Naivni Bajes

STATISTIČKA KLASIFIKACIJA

Seminarski rad - MOAR

Student: Rastko Lazarević 2022/0247

Mentori: prof. dr Tatjana Lutovac
prof. dr Bojana Mihailović

Statističko učenje

- **Problemi u kojima vlada visok stepen neizvesnosti.**
- **Različiti tipovi neivesnosti i subjektivnosti.**
- **Standardni logički zakoni nisu dovoljni.**
- **Probabalistički pristup za modelovanje neizvesnosti.**
- **Bajesove mreže.**
- **Kvantifikacija sigurnosti (intervali poverenja, varijansa).**

- **Pogled u budućnost.**
- **Donošenje odluke.**

Statističko učenje

Primene:

- **Decision-support sistemi - agent koji donosi odluku.**
  - Modelovanje rezonovanja. Konačan odgovor sistema ili sugestija.
- **Sistemi za detekciju – statičke ili dinamičke prirode**
  - Sistemi za detekciju otkaza.
  - Detekcija različitih obrazaca (tekst itd.)
- **Klasifikatorski sistemi – najčešće statičke prirode,**
  može i dinamička
  - Tekstualni podaci, fotografije, sekvence signala.
- **Prediktivni sistemi – najčešće dinamičke prirode,**
  može i statičke
  - Predikcija budućih stanja sistema (kvarovi).

Statističko učenje

Primene i česti modeli realizacije:

- **Klasifikatorski sistemi**
  - **Naivni Bajes**
- **Prediktivni sistemi i sistemi za detekciju**
  - **HMM – Skriveni Markovljevi lanci**
- **Decision support sistemi**
  - **Različiti tipovi Bajesovih mreža**

  (zavisno od konkretnog problema)

Zavisno od literature, pomenuti modeli se mogu različito okarakterisati. Često se svi smatraju Bajesovim mrežama, ali zbog specifičnosti problema koji rešavaju i ustaljenosti u konkretnim primenama, u nekoj literaturi su okarakterisani kao posebna kategorija.

Verovatnoća i statistika

- U pozadini svih ovih modela koji su naizgled jednostavni i intuitivni, leži strogo fundirana Teorija verovatnoće i statistike.
- U kontekstu Bajesovih mreža iznosimo važne elemente ove teorije:

- Slučajne promenljive: $A = \{a_1, \dots, a_n\}$
- Fundamentalna teorema za slučajne promenljive (uslovna verovatnoća):

$$P(A, C) = P(A|C)P(C)$$

- (Uslovna) Nezavisnost slučajnih promenljivih:

$$P(A \mid C) = P(A) \qquad P(A, C) = P(A)\,P(C) \qquad P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

- Bajesova formula:

$$P(A \mid C) = \frac{P(C|A)P(A)}{P(C)}$$

- Apriorna i aposteriorna verovatnoća: $P(A), P(A|\,C)$
- Marginalizacija:

$$P(A) = \sum_C P(A, C) = \sum_C P(A|C)P(C)$$

Bajesove mreže

- **Vizuelizujemo tok našeg rezonovnja pomoću grafovske strukture (DAG).**
- **Čvorovi – slučajne promenljive**
- **Usmerene grane – tok rezonovanja i prenosa informacija**

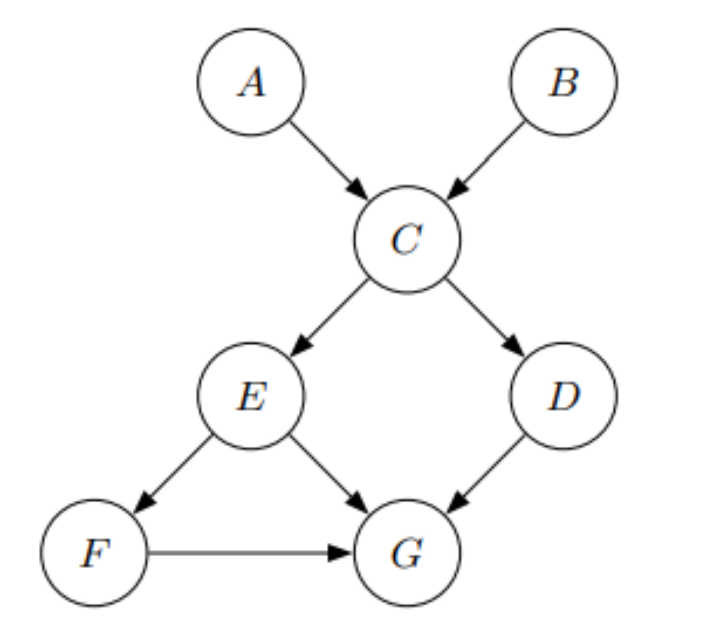
Dobra praksa je da veze (smer usmerene grane grafa) reprezentuju kauzalan uticaj među promenljivim, mada to ne mora biti slučaj. Ukoliko veze modelujemo kauzalnim uticajima, postoji teorija koja nam daje formalizam provere i validacije, da li naša mreža zaista odgovara fenomenu (fizici) koju želimo da modeliramo. Dati formalizam je deo teorije kauzalnih mreža i daje nam mogućnost provere (ne)zavisnosti i protoka informacija među sluačajnim promenljivim mreže *(d-razdvojenost)*.

Formalizam kauzalnih mreža i *d-razdvojenosti* bi se, intuitivno govoreći, mogao okarakterisati kao pandan teoriji uslovne verovatnoće za analizu zavisnosti između pojedinih događaja, pojava i objekata, ali ovde u grafovskim strukturama kojim modelujemo naše rezonovanje.

Bajesove mreže

$$P(U) = \prod_{A_i \in U} P(A_i \mid p(A_i))$$

- **Pravilo lanca za Bajesove mreže, gde U *univerzum slučajnih promenljivih mreže i* $p(A_i)$ skup roditeljskih promenljivih promenljive A_i.**
- **Prostorna optimizacija.**
- **Koenzistentnost.**



Naivni Bajes

H – **promenljiva hipoteza** **(njenu raspodelu želimo)**
$I = \{I_1, \dots, I_n\}$ – **informacione promenljive** **(opservabilne)**

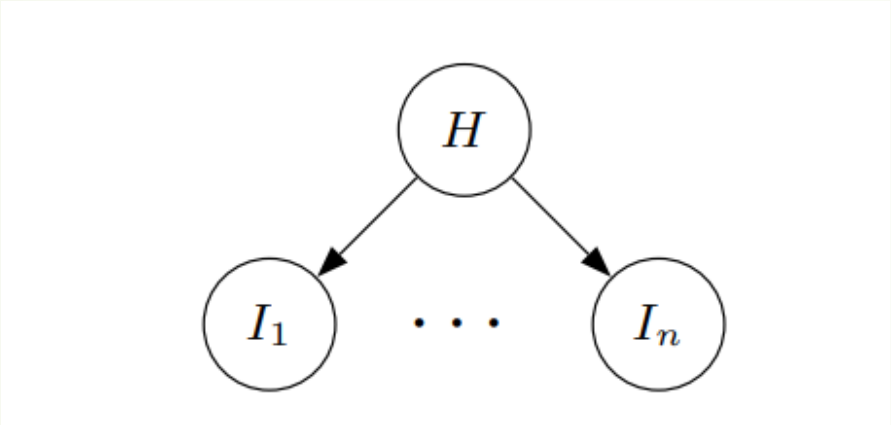
- **Pretpostavka o uslovnoj nezavisnosti:** $P(I_1, \dots, I_n \mid H) = \prod_{i=1}^{n} P(I_i \mid H)$

$$P(H \mid I_1, \dots, I_n) = \frac{P(I_1, \dots, I_n \mid H)P(H)}{P(I_1, \dots, I_n)}$$

$$P(H \mid I_1, \dots, I_n) = \frac{1}{P(I_1, \dots, I_n)} P(H) \prod_{i=1}^{n} P(I_i \mid H)$$

Naivni Bajes

H – **promenljiva hipoteza** **(njenu raspodelu želimo)**
$I = \{I_1, \dots, I_n\}$ – **informacione promenljive** **(opservabilne)**



- $\sum_H P(U)$ - **ovakav proračun je teško sprovesti.**
- **Vidimo da mreža zaista ispoljava svojstva Naivnog Bajesa o uslovnoj nezavisnosti.**

$$P(U) = P(H)P(I_1 | H) \dots P(I_n | H)$$

$$P(H | I_1, \dots, I_n) = \frac{P(U)}{\sum_H P(U)}$$

$$P(H | I_1, \dots, I_n) = \frac{1}{P(I_1, \dots, I_n)} P(H) \prod_{i=1}^{n} P(I_i | H)$$

Naivni Bajes - klasifikator

$H = K = \{k_1, \dots, k_m\}$ – **Skup klasa**
$I = X = \{I_1(x_1), \dots, I_n(x_n)\}$ – **Skup obeležja**
$X_{obs} = \{x_1, \dots, x_n\}$ – **Konkretna opservacija**

- **Definicija klasifikatora:**

$$\boldsymbol{k_{out} = argmax_i \left(P(K = k_i \mid X)\right)}$$

- **Koraci projektovanja i klasifkacije:**

1. **Ocena apriorne verovatnoće** $P(K)$.
2. **Ocena modela za svaku klasu, tj. računamo** $P(X \mid K)$.
3. **Računanje verovatnoće** $P(X_{obs} \mid K)$. **Računamo da nova opservacija** X_{obs} **pripada svakoj klasi tj.** $P(X_{obs} \mid K = k_i)$.
4. **Određivanje maksimalne verovatnoće** $P(K = k_i \mid X_{obs})$. **Primena Bajesove formule.**

Prednosti i mane

- **Visoka brzina obuke**, čak i sa manjim trenirajućim skupom. **Štedimo procesorsku moć.**
- Dobri rezultati i kada su **podaci visoke dimenzionalnosti.**
- Robusnost u prisustvu **nedostajućih podataka.**
- **Laka implementacija i interpretacija.**

- **Pretpostavka uslovne nezavisnosti je nekada nerealna.**
- Za **klase koje su opisane sličnim parametrima** (za dati skup obeležja) poput varijanse i srednje vrednosti, NB klasifikator neće uopšte raditi.
- **Problem nultih verovatnoća.**
- **Laplasovo izglađivanje.**

* Bajesove mreže daju mogućnost ukrštanja ekspertskog znanja i dostupnih podataka.
* Sa druge strane, donošenje odluka pomoću fazi logike se primarno zasniva na ekspertskom znanju.
* Složeni modeli, poput neuralnih mreža, se primarno oslanjaju na podatke.

Naivni Bajesov klasifikator - primer

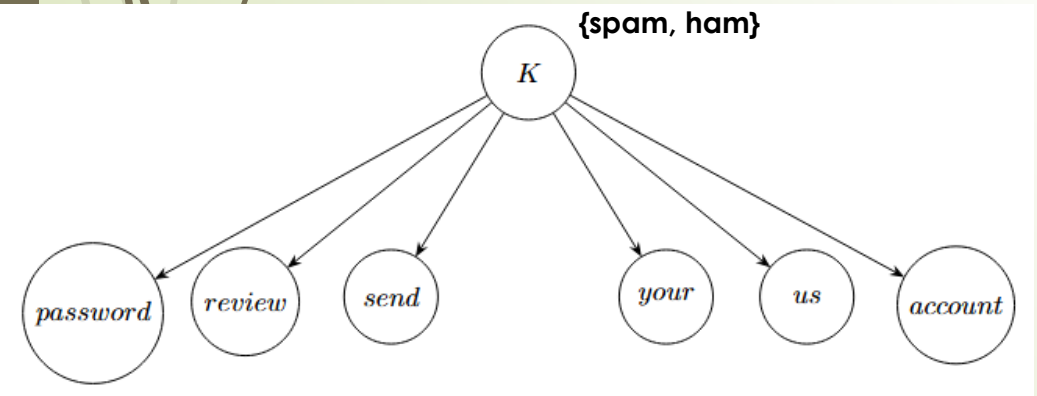
Naslov mejla	Vrsta
send us your password	spam
send us your account review	ham
review your password	ham
review us	spam
send your password	spam
send us your account	spam

Tabela 1: Obučavajući skup

spam	ham	reči
2/4	1/2	password
1/4	2/2	review
3/4	1/2	send
3/4	1/2	us
3/4	1/2	your
1/4	1/2	account

Definišemo slučajnu promenljivu $K = \{spam,\ ham\}$, koja predstavlja naše klase. Lako se računa da su apriorne verovatnoće date sa $P(K = spam) = \frac{4}{6}$ i $P(K = ham) = \frac{2}{6}$.

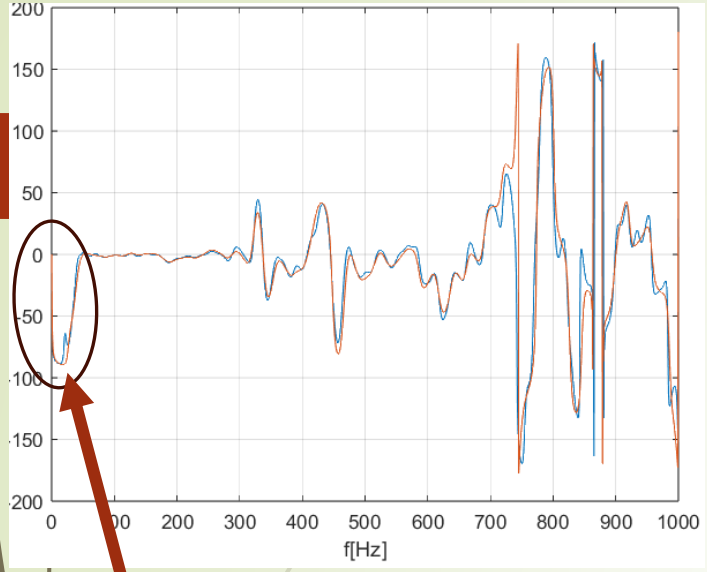
Normalizacijom vrednosti iz tablice dolazimo do uslovnih raspodela, precizcnije njihovih estimacija: $P(password \mid K) = (\frac{2}{4},\ \frac{1}{2})$, $P(review \mid K) = (\frac{1}{5},\ \frac{4}{5})$, $P(send \mid K) = (\frac{3}{5},\ \frac{2}{5})$, $P(us \mid K) = (\frac{3}{5},\ \frac{2}{5})$, $P(your \mid K) = (\frac{3}{5},\ \frac{2}{5})$, $P(account \mid K) = (\frac{1}{3},\ \frac{2}{3})$. Ovde važi $P(\text{reč} \mid K) = (P(\text{reč} \mid K = spam),\ P(\text{reč} \mid K = ham))$.
Konačno, mreža koja odgovara našem modelu je data na **Slici 4.**

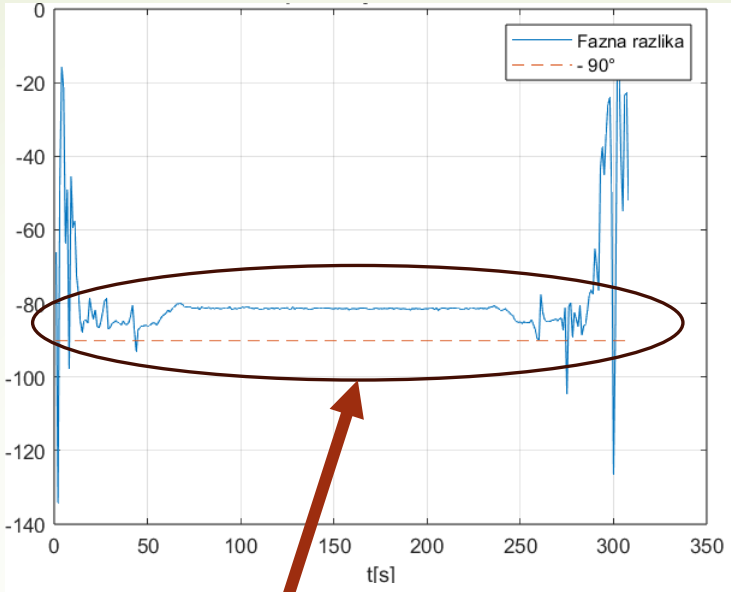


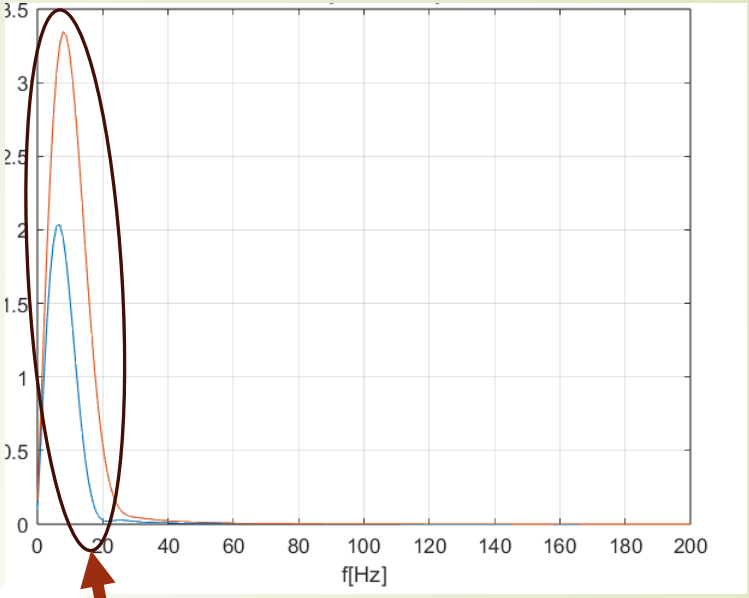
$P(review\ password \mid K = spam) = P(110000 \mid spam) = \frac{2}{4}\frac{1}{5}(1-\frac{3}{5})(1-\frac{3}{5})(1-\frac{3}{5})(1-\frac{1}{3}) = 0.00427$, gde nam 1 i 0 opisuje da li se reči iz novog naslova pripadaju skupu "korisnih"reči. Sličnim rezonovanjem $P(review\ password \mid K = ham) = P(110000 \mid ham) = \frac{1}{2}\frac{4}{5}(1-\frac{2}{5})(1-\frac{2}{5})(1-\frac{2}{5})(1-\frac{2}{3}) = 0.0288$.

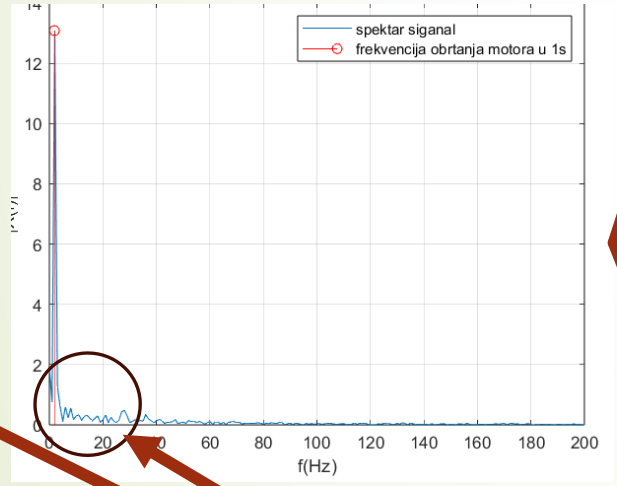
$P(review\ password \mid K = spam)\ P(K = spam) = 0.00427 \cdot \frac{4}{6} = 0,00285.$
$P(review\ password \mid K = ham)\ P(K = ham) = 0.0288 \cdot \frac{2}{6} = 0.0096.$

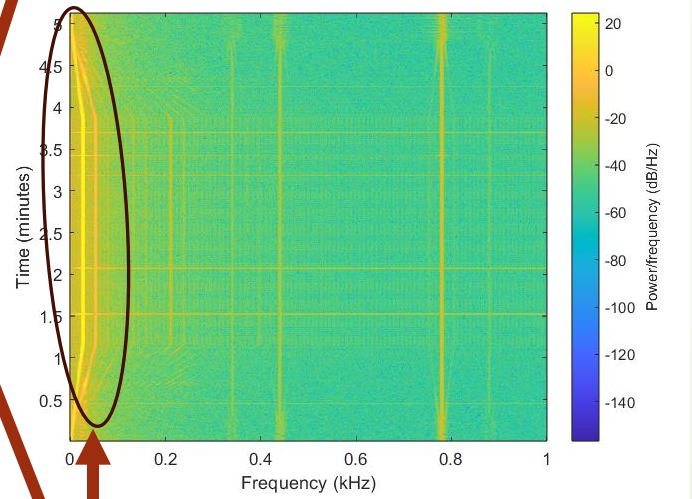
Odatle lako dolazimo da važi $P(K = spam \mid review\ password) = 0.23$ i $P(K = ham \mid review\ password) = 0.73.$ Dakle, odluka našeg klasifikatora je da je mejl sa naslovom **"review password"** ham.









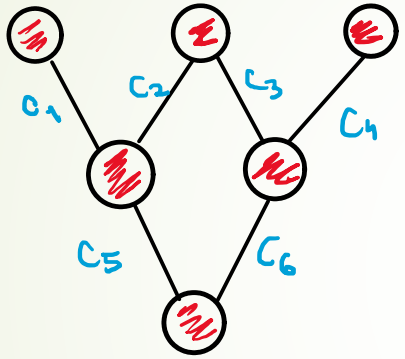


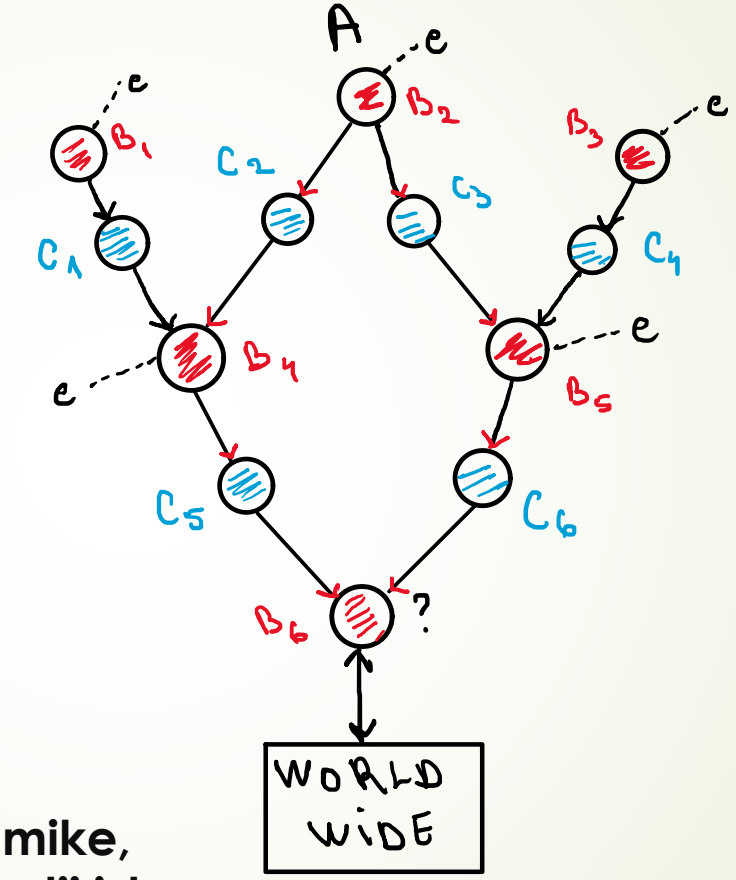
Fault

-Visoka dinamika.
-Rad u realnom vremenu.
- Klasifikacija?
- Detekcija?
- Predikcija?

{balansiranost, savijanje vratila, neusklađenost osa, …}

Bajesove mreže u mrežnim sistemima





Za osnovni model je u redu.
Internet mreže su sistemi visoke dinamike,
Pa klasične Bajesove mreže nisu najbolji izbor.
Bolja opcija su stohastički modeli koji se bolje nose sa dinamikom.

$$q_i = \frac{trenutna\ opterećenost}{kapacitet}$$

$$p_i = 1 - q_i$$

- Optimizacija mrežne opterećenosti.

- Biparitni graf.

K R A J