Kako odrediti $P(x|y)$?

- $P(x|y)$ - koliko je verovatno da uočimo kombinaciju vrednosti atributa *x* ako taj primer pripada klasi *y*?
- Razmotrimo skup podataka sa 16 binarnih atributa gde je *y* takođe binarno. Koliko nam primera treba da bismo u potpunosti odredili $P(x|y)$?

Congressional Voting Records Data Set

Podaci o glasanju US kongresmena o 16 ključnih pitanja uz klasnu oznaku da li je kongresmen republikanac ili demokrata

1. Class Name: 2 (democrat, republican)
2. handicapped-infants: 2 (y,n)
3. water-project-cost-sharing: 2 (y,n)
4. adoption-of-the-budget-resolution: 2 (y,n)
5. physician-fee-freeze: 2 (y,n)
6. el-salvador-aid: 2 (y,n)
7. religious-groups-in-schools: 2 (y,n)
8. anti-satellite-test-ban: 2 (y,n)
9. aid-to-nicaraguan-contras: 2 (y,n)
10. mx-missile: 2 (y,n)
11. immigration: 2 (y,n)
12. synfuels-corporation-cutback: 2 (y,n)
13. education-spending: 2 (y,n)
14. superfund-right-to-sue: 2 (y,n)
15. crime: 2 (y,n)
16. duty-free-exports: 2 (y,n)
17. export-administration-act-south-africa: 2 (y,n)

Kako odrediti $P(x|y)$?

- Ukoliko se x sastoji od *n* binarnih atributa treba da estimiramo parametre:

$$\theta_{i,j} = P\left(x = x_i | y = y_j\right)$$

  gde x_i uzima 2^n mogućih vrednosti, a y_j 2 moguće vrednosti
- Broj nezavisnih parametara je $2(2^n - 1)$ (jer za bilo koje y_j mora da važi $\sum_i P\left(x = x_i | y = y_j\right) = 1$)
- Da bismo dobili pouzdanu ML ocenu, svaku kombinaciju moramo uočiti više puta
  - oko stotinu nezavisno izvučenih primera da bismo dobili ocenu koja je u nekoliko procenata od svoje stvarne vrednosti
- Određivanje vrednosti pune uslovne verovatnoće bi zahtevao ogromnu količinu podataka!
  - Npr. da je *x* vektor od 16 binarnih varijabli trebalo bi da estimiramo više od milion nezavisnih parametara
- A u opštem slučaju *x* i *y* ne moraju biti binarni

Naïve Bayes

- **Naivni** Bajes:
  - uvešćemo pretpostavku da su obeležja skupa podataka uslovno nezavisna za zadatu klasu:

$$P(x|y) = \prod_{d=1}^{D} P(x_d|y)$$

Specifičan model za obeležje *d*

- Klasifikaciono pravilo postaje:

$$\mathrm{h}_{MAP} = \arg\max_c P(x|y=c)P(y=c) = \arg\max_c \prod_{d=1}^{D} P(x_d|y=c)\,P(y=c)$$

- Naivna pretpostavka dovodi do dramatične redukcije broja nezavisnih parametara koje moramo estimirati
  - Npr., u primeru gde se se *x* sastoji od *n* binarnih varijabli i *y* je binarno, uz naivnu pretpostavku treba da estimiramo svega $2n$ parametara (nasuprot $2(2^n - 1)$ koliko bismo imali bez ove pretpostavke)

Naivna pretpostavka

- Ovo je slabija pretpostavka od nezavisnosti atributa

- Primer:
  - 3 binarne slučajne varijable – kiša (*K*), grmljavina (*G*) i Munja (*M*)
  - Razumna pretpostavka: grmljavina je nezavisna od kiše pod uslovom munje, tj. $P(G|K,M) = P(G|M)$
    - Znamo da munja uzrokuje grmljavinu, pa jednom kada znamo vrednost *M* (desila se ili ne), vrednost *K* nam ne pruža nikakve dodatne informacije o vrednosti *G*
    - Naravno, postoji jasna zavisnost vrednosti *G* od vrednosti *K*
    - Ali ne postoji *uslovna* zavisnost jednom kada znamo vrednost *M*

Naïve Bayes

- „Individualne“ verodostojnosti svakog obeležja možemo odrediti putem ML ocene, što je, u slučaju diskretnih vrednosti obeležja, prosto frekvencija pojave:

$$P(x_d|y=c) = \frac{N_{x_d,c}}{N_c}$$

Broj pojave obeležja x_d u primerima iz klase *c*

Broj primera iz klase *c*

- Primer: kategorizacija emailova na klase *spam*/*ham*
  - Imamo kolekciju od 500 emailova, od čega je 100 *spam*, a 400 *ham*
  - Imamo email „hello world“. Među *spam* emailovima, reč „hello“ se pojavljuje 20 puta, a reč „world“ 2 puta

$$P(x = [\text{hello}, \text{world}]|y = \text{spam})$$
$$= P(\text{hello}|y = \text{spam})\text{P}(\text{world}|y = \text{spam}) = \frac{20}{100} \cdot \frac{2}{100} = 0.004$$

Naïve Bayes

- Naivna pretpostavka:
  - Pretpostavili smo da je pojava reči *hello* nezavisna od pojave reči *world*
  - Šta je sa rečima *peanut*, *butter*, *alergy*? Intuicija nam govori da tekst koji sadrži reč *peanut* verovatnije sadrži reči *butter* ili *alegy* od nekih drugih reči – naivna pretpostavka je narušena

- U praksi, uvedena naivna pretpostavka je zaista često narušena, ali pokazano je da NB i u tim slučajevima može da ima dobre performanse

  [H. Zhang, "The optimality of naive bayes," AA, vol. 1, no. 2, p. 3, 2004.]

- Najveća prednost NB jeste što zahteva relativno malo podataka za estimaciju parametara
- Pretpostavka o uslovnoj nezavisnosti obeležja za datu klasu znači da parametre za svako obeležje možemo naučiti nezavisno