SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Gibbsovo uzorkovanje

Filip Boltužić

Voditelj: *Prof. dr. sc. Bojana Dalbelo Bašić*

Zagreb, veljača 2014.

SADRŽAJ

1. Uvod

Gibbsovo uzorkovanje pripada Monte Carlo Markovljevim (engl. *Monte Carlo Markov Chain, MCMC*) metodama. Markovljevim lancima modelira se bezmemorijski matematički sustav stanja i prijelaza između stanja (Kass et al., 1998). Markovljev lanac je niz vrijednosti generiranih u procesu s Markovljevim svojstvom. Prema Markovljevom svojstvu međuovisnost postoji samo između susjednih vrijednosti, tj. vrijednost u Markovljevom lancu generira se samo na temelju prethodne vrijednosti, zbog čega se nazivaju bezmemorijskim.

Gibbsovo uzorkovanje zadovoljava obilježja Monte Carlo metoda i Markovljevih lanaca. Metoda generira niz uzoraka, gdje su generirani susjedi međusobno zavisni (korelirani). Početni uzorci (engl. *burn out period*) Gibbsovog uzorkovanja se često zanemaruju jer ne predstavljaju ciljanu distribuciju.

1.1. Pretpostavke

Distribucija iz koje je potrebno uzorkovati je "posebna", jer nije moguće uzorkovati izravno. Izračun vjerojatnosti sunčanog ili kišovitog vremena tijekom sutrašnjeg dana proces je koji je moguće procijeniti Gibbsovim uzorkovanjem. No, informacije koje su potrebne za Gibbsovo uzorkovanje su informacije o uvjetnim distribucijama vjerojatnosti sunčanog, odnosno kišovitog vremena. Potrebno je poznavati vjerojatnosti o vremenu temeljem današnjeg vremena. Odnosno, ukoliko nas zanimaju $P(sutra = kiša)$ i $P(sutra = sunce)$, potrebno je poznavati vjerojatnosti:

- $P(sutra = kiša|danas = kiša)$
- $P(sutra = kiša|danas = sunce)$
- $P(sutra = sunce|danas = kiša)$
- $P(sutra = sunce|danas = sunce)$

2. Monte Carlo metode

2.1. Povijest

Monte Carlo metode obuhvaćaju računalne modele zasnovane na stohastičkoj matematici, točnije, uporabi nasumičnih brojeva u izračunima. Moderne Monte Carlo metode osmislio je Stanislaw Ulam (Kass et al., 1998), a ime su dobile po omiljenom odredištu zabave Ulamovog ujaka – Monte Carlo kockarnicama. Originalno su zamišljene kako bi pomogle pri difuziji neutrona. Svrha Monte Carlo metoda bila je izraditi umjetno stvoriti nasumičan proces bez mogućnosti upravljanja njime. John von Neumann, slavni znanstvenik, uvidio je potencijal Monte Carlo metode i implementirao ih na računalu ENIAC. Intenzivnija primjena Monte Carlo metoda počela je s pojavom snažnijih računala.

2.2. Mehanizam

Monte Carlo metode pokušavaju oponašati slučajne procese u prirodi (npr. bacanje novčića). Na taj način pokušavaju se predvidjeti svi mogući ishodi i vjerojatnosti događaja unutar okvira zadanog procesa.

Monte Carlo metoda je probabilistički računalni algoritam koji pokušava predvidjeti sve moguće ishode i vjerojatnosti procesa na koji je primijenjen. Temelji se na slučajnim varijablama, koje je potrebno zadati u obliku funkcije gustoće (Gilks et al., 1996). Na temelju funkcija gustoće $P(x)$ Monte Carlo metodama moguće je riješiti probleme:

- generiranja R uzoraka $\{x^{(r)}\}_{r=1}^{R}$ iz $P(x)$,
- izračun očekivanja funkcija s distribucije $P(x)$.

Monte Carlo metodama se simuliraju sustavi s mnogo neizvjesnosti. Pokazale su se iznimno učinkovite prilikom modeliranja složenih vjerojatnosti i strategija odlučivanja kojima nije moguće upravljati, ili ih je iznimno teško ili zahtjevno izgraditi.

Izračun osiguranja, modeliranje kreditnih rizika u financijskim institucijama, rekonstrukcija eksplozija samo su neki primjeri u kojima se koriste Monte Carlo metode, zbog činjenice da je navedene procese u stvarnosti teško simulirati i pratiti.

U Monte Carlo analizama potrebno je definirati distribucije vjerojatnosti slučajnih varijabli. Markovljevim lancima moguće je modelirati međuovisnost, što nije moguće Monte Carlo metodama. Ukoliko je potrebno dobiti uzorke iz aposteriori distribucije vjerojatnosti, Markovljevim lancima moguće je uzastopno uzorkovati dok uzorkovanje ne postane stabilan proces. Tada je moguće dobiti nezavisne uzorke iz aposteriori distribucije vjerojatnosti.

Monte Carlo integracija pripada Monte Carlo metodama, a služi za izračun integrala.

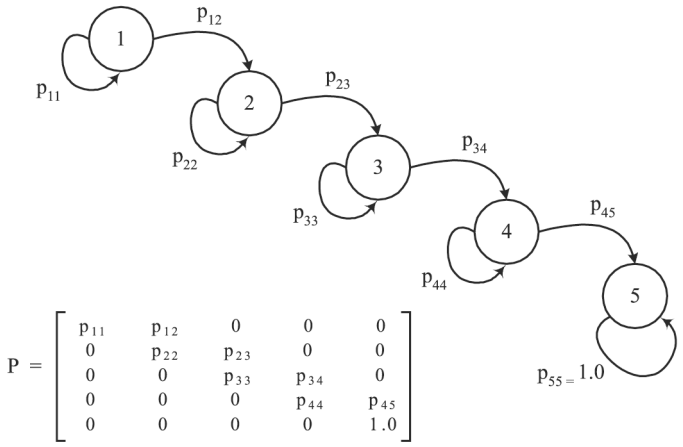
Monte Carlo metode imaju svojstvo crne kutije (engl. *black box*). Njima je moguće dobiti odgovore na razna pitanja procjene vjerojatnosti, kao: kolika je vjerojatnost kiše sutra, kolika je vjerojatnost da će Googleova dionica izgubiti na vrijednosti, kolika je vjerojatnost da će klijent vratiti kredit ... No, nije moguće saznati koji su razlozi na temelju kojih je dobivena određena vjerojatnost.

3. Markovljev lanac

Markovljeve lance inicijalno je predstavio ruski znanstvenik Andrey Markov 1912. godine. Markov je pokušao modelirati slavno remek djelo Aleksandra Puškina *Jevgenij Onjegin*. Cilj modela bio je predviđanje idućih slova temeljem trenutnih. Markovljeva ideja je po prvi puta predstavila koncepte zavisne varijable i uvjetne vjerojatnosti.

Markovljevi lanci se, također, primjenjuju nad širokim spektrom problema. Koriste se za modeliranje događaja i njihovih vjerojatnosti kada nije potrebno pamtiti prošle događaje. Markovljevi lanci sastoje se od niza stanja i vjerojatnosti prijelaza između stanja. Primjer Markovljevog lanca prikazan je slikom 3.1. Na slici je vidljivo pet stanja označenim brojevima u krugovima, jednosmjernih strelica kojima se označavaju mogućih prijelazi te matrica prijelaza P, matrični zapis Markovljevog lanca. Markovljev lanac ima oblik usmjerenog grafa. Stanja na burzama dionica, vremenska prognoza, prepoznavanje govora su neke od problema koji se često modeliraju Markovljevim lancima.

Markovljevi lanci pretpostavljaju kako je sustav koji se modeliraju stabilan. U praksi, to često nije tako. Kompleksni sustavi, kao što su burze dionica, sastoje se od pravilnosti i šuma. No, često moguće procijeniti jesu li vjerojatnosti dobivene danas



Slika 3.1: Primjer Markovljevog lanca. Preuzeto iz (Wirahadikusumah i Abraham, 2003).

relevantne za nekoliko godina.

4. Gibbsovo uzorkovanje

Gibbsovo uzorkovanje je dobilo ime po Josiahu Willardu Gibbsu, američkom znanstveniku 19. stoljeća koji je izumio Gibbsova nasumična polja. Stuart i Donald Geman su prvi puta opisali postupak Gibbsovog uzorkovanja (Geman i Geman, 1984). Braća Geman bavili su se izradom modela za analizu slike. Gibbsovo uzorkovanje u njihovom radu je poseban slučaj Metropolis-Hastings algoritma (Metropolis et al., 1953). (Gelfand i Smith, 1990) su pokazali potencijalne primjene Gibbsovog uzorkovanja prilikom rješavanja velikog broja statističkih problema.

Gibbsovo uzorkovanje koristi Monte Carlo tehnike za procjenu vjerojatnosti u modelu zasnovanom na Markovljevom lancu. Prema tome, Gibbsovo uzorkovanje primjenjuje se u kompleksnim sustavima visokog stupnja entropije, gdje pretpostavljamo da iduće stanje ovisi samo o trenutnom stanju. Najčešće se koristi za izračune vrijednosti određenih integrala, posebice u višedimenzionalnim slučajevima.

Metropolis Hastings algoritam sličan je algoritmu Gibbsovog uzorkovanja. Metropolis Hastings algoritam ne donosi odluke temeljem svih uvjetnih distribucija vjerojatnosti, već donosi odluku o prihvaćanju ili odbijanju učinjenog koraka (odbijanje vodi u prethodni korak).

Ako je moguće dobiti nezavisne uzorke izravno iz distribucije, dovoljno je koristiti Monte Carlo mehanizme. Ukoliko su poznate samo uvjetne vjerojatnosti, a potrebno je uzorkovati iz zajedničke distribucije vjerojatnosti nužno je koristiti Metropolis Hastings algoritam ili Gibbsovo uzorkovanje. Gibbsovo uzkorkovanje daje uzorak za svaki korak, ali zahtjeva potpunu informaciju o uvjetnim distribucijama vjerojatnosti, što Metropolis Hastings algoritam ne zahtjeva, već odbacuje dobivene uzorke ukoliko su dobiveni na temelju izrazito niske vjerojatnosti.

Gibbsovo uzorkovanje neće uvijek konvergirati.

5. Metoda Gibbsovog uzorkovanja

Zajednička distribucija (engl. *joint distribution*) definirana je jednadžbom:

$$f(x, y_1, y_2, \ldots, y_p). \tag{5.1}$$

Potrebno je izračunati svojstva marginalne distribucije (engl. *marginal distribution*)

$$f(x) = \int \ldots \int f(x, y_1, y_2, \ldots, y_p) dy_1 dy_2 \ldots dy_p \tag{5.2}$$

kao što su srednja vrijednost (engl. *mean*) ili standardna devijacija (engl. *standard deviation*). Analitičkim izračunom integrala 5.2 dobije se $f(x)$, nakon čega je moguće izračunati željena svojstva. Analitički (ili numerički) izračun integrala može biti izuzetno složen. Gibbsovo uzorkovanje je alternativan način računjanja marginalne distribucije $f(x)$.

Gibbsovim uzorkovanjem generiraju se uzorci $X_1, ..., X_m$ $f(x)$ bez poznate funkcije $f(x)$. Generiranjem dovoljno velikog uzorka, moguće je izračunati svojstva, kao što su srednja vrijednost ili standardna devijacija, funkcije $f(x)$ s određenom preciznošću.

5.1. Dvodimenzionalni slučaj

Prvi primjer Gibbsovog uzorkovanja bit će objašnjen za dvodimenzionalni slučaj. Gibbsovim uzorkovanjem se za par slučajnih varijabli (X, Y) želi dobiti $f(x)$. Poznate su uvjetne distribucije $f(x|y)$ i $f(y|x)$. Generira se Gibbsova sekvenca:

$$Y_0^{'}, X_0^{'}, Y_1^{'}, X_1^{'}, \ldots, Y_k^{'}, X_k^{'}. \tag{5.3}$$

Postavlja je inicijalna vrijednost $Y_0^{'} = y_0^{'}$, dok se sve ostale vrijednosti generiraju prema

$$\begin{aligned} X_j^{'} &\sim f(x|Y_j^{'} = y_j^{'}) \\ Y_{j+1}^{'} &\sim f(y|X_j^{'} = x_j^{'}). \end{aligned} \tag{5.4}$$

Generiranje niza (5.3) prema formuli (5.4) naziva se **Gibbsovo uzorkovanje**. (Gelfand i Smith, 1990) su predložili generiranje m nezavisnih Gibbsovih sekvenci duljine k. Posljednje vrijednosti $X_k^{'}$ svake od m sekvenci se potom koriste za aproksimaciju $f(x)$. Ako je k dovoljno velik, uzorak $X^{'}$ je nezavisna i jednako distribuirana varijabla (engl. *independent and identically distributed*) kao i inicijalna nasumična varijabla X. Primjere dvodimenzionalnog slučaja Gibbsovog uzorkovanja pokazali su (Casella i George, 1992).

5.1.1. Primjer 1

Primjer zajedničke distribucije nasumičnih varijabli X i Y:

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$
$$x = 0, 1, \ldots, n$$
$$0 \leq y \leq 1. \tag{5.5}$$

Potrebno je izračunati svojstva marginalne distribucije $f(x)$ slučajne varijable X. Uvjetne distribucije su poznate:

$$f(x|y) = \binom{n}{k} y^k (1-y)^{n-k} \tag{5.6a}$$

$$f(y|x) = \frac{\Gamma(\alpha+n+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \tag{5.6b}$$

Generiranjem Gibbsove sekvence formulom (5.4) pomoću uvjetnih distribucija (5.6a) i (5.6b) dobivaju se $X_1, X_2, \ldots, X_m$ iz $f(x)$. Dobiveni $f(x)$ je aproksimacija pravog $f(x)$ kojeg je moguće analitički ili numerički izračunati iz zajedničke distribucije (5.5). U ovome primjeru analitičkim izračunom dobiva se da je

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}$$
$$x = 0, 1, \ldots, n. \tag{5.7}$$

Ovdje je moguće usporediti koliko je precizno Gibbsovo uzorkovanje.

5.1.2. Primjer 2

Uvjetne distribucije slučajnih varijabli X i Y su eksponencijalne distribucije

$$f(x|y) \propto ye^{-yx}, 0 < x < B < \infty$$
$$f(y|x) \propto xe^{-xy}, 0 < y < B < \infty, \tag{5.8}$$

gdje je B poznata konstanta veća od nule. Ograničenje uvjetnih distribucija na interval $(0, B)$ je dovoljan uvjet za postojanje marginalne distribucije $f(x)$.

Prosjek konačnih vrijednosti $Y_k^{'}$ i $X_k^{'}$ Gibbsovih sekvenci može poslužiti za izračun prave marginalne distribucije. Ako se generira m sekvenci Gibbsovim uzorkovanjem onda se vrijednost f(x) može aproksimirati

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x|y_i). \tag{5.9}$$

Jednadžba (5.9) je procjena gustoće. Prilikom izračuna $f(x)$ koristi se informacija o prethodnom stanju y_1, ...,y_m iz m Gibbsovih sekvenci. Procjena sadrži više informacija od procjene s vrijednostima x_1, ... x_m. Reo-Blackwell teorem sadrži dokaz (Casella i Robert, 1996).

6. Dokaz konvergencije

Potreban je dokaz da Gibbsova sekvenca (5.3) proizvodi konvergentne nizove za nasumična varijabla distribucije $f(x)$.

X i Y su nasumične varijable, sa zajedničkom raspodjelom

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix} \tag{6.1}$$

Marginalna distribucija x je

$$f_x = \begin{bmatrix} f_x(0) & f_x(1) \end{bmatrix} = \begin{bmatrix} p_1 + p_3 & p_2 + p_4. \end{bmatrix} \tag{6.2}$$

Prema tome, uvjetne distribucije $X|Y = y$ i $Y|X = x$ iznose:

$$A_{y|x} = \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix}, A_{x|y} = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix} \tag{6.3}$$

Dobivene matrice sliče matricama prijelaza karakterističnim za Markovljeve lance (Gilks et al., 1996). Generiranje Gibbsove sekvence (5.3) zahtjeva uvjetne distribucije, što je prikazano (6.3). U ovom slučaju Gibbsova sekvenca bit će niz nula i jedinica. Prema (5.4) potrebno je povezati uvjetne distribucije za dobivanje koraka Gibbsove sekvence, iz čega nastaje

$$P(X_1^{'} = x_1 | X_0^{'} = x_0) = \sum_y P(X_1^{'} | Y_1^{'} = y_1) \cdot P(Y_1^{'} = y_1 | X_0^{'} = x_0) \tag{6.4}$$

, u matričnom obliku

$$A_{x|x} = A_{y|x} A x|y. \tag{6.5}$$

Vrijedi:

$$f_k = f_0 A_{x|x}^k = (f_0 A_{x|x}^{k-1}) A_{x|x} = f_{k-1} A_{x|x} \tag{6.6}$$

Korak k Gibbsove sekvence se dobije kao $(A_{x|x}^k$. Ako su vrijednosti u $A_{x|x}$ pozitivne, onda (6.6) za bilokoju inicijalnu vjerojatnost f_0 i kada $k \to \infty$, f_k konvergiira

distribuciji f koja je stacionarna točna niza (6.6) i zadovoljava jednakost

$$fA_{x|x} = f. \tag{6.7}$$

Ako se generiranje Gibbsove sekvence zaustavi kod dovoljno velikog broja koraka k, pretpostavlja se kako je distribucija $X_k^{'}$ približno f_x.

Sve navedeno ne vrijedi samo u slučaju $2x2$, već i u općem slučaju slučajnih varijabli X i Y s n i m mogućih vrijednosti.

6.1. Matematika za slučaj dvije varijable

Dvije slučajne varijable X i Y. Poznate su uvjetne vjerojatnosti $f_{X|Y}(x|y)$ i $f_{Y|X}(y|x)$. Moguće je izračunati marginalnu distribuciju varijable X: $f_X(x)$, kao i zajedničku distribuciju X i Y preko:

$$f_X(x) = \int f_{XY}(x,y)dy, \tag{6.8}$$

gdje je $f_{XY}(x,y)$ zajednička distribucija.

$$f_{XY}(x,y) = f_{X|Y}(x|y)f_Y(y) \tag{6.9}$$

$$f_X(x) = \int f_{X|Y}(x|y)f_Y(y)dy$$

$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x|y) \int f_{Y|X}(y|t)f_X(t)dtdy \\ &= \int \left[\int f_{X|Y}(x|y)f_{Y|X}(y|t)dy \right] f_X(t)dt \\ &= \int h(x,t)f_X(t)dt, \end{aligned}$$

gdje je

$$h(x,t) = \left[\int f_{X|Y}(x|y)f_{Y|X}(y|t)dy \right]. \tag{6.10}$$

6.2. Slučaj s više od dvije varijable

U slučaju više od dvije varijable generiranje Gibbsove sekvence radi se uzorkovanje supstitucijom (engl. *substitution sampling*).

U slučaju dvije varijable uzorkovanje supstitucijom je uvijek isto.

Za tri slučajne varijable X, Y i Z potrebno je izračunati marginalnu distribuciju $f_X(x)$. Ako se Y i Z promatraju kao jedna varijabla moguće je jednadžbom

$$f_X(x) = \int \left[\int \int f_{X|YZ}(x|y,z) f_{YZ|X}(y,z|t) dy dz \right] f_X(t) dt \tag{6.11}$$

izračunati marginalnu distribuciju. Gibbsova sekvenca bi za j-ti korak bila:

$$\begin{aligned} X_j^{'} &\sim f(x|Y_j^{'} = y_j^{'}, Z_j^{'} = z_j^{'}) \\ Y_{j+1}^{'} &\sim f(y|X_j^{'} = x_j^{'}, Z_j^{'} = z_j^{'}) \\ Z_{j+1}^{'} &\sim f(z|X_j^{'} = z_j^{'}, Y_{j+1}^{'} = y_{j+1}^{'}) \end{aligned}$$

7. Praktični dio

Praktični dio napravljen je temeljem rada (Resnik i Hardisty, 2010). Resnik i Hardisty objašnjavaju ostvarenje Naivnog Bayesa (engl. *Naive Bayes*) pomoću Gibbsovog uzorkovanja na primjeru klasifikacije polariteta dokumenata prema riječima u dokumentima.

U praktičnom dijelu napravit će se sustav koji izvodi algoritam Naivnog Bayesa, a potrebne vjerojatnosti računa Gibbsovim uzorkovanjem.

7.1. Skup podataka

Besplatno dostupna biblioteka *Natural Language Toolkit, NLTK* [1] za programski jezik *Python* sadrži pripremljene korpuse teksta. Za potrebe Gibbsovog uzorkovanja korišten je skup podataka *movie_reviews* koji sadrži osvrte na filmove. Format zapisa u korpusu je $< T, S >$, gdje je T tekst osvrta, a $S \in \{'pos', 'neg'\}$ označeni polaritet osvrta. U nastavku će se za *'pos'* kritike koristiti broj 1, a za *'neg'* kritike broj 0. Pozitivne kritike označene su oznakom *'pos'*, a negative *'neg'*. Primjer negativne kritike prikazan je unutar slike 7.1.

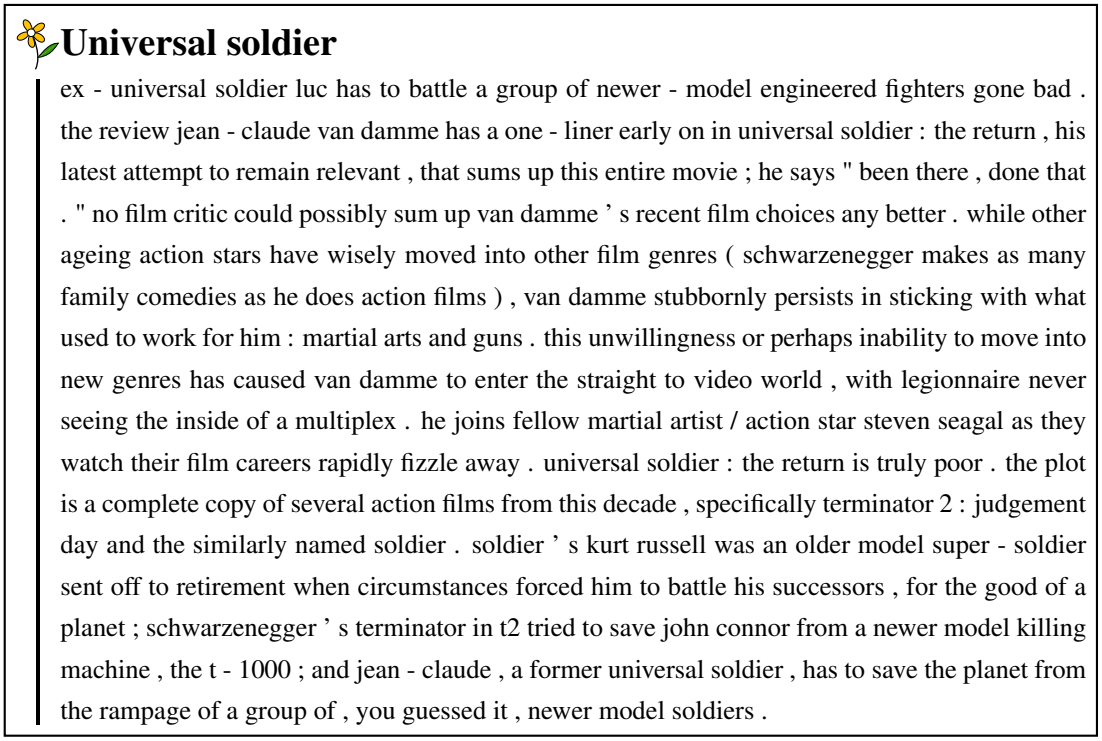
Programski jezik korišten za *Python*, verzija 2.7.3, u 64-bitnom okruženju. Dodatne *Python* biblioteke korištene prilikom izrade programa su *numpy*[2] i već spomenuti *nltk*.

7.2. Matematički model

U ovom slučaju, dokument je skup riječi koje sadrži, tzv. (engl. *bag of words*) princip. Za dokument W_j potrebno je dodijeliti adekvatan polaritet $L_j = 0$ ili $L_j = 1$. Skup dokumenata $\mathbb{C}_k$ pripada skupini klase $L_j = k$, a dobije se tako da se prebroje svi dokumenti W_j s $L_j = k$, prema tome $\mathbb{C}_k = \{W_j | L_j = k\}$. Potrebno je pronaći polaritet L_j

[1] Dostupno na `http://www.nltk.org/` .

[2] Dostupno na `http://www.numpy.org/`.



Slika 7.1: Primjer negativnog osvrta iz movie_reviews baze podataka

koji, za poznati dokument W_j, pronalazi maksimalnu vjerojatnost $P(L_j|W_J)$. Prema Bayesovom pravilu vrijedi:

$$L_j = \underset{L}{\operatorname{argmax}}\, P(L|W_j) = \underset{L}{\operatorname{argmax}}\, \frac{P(W_j|L)P(L)}{P(W_j)}. \tag{7.1}$$

Moguće je izostaviti nazivnik $P(W_j)$ jer nije ovisan o L_j. Na ovaj način nastoji se modelirati način na koji su dokumenti nastali, što se naziva generativnim modelom (engl. *Generative model*). Odabir polariteta L_j modelira se Bernoulijevom raspodjelom s parametrom π:

$$L_j \sim Bernoulli(\pi), \tag{7.2}$$

Potrebno za svaku poziciju riječi u dokumentu R_i odabrati riječ w_j temeljem distribucije vjerojatnosti riječi. Odabir distribucije vjerojatnosti iz koje se uzorkuje ovisan je dodijeljenom polaritetu dokumenta L_j. Moguće distribucije označavat će se θ_0 i θ_1. Dokument W_j gradit će se temeljem multinomijalne distribucije:

$$W_j = Multinomijalna(R_j, \theta_{L_j}). \tag{7.3}$$

Pretpostavlja se da je uzorkovanje međusobno neovisno. Distribucijama L_j i W_j nastoji se aproksimirati način na koji su dobiveni stvarni podaci.

7.2.1. Apriori parametri

Gore spomenute parametre distribucija π i θ nužno je imati prije generiranja raspodjela za W_j i L_j. Dobivanje početnih vrijednosti za π i θ će se dobiti iz jednolike raspodjele. Konkretno, π će se generirati Beta distribucijom s parametrima $\gamma_{\pi 1} = 1$ i $\gamma_{\pi 2} = 1$:

$$\pi \sim Beta(\gamma_\pi). \tag{7.4}$$

Parametri apriori vrijednosti nazivaju se hiperparametrima. Kako oba parametra Beta distribucije ovdje iznose 1, svi događaji su jednako vjerojatna, što znači da je apriori znanje o sustavu nedostupno. U slučaju θ parametra, on je modeliran Dirichlechtovom distribucijom, generaliziranom Beta distribucijom važećoj u više od dvije dimenzije:

$$\theta \sim Dirichlet(\gamma_\theta) \tag{7.5}$$

7.2.2. Zajednička distribucija

Prostor stanja u ovom problemu sastoji se skalarne varijable parametra π, dva parametarska vektora θ_1 i θ_2, oznaka klase L_N (za svaki od N dokumenata) i vektora W_J (J je broj riječi). Gibbsovo uzorkovanje kreće se kroz k-dimenzionalni prostor definiran ovim varijablama. Uvjetne distribucije definiraju matricu prijelaza između stanja, a zajednička distribucija je ciljna distribucija koju želimo aproksimirati. U našem slučaju zajednička distribucija iznosi:

$$P(\pi|\gamma_{\pi 1}\gamma_{\pi 2})P(L|\pi)P(\theta_0|\gamma_0)P(\theta_1|\gamma_1)P(\mathbb{C}_0|\theta_0, L)P(\mathbb{C}_1|\theta_1, L). \tag{7.6}$$

Postupak matematičke preobrazbe dobivene jednadžbe opisan je u (Resnik i Hardisty, 2010). Konačan rezultat je:

$$P(\mathbb{C}, L, \pi, \theta_0, \theta_1; \mu) \propto \pi^{C_1+\gamma_{\pi 1}-1}(1-\pi)^{C_0+\gamma_{\pi 0}-1} \prod_{i=1}^{V} \theta_{0,i}^{N_{\mathbb{C}_\nu}(i)+\gamma_{\theta i}-1} \theta_{1,i}^{\mathbb{C}_\nu}(i) + \gamma_{\theta i} - 1. \tag{7.7}$$

Dodatno, moguće je integrirati po varijabli π i tako reducirati zajedničku distribuciju čime se dobije konačno rješenje:

$$P(L, \mathbb{C}, \theta_0, \theta_1; \mu) \propto \frac{\Gamma(\gamma_{\pi 1} + \gamma_{\pi 0}}{\Gamma(\gamma_{\pi 1})\Gamma(\gamma_{\pi 0})} \frac{\Gamma(C_1 + \gamma_{\pi 1}\Gamma(\mathbb{C}_\nu + \gamma_{\pi 0})}{\Gamma(N + \gamma_{\pi 1} + \gamma_{\pi 0})} \prod_{i=1}^{V} \theta_{0,i}^{N_{\mathbb{C}_\nu}(i)+\gamma_{\theta i}-1} \theta_{1,i}^{\mathbb{C}_\nu}(i) + \gamma_{\theta i} - 1. \tag{7.8}$$

7.2.3. Gibbsovo uzrokovanje

Gibbsovim uzorkovanjem se u općem slučaju dodjeljuje vrijednost varijabli Z_i uzorkovanjem iz uvjetne distribucije:

$$P(Z_i|z_i^{(t+1)}, \ldots, z_{i-1}^{t+1}, \ldots, z_r^t). \tag{7.9}$$

Uzorkovanje vrijednosti u trenutku $(t+1)$ moguć je nakon dobivanja uzorka svih varijabli u trenutku t. U našem slučaju, u trenutku t poznat je broj riječi u svakom dokumentu, broj dokumenata označen s negativnim odnosno pozitivnim sentimentom, broj riječi po dokumentima označenim negativnim odnosno pozitivnim sentimentom, oznake sentimenta po dokumentu i trenutne vrijednosti hiperparametara θ_1 i θ_2. Izračun sentimenta L se dobije uzorkovanjem prema vjerojatnostima dobivenih fiksiranjem uvjetnih vjerojatnosti za $L = 0$ (negativni sentiment) i $L = 1$ (pozitivni sentiment). Prema tome, $L_i^{(t+1)}$ se računa iz uvjetne distribucije:

$$P(L_i|L_1^{t+1}, \ldots, L_{i-1}^{t+1}, \ldots, L_{i+1}^t t, L_N^t, \mathbb{C}, \theta_0^t, \theta_1^t; \mu). \tag{7.10}$$

Na sukladan način se dobiju i vrijednosti θ_0 i θ_1:

$$P(\theta_0^{(t+1)}|L_1^{(t+1)}, \ldots, L_N^{(t+1)}, \mathbb{C}, \theta_1^{(t)}; \mu) \tag{7.11}$$

$$P(\theta_1^{(t+1)}|L_1^{(t+1)}, \ldots, L_N^{(t+1)}, \mathbb{C}, \theta_0^{(t+1)}; \mu) \tag{7.12}$$

7.3. Programska implementacija

7.3.1. Priprema podataka

Prilikom programske implementacije korištene su brojne biblioteke spomenute u 7.1. Uvoz biblioteka napravljen je kao na slici 7.2. U skupu podataka *movie_reviews* ima 2000 osvrta i približno 32000 različitih riječi. Zbog pojednostavljena uzorkovanja i smanjivanja prostora stanja napravljen je Naivni Bayesov klasifikator te su njime izvučene najveće vjerojatnosti određene oznake za $P(L|w)$, gdje je L oznaka sentimenta, a w riječ. Korišteno je 500 od 32000 mogućih riječi. Deset riječi s tog popisa prikazano je slikom 7.3. Primjerice, korištenje riječi seagal (glumac) u kritici znači da će vjerojatnost da je kritika biti negativna je 13.2 veća od vjerojatnosti pozitivne vjerojatnosti. Izračun tih riječi prikazan je programskim kodom na slici 7.4. U konačnici gradi se matrica frekvencije riječi i dokumenata (kod na slici 7.5). Element x na indeksu i, j govori da se riječ j pojavljuje u dokumentu i x puta. Matrica je prikazana primjerom 7.6.

```
from nltk.corpus import movie_reviews
import nltk
import random
from numpy import array, empty, ones, nonzero, repeat, zeros
import numpy
```

Slika 7.2: Uvoz bibilioteka

```
      contains(seagal) = True      neg : pos    =     13.2 : 1.0
contains(outstanding) = True      pos : neg    =     10.7 : 1.0
       contains(mulan) = True      pos : neg    =      8.9 : 1.0
       contains(damon) = True      pos : neg    =      7.7 : 1.0
contains(wonderfully) = True      pos : neg    =      7.6 : 1.0
      contains(wasted) = True      neg : pos    =      5.9 : 1.0
       contains(awful) = True      neg : pos    =      5.4 : 1.0
       contains(waste) = True      neg : pos    =      5.4 : 1.0
        contains(lame) = True      neg : pos    =      5.3 : 1.0
       contains(flynt) = True      pos : neg    =      4.9 : 1.0
```

Slika 7.3: Riječi s najvišim $P(L|w)$

7.3.2. Inicijalizacija

Matrica frekvencija riječi po dokumentima i pripadajuće oznake sentimenta predstavljaju ulazni skup podataka programu Gibbsovog uzorkovanja. Idući nužni korak je dobivanje inicijalnih vrijednosti hiperparametara π i vektora θ. Zahvaljujući *Pythonovim* paketima uzorkovanje iz Dirichletove distribucije moguće je provesti iznimno jednostavno. Temeljem hiperparametara uzorkuju se parametri nužni za Gibbsovo uzorkovanje. Inicijalizacija je prikazana slikom 7.7. Gibbsovo uzorkovanje dodatno je parametrizirano ukupnim brojem iteracija, brojem iteracija odbacivanja, udaljenosti između uzoraka, objašnjenih u poglavlju **??**.

```
def document_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains(%s)' % word] = (word in document_words)
    return features

documents = [(list(movie_reviews.words(fileid)), category)
                for category in movie_reviews.categories()
                for fileid in movie_reviews.fileids(category)]
random.shuffle(documents)

all_words = nltk.FreqDist(w.lower() for w in movie_reviews.words())
word_features = all_words.keys()[:2000]

featuresets = [(document_features(d), c) for (d,c) in documents]
train_set, test_set = featuresets[100:], featuresets[:100]
classifier = nltk.NaiveBayesClassifier.train(train_set)

vocabulary_size = 500

most_informative = classifier.most_informative_features(vocabulary_size)
```

Slika 7.4: Algoritam naivnog Bayesa koji računa $P(L|w)$

```
document_num = 10
i = 0
true_labels = empty(document_num, int)
corpus = empty((document_num, vocabulary_size-1), int)

while i < document_num:
        if documents[i][1] == 'pos':
                true_labels[i] = 1
        else:
                true_labels[i] = 0

        for w in documents[i][0]:
                if w.lower() in freq_dict:
                        freq_dict[w.lower()]+=1

        print len(freq_dict)
        j = 0
        for key in sorted(freq_dict):
                corpus[i][j] = freq_dict[key]
                j+=1

        freq_dict = dict.fromkeys(vocabulary_words,0)
        i+=1
```

Slika 7.5: Izračun matrice frekvencija riječi i dokumenata

```
[[56  0  1 ...,  1  0  0]
 [12  0  0 ...,  1  0  0]
 [30  0  0 ...,  0  0  0]
 ...,
 [18  0  0 ...,  1  0  0]
 [75  0  1 ...,  0  0  0]
 [25  0  0 ...,  0  0  0]]
```

Slika 7.6: Primjer matrice frekvencije riječi i dokumenata

```
categories = 2          # broj kategorija L
vocabulary = 499        # broj rijeci V
documents = 10          # broj dokumenata W
hyp_pi = ones(categories, int)
hyp_thetas = ones((categories, vocabulary), int)
pi = log(dirichlet(hyp_pi, 1)[0])
thetas = dirichlet(hyp_thetas, categories)
```

Slika 7.7: Inicijalizacija hiperparametara

8. Zaključak

Zaključak.

9. Literatura

George Casella i Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

George Casella i Christian P Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

Alan E Gelfand i Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Stuart Geman i Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Walter R Gilks, Sylvia Richardson, i David J Spiegelhalter. *Markov chain Monte Carlo in practice*, svezak 2. CRC press, 1996.

Robert E Kass, Bradley P Carlin, Andrew Gelman, i Radford M Neal. Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician*, 52(2): 93–100, 1998.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, i Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

Philip Resnik i Eric Hardisty. Gibbs sampling for the uninitiated. Technical report, DTIC Document, 2010.

Reini Wirahadikusumah i Dulcy M Abraham. Application of dynamic programming and simulation for sewer management. *Engineering, Construction and Architectural Management*, 10(3):193–208, 2003.