i,\sigma^2}(x_i)$$

ja

$$\ln f_{y_i,\sigma^2}(x_i) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-y_i)^2}{2\sigma^2},$$

saame, et MAP vektor maksimiseerib järgmist avaldist

$$-\sum_{i=1}^{n} \frac{(x_i-y_i)^2}{2\sigma^2} + \ln q(y_1,\dots,y_n).$$

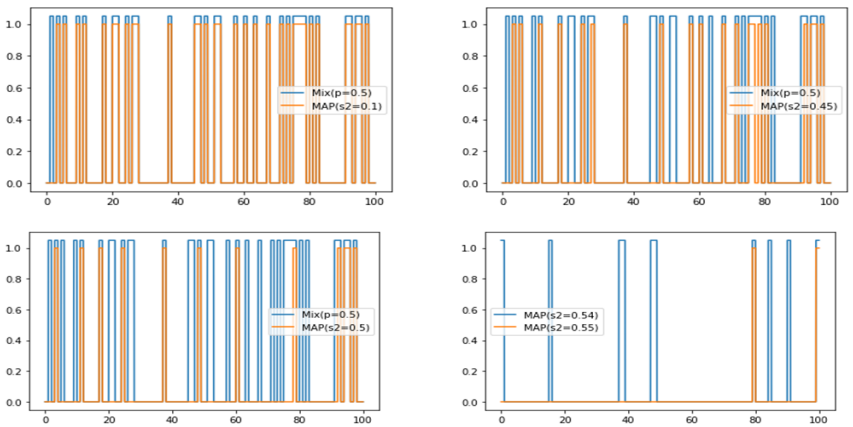
Korrutades ülaltoodud avaldise läbi konstandiga $-2\sigma^2$, saame, et otsitav MAP hinnang $(\hat{y}_1,\dots,\hat{y}_n)$ on järgmine:

$$(\hat{y}_1,\dots,\hat{y}_n) = \arg\min_{(y_1,\dots,y_n)\in\{0,1\}^n}\Big[\sum_{i=1}^{n}(x_i-y_i)^2 - 2\sigma^2\ln q(y_1,\dots,y_n)\Big]. \quad (6)$$

Ülesanne (6) on tüüpiline tehisõppe probleem. Sulgudes olevast avaldisest esimene on pelgalt andmetest sõltub minimiseeritav **kaofunktsioon** ja $-\ln q(y_1,\dots,y_n)$ on ainult mudelist sõltuv nn **karistusliige**. Seega MAP hinnang üritab samaaegselt minimiseerida kaofunktsiooni ja maksimiseerida tõenäosust $q(y_1,\dots,y_n)$. Konstant $2\sigma^2$ – nn **regulariseerimiskonstant** – reguleerib nende osade vahekorda. Kui σ^2 on väga väike, siis mudelil pole mõju ja MAP minimiseerib vaid summat $\sum_{i=1}^{n}(x_i-y_i)^2$. Sellisel juhul $\hat{y}_i = 1$ parajasti siis, kui $x_i > 1/2$ ning meil on jälle reegel (3). Selline olukord on ka siis, kui $Y_1,\dots,Y_n$ on Bernoulli mudelist parameetriga 0,5. Tõepoolest, sellisel juhul on iga vektori tõenäosus võrdne $q(y_1,\dots,y_n) = (0,5)^n$ ja sellisel juhul pole karistusliikmel mingit mõju. Kui aga σ on väga suur, siis kaob andmete mõju ning MAP hinnang maksimiseerib vaid tõenäosust $q(y_1,\dots,y_n)$. Nii Bernoulli kui ka Beta-Bernoulli mudeli korral

$$\max_{(y_1,\dots,y_n)\in\{0,1\}^n} q(y_1,\dots,y_n) = \max\{q(1,1\dots,1), q(0,0\dots,0)\},$$

st suurima tõenäosusega vektor on alati konstantne. See tähendab, et (fikseeritud andmete korral) arvu σ^2 suurendades muutub hinnang järjest konservatiivsemaks, st blokid muutuvad pikemaks ja nende arv väheneb. Edaspidi nimetame hinnangut (6) **Bayesi MAP** hinnanguks, kui q vastab Beta-Bernoulli protsessile (valem (5) ja σ^2 on tegelik parameeter ning hinnangut (6) nimetame **mudelipõhiseks (MAP) hinnanguks**, kui q vastab Bernoulli protsessile tegeliku parameetriga p (st $q(y_1,\dots,y_n) = p^k(1-p)^{n-k}$, kus $k = \sum_i y_i$). Mudelipõhine MAP hinnang on sama, mis (3).



Joonis 4: Andmed on genereeritud parameetritega $\sigma^2 = 1$ ja $p = 0,2$, andmeid joonistel pole. Sinine joon minimiseerib $\sum_{i=1}^{n}(x_i - y_i)^2$, kollane on hinnang (6) Beta-Bernoulli mudeli korral. Üleval vasakul: $\sigma^2 = 0,1$, üleval paremal $\sigma^2 = 0,45$, all vasakul: $\sigma^2 = 0,5$, all paremal on jooned (6) $\sigma^2 = 0.54$ ja $\sigma^2 = 0,55$ korral. Näeme kuidas σ^2 kasvades hinnang (6) muutub konservatiivsemaks, alates $\sigma^2 = 0,66$ on hinnang (6) konstantne. Seega Bayesi MAP hinnang, mis vastab tegelikule parameetrile $\sigma^2 = 1$ on konstantselt 0.

Bayes ja mudelipõhine MAP. Järgnevas uurime Bayesi lähenemisviisi mõju MAP hinnangule. Järgnevad joonised 4 ja 5 võrdlevad kahte erinevat joondust samadel andmetel: Bayesi MAP ning mudelipõhine MAP. Näeme, et Bayesi MAP on palju konservatiivsem kui mudelipõhine. Konservatiivsus siinkohal tähendab, et blokkide arv on väiksem ja blokid seetõttu pikemad. Joonis 6 illustreerib seda eriti ilmekalt. Sellel joonisel on genereeritud iga σ korral vahemikust $(0;2]$ (teatava sammuga) 100 erinevat valimit pikkusega 101 parameetriga $p = 0,2$. Iga valimi korral on leitud Bayesi MAP ja mudelipõhine MAP ning arvutatud nende hinnangute blokkide arv n. Ütleme siis, et $n-1$ on **hüpete arv**. Konstantse hinnangu korral on see arv 0. Seejärel on hüpete arv keskmistatud üle 100 korduse. Mida väiksem on blokkide ja hüpete arv, seda konservatiivsem on hinnang. Joonis näitab ilmekalt, et σ suurenemisel muutuvad mõlemad hinnangud konservatiivsemaks, kuid

Bayesi MAP teeb seda palju kiiremini ning kui σ ületab 0,76, siis on kõik Bayesi MAP hinnangud konstantsed ja seda sõltumata andmetest! See tähendab, et statistilise analüüsi tulemus (Bayesi meetodil) ei pruugi üldse sõltuda andmetest (välja arvatud ehk see, kas hinnang on konstantselt 0 või 1)! Millest see tuleb? Põhjus on Beta-Bernoulli protsessi mälus, mistõttu on konstantsetel ja konservatiivsematel jadadel Beta-Bernoulli protessi korral märksa suurem kaal kui Bernoulli protsessi korral. Veendume selles. Olgu $q(y_1,\dots,y_n) = p^k(1-p)^{n-k}$ (k on ühtede arv vektoris), st $(Y_1,\dots,Y_n)$ jaotus Bernoulli protsessi korral. Kui $p < 0{,}5$, siis kõige suurema tõenäosusega vektor on $(0,\dots,0)$ ja vastav tõenäosus on $q(0,\dots,0) = (1-p)^n$. Tõenäosuselt järgmine vektor on selline, kus on täpselt üks 1 ning suurima ja suuruselt teise tõenäosusega vektorite tõenäosuste suhe on konstantne: $q(0,0,\dots,0)/q(1,0,\dots,0) = (1-p)/p$. Kehtigu nüüd $q(y_1,\dots,y_n) = \left((n+1)\binom{n}{k}\right)^{-1}$, st $(Y_1,\dots,Y_n)$ jaotus Beta-Bernoulli protsessi korral. Sellisel juhul on kõige suurema tõenäosusega vektoreid kaks: $(0,\dots,0)$ ja $(1,\dots,1)$, mõlema tõenäosus on $1/(n+1)$. Tuletame meelde, et kokku on 2^n vektorit, mille tõenäosuste summa peab olema 1. Kuigi suure n korral on nii $(1-p)^n$ kui ka $1/(n+1)$ väga väikesed arvud, on nende vahe üüratu, sest $(1-p)^n$ kahaneb eksponentsiaalselt ning isegi kui $1-p$ on ligikaudu 1 ja seetõttu on vähegi suure n korral $(1-p)^n$ kordades väiksem kui $1/(1+n)$. Seega Bayesi lähenemisviis annab konstantsetele jadadele oluliselt suurema kaalu. Uurime suuruselt järgmist vektorit, kus on täpselt üks 1 (või täpselt üks 0). Sellise vektori tõenäosus on $q(1,0,\dots,0) = 1/n(n+1)$, mistõttu suhe $q(0,0,\dots,0)/q(1,0,\dots,0) = n$ (Bernoulli protsessi korral see suhe on konstantne). See tähendab, et Beta-Bernoulli protsessi korral konstantse vektori suhteline kaal on (vähegi suure n korral) kordades suurem kui tõenäosuselt järgmiste vektorite kaal. Bernoulli protsessi korral see nii pole. Ja see selgitabki, miks Beta-Bernoulli protsessi korral näemegi enamasti konstantseid jadasid – mudeli mõju ületab andmete mõju. Bernoulli protsesi korral seevastu on konstantsete ja väga konservatiivsete vektorite suhteline kaal väiksem ja seetõttu ei näe me neid nii tihti.

Kokkuvõte ja järeldused

Käesolevas artiklis tutvusime lihtsaima varjatud tunnustega mudeliga – segumudeliga. Seejärel vaatlesime tehisõppe üht tüüpülesannet – segmenteerimine – ja veendusime, et segumudeli abil (ehk eeldusel, et vaatlused on sõltumatud ja pärinevad samast segumudelist) on hea hinnangu (ja seda mitmes mõttes) leidmine imelihtne – iga vaatlust tuleb klassifitseerida eraldi ja vaatluse x_i klassifitseerimiseks tuleb seda lihtsalt võrrelda arvuga t. Arv t sõltub müra dispersioonist σ^2 ja klasside tõenäosusest p. Praktikas pole need numbrid enamasti teada ja sellisel juhul tuleb neid andmete põhjal hinnata. Käesolevas tekstis eeldame lihtsuse mõttes, et σ^2 on teada, klasside tõenäosus aga mitte. Kui andmed oleksid kujul $(x_1, y_1), \ldots, (x_n, y_n)$, kus y_i on teadaolevad klassid ja eesmärk klassita vaatluse x klassiftseerimine (*supervised learning*), siis $\hat{p} = \frac{1}{n}\sum_i y_i$ on väga mõistlik hinnang, kuid meil on andmed klassideta (*unsupervised learning*) ja sellisel juhul pole p hindamine lihtne. Samas sõltub t väga arvust p ja seetõttu selle arvu fikseerimine mingil (uurijale õigena tunduval) väärtusel on väga tugev eeldus, mis kindlasti mõjutab tulemust. Iga eeldus on kitsendus ja enamasti on eelduste kontrollimine praktikas keeruline. Valedel eeldustel tehtud analüüs võib aga viia valedele järeldustele ja seetõttu eelistakse meetodeid, mis eeldavad vähem. Tuletame meelde meie eeldused: vaatluste sõltumatus, (tinglik) normaaljaotus ning σ^2 ja p. Normaalne müra ja sõltumatus on väga loomulikud eeldused, σ^2 on meil fikseeritud artikli loetavuse huvides, jääb aga tugev eeldus p. Sel ja paljudel muudel põhjustel kasutatakse tihti Bayesi lähenemisviisi, kus arvu p käsitletakse mingi eeljaotusega juhusliku suurusena ja uurija ei eelda nüüd enam mingit konkreetset arvu vaid jaotust. Kui ta arvab, et p on pigem suur, siis võiks see jaotus panna suurtele väärtustele suuremad tõenäosused; kui aga uurijal mingisuguseid eelteadmisi pole, valitakse tihti ühtlane jaotus. Ühtlane jaotus ei anna ühelegi väärtusele eelist ning nii tundubki, et mingeid eeldusi p kohta enam pole. Seega üks eeldus vähem, kõik muu sama! Paraku on see illusioon, sest juhtub midagi üpris ootamatut – asendades fikseeritud p ühtlase jaotusega muudame oluliselt varjatud tunnuste jaotust ning asendame Bernoulli protsessi Beta-Bernoulli protsessiga.

See aga muudab kõik. Eelkõige seda, et klassid $Y_1, \ldots, Y_n$ ja seetõttu ka vaatlused $X_1, \ldots, X_n$ pole enam sõltumatud. Seega loobudes ühest eeldusest muutsime väga oluliselt teist – ja kõige olulisemat eeldust – sõltumatus. See tähendab, et meil on nüüd hoopis teine mudel, mida uurija tegelikult ei eelda, sest tegelikult on uurija väga rahul sõltumatusega, ta lihtsalt ei tea p väärtust ja üritab Bayesi lähenemisviisi abil p hindamist/eeldamist vältida.

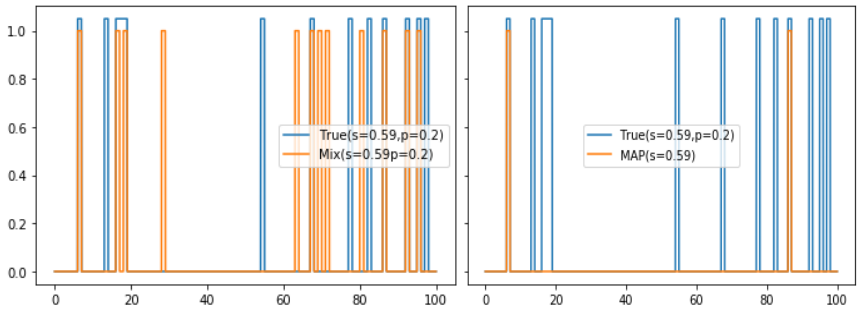
Mida uus mudel ehk Bayesi lähenemine segmenteerimisel enesega kaasa toob? Nagu nägime, tekib kohe hulk probleeme:

- Segmenteerimise eesmärk pole enam ühene. Sõltumatu mudeli korral kasutatav hinnang on väga hea mitmes mõttes: iga konkreetse vaatluse x_i korral minimiseerib hinnang $\hat{y}_i$ klassifitseerimisviga ja sellest järeldub omadus **S1**. Samuti on saadud hinnang MAP hinnang ehk kehtib omadus **S2**. See tähendab seda, et praktik ei pea nuputama, mis on segmenteerimise eesmärk. Sõltuvuse korral see aga nii pole ja hinnang, mis rahuldab omadust **S1**, ei ole üldiselt MAP hinnang. On palju muid võimalusi ja see tähendab, et praktik peab väga täpselt teadma, mida ta tahab. Enamasti, paraku, pole tal õrna aimugi. Käesolevas artiklis oleme valinud MAP hinnangu, praktikas levinud valik.

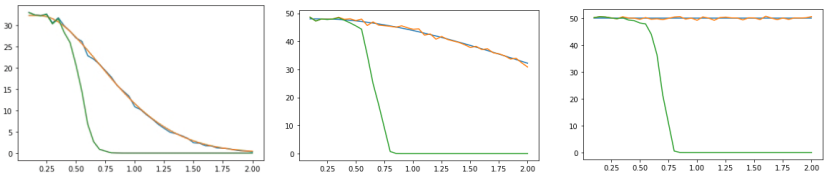
- Hinnangute leidmine on nüüd arvutuslikult väga keeruline, tihti võimatu. Sõltumatuse korral klassifitseeriti iga vaatlus eraldi võrreldes seda arvuga t. Beta-Bernoulli mudeli korral aga sellist lihtsat eeskirja MAP hinnangu leidmiseks pole ja selle leidmine võib osutuda väga keerukaks. Kuigi kõikvõimalikke vektoreid on 2^n ja iga vektori korral tingliku tõenäosuse arvutamine on väga lihtne, siis maksimumi leidmine 2^n arvu seast käib ka suhteliselt väikese n korral arvutile üle jõu. Piltidel saadud joonte jaoks kasutati teatavat iteratiivset algoritmi, et maht $n = 101$ on suhteliselt väike, siis on alust arvata, et hinnangud on korrektsed. Kuidas leida omadust **S1** rahuldavat hinnangut Beta-Bernoulli protsessi korral, on autorile teadmata.

- Hinnangute konservatiivsus. Veendusime, et Bayesi meetodil saadud MAP hinnangud on tunduvalt konservatiivsemad ja tihti konstantsed (st üks blokk ja mitte ühtegi hüpet). Simulatsioonid näitasid, et kui σ ületab teatud piiri, siis hinnang on konstantne iga vaadeldud valimi korral, st andmed ei mõjuta hinnangut! Nägime, et selle konservatiivsuse taga on Beta-Bernoulli protsessi mälu ja seetõttu on konservatiivsetel vektoritel Bernoulli protsessiga võrreldes oluliselt suurem kaal. Mida see praktikas tähendab? Esmapilgul ehk suurt pettumust. Tõepoolest, mis iganes tegelik p ka poleks, tüüpiline jada koosneb blokkidest, mille pikkus sõltub p-st. Nägime, et isegi kui MAP jada hüppab, on blokke vähem ja keskmine pikkus erinev (pikad blokid pikemad). Sama muidugi kehtib ka mudelipõhise MAP korral, aga Bayesi lähenemisel on see erinevus eriti reljeefne. Siit järeldus: MAP hinnang, eriti Bayesi lähenemisviisil, ei peegelda kuidagi varjatud protsessi tüüpilist käitumist! Tihti aga seda eeldatakse (kas või alateadlikult) ning ebatüüpiline, näiteks konstantne hinnang tundub vale. Tegelikult on ju kindlasti rohkem hüppeid! Iseäranis nadi tundub olukord siis, kui hinnang on kõikide andmete korral konstantne – statistiline analüüs, mis ei sõltu põrmugi andmetest, tundub kui mitte vale, siis vähemalt mõttetu. Kusagil peab olema mingi viga! Aga tuletame meelde – MAP hinnang on teatavas mõõtes parim ja parim pole kunagi tüüpiline. Muuhulgas tähendab hinnangute konservatiivsus ka seda, et MAP hinnangu põhjal ei saa ega tohi hinnata tõenäosust p (see on viga, mida tihti tehakse). Samas kui MAP hinnang hüppab (kas või ühes kohas), tähendab see seda, et andmetes peab tõepoolest olema mingi oluline muutus.

Kas see kõik tähendab, et Bayesi MAP on kasutu hinnang? Sugugi mitte, isegi kui konstantne, teeb hinnang just seda, mida talt oodatakse – mitte ühegi parameetri p väärtuse korral ei ole ta ilmselt suurima tõenäosusega (parim), aga keskmistades üle parameetri tingliku jaotuse on seesama konstantne hinnag kõige parem.



Joonis 5: Andmed on genereeritud parameetritega $\sigma = 0,59$ ja $p = 0,2$, andmeid joonistel pole. Sinine joon on tegelikud klassid, kollane joon vasakul on mudelipõhine MAP ja kollane joon paremal on Bayesi MAP.



Joonis 6: x-telg: σ; roheline joon: keskmine (üle 100 korduse, iga kord on genereeritud uued andmed pikkusega $n = 101$ ja leitud Bayesi ning mudelipõhine MAP) hüpete arv Bayesi MAP hinnangul, sinine joon: keskmine hüpete arv mudelipõhisel hinnangul; kollane joon: teoreetiline keskmine hüpete arv mudelipõhisel hinnangul. Vasakul $p = 0,2$, keskel $p = 0,4$, paremal $p = 0,5$. Kui $p = 0,5$, siis iga σ korral $P(\hat{Y} = 1) = 0,5$ ja seega keskmine blokipikkus on 2 ning keskmine hüpete arv seega iga σ korral on $101/2 - 1$. Näeme, et Bayesi MAP on selgelt konservatiivsem kui mudelipõhine hinnang, veel enam – kui $\sigma > 0,8$, on Bayesi MAP konstantne, seda iga p ja kõikide andmete korral.

Lõpetuseks – käesolevas artiklis käsitlesime ehk kõige lihtsamat varjatud tunnustega mudelit ja nägime, kuidas katse vaid ühest lihtsast eeldusest Bayesi lähenemisviisi abil loobuda muudab kardinaalselt kogu mudelit ja rikub ära statistika kõige ihaldatuma eelduse – sõltumatuse. Samuti nägime, kuidas mudel tihti domineerib andmete üle ja analüüsi tulemus

on andmetest suuresti sõltumatu. Aga meie mudel on väga lihtne ja nii oskame tulemusi interpreteerida ning mudeli mõju hinnata. Paraku kasutatakse praktikas tänapäeval väga keerulisi mudeleid, kus on palju varjatud tunnuseid ja väga palju parameetreid. Bayesi lähenemisviis on populaarne ja seetõttu pannakse parameetritele eelmõõdud. Nii saadud mudel on enamasti ülikeeruline ja selle teoreetiline uurimine ja millegi täpne välja arvutamine pea võimatu. Aga arvutid ja Monte-Carlo meetodid teevad oma töö ning tulemusi ja publikatsioone tuleb kui Vändrast saelaudu. Aga mida nad tegelikult näitavad?
[1]

[1] PRG865

