$$P(y = c|x) = \frac{P(x|y = c)P(y = c)}{P(x)}$$

Naivni Bajesov model

Kontinualna obeležja

Kontinualne varijable – opcije

1. Diskretizacija
   - Pretvorićemo kontinualnu varijablu u kategoričku tako što ćemo njen opseg podeliti na intervale (npr. 0-10, 10-20,...)
   - Kako odabrati granice intervala?

2. Gausov kernel

Gausov kernel

- Pretpostavka: x_d prati normalnu (Gausovu) raspodelu:

$$P(x_d|y=c) = \frac{1}{\sqrt{2\pi\sigma_{d,c}^2}} \exp\left\{-\frac{\left(x_d - \mu_{d,c}\right)^2}{2\sigma_{d,c}^2}\right\}$$

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

Gausov kernel – treniranje modela

- Za svako x_d i svaku klasu c treba proceniti parametre μ_{dc}, σ_{dc} na osnovu trening podataka
  - Ukupno $2 \cdot D \cdot C$ nezavisnih parametara
  - Možemo uvesti pretpostavke da smanjimo ovaj broj parametara – ako verujemo da šum u opservacijam x_d dolazi iz istog izvora možemo pretpostaviti da su svi σ_{dc} identični

- ML ocena iz skupa podataka:

$$\hat{\mu}_{dc} = \frac{1}{\sum_n \mathbb{I}(y^{(n)} = c)} \sum_n x_d^{(n)} \mathbb{I}(y^{(n)} = c) \qquad \mathbb{I}(y^{(n)} = c) \begin{cases} 1, & y^{(n)} = c \\ 0, & y^{(n)} \neq c \end{cases}$$

$$\hat{\sigma}_{dc}^2 = \frac{1}{(\sum_n \mathbb{I}(y^{(n)} = c)) - 1} \sum_n \left(x_d^{(n)} - \hat{\mu}_{dc} \right)^2 \mathbb{I}(y^{(n)} = c)$$

Gausov kernel primer – golf dataset

Temperature	Humidity	Play
85	85	no
80	90	no
65	70	no
72	95	no
71	80	no
83	78	yes
70	96	yes
68	80	yes
64	65	yes
69	70	yes
75	80	yes
75	70	yes
72	90	yes
81	75	yes

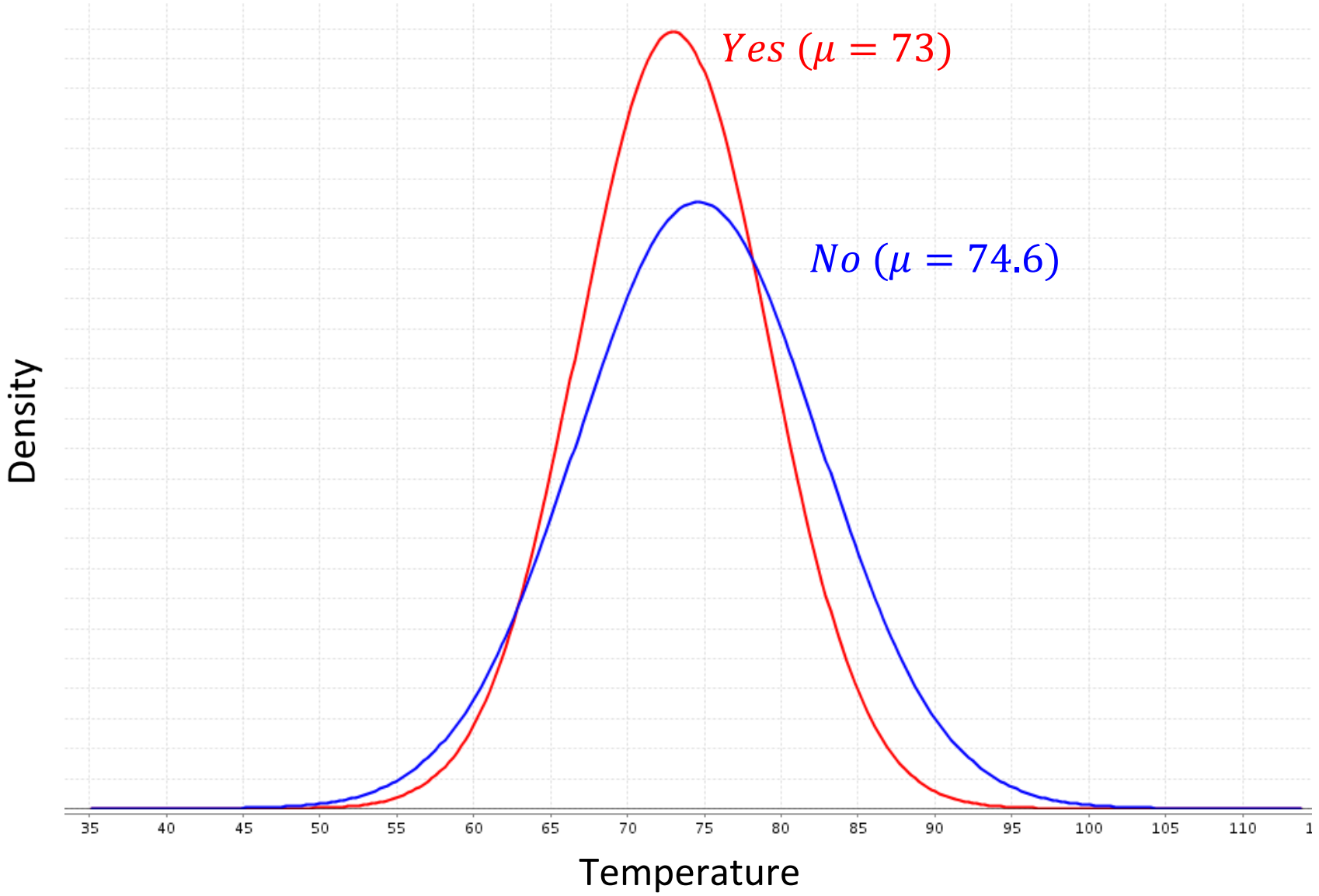
$$P(Temperature = 77|yes) = ?$$

$$\mu_{T,yes} = \frac{83 + 70 + 68 + 64 + 69 + 75 + 75 + 72 + 81}{9} = 73$$

$$\sigma^2_{T,yes} = \frac{1}{9-1}\left((83-73)^2 + \cdots + (81-73)^2\right) = 38$$

$$P(T = 77|yes) = \frac{1}{\sqrt{2\pi \cdot 38}} \exp\left\{-\frac{(77-73)^2}{2 \cdot 38}\right\} = 0.0524$$

Gausov kernel primer – golf dataset



Naïve Bayes u praksi

- U slučaju kontinualnih varijabli x_d, NB sa Gausovim kernelom će imati bolje performanse ako su raspodele x_d bliske Gausovoj raspodeli

- Ukloniti outlier-e (npr. vrednosti koje su 3 ili 4 σ udaljene od srednje vrednosti)

- Alternativno, koristiti druge raspodele koje bolje odgovaraju podacima

Naïve Bayes u praksi

- Verovatnoće $P(x|y)$ su obično mali brojevi. Ako množimo mnogo malih brojeva, može doći do numeričkog underflow-a
  - Logaritmovati verovatnoće $P(x|y)$
  - Ovo ne menja predikcije jer je bitno samo koja klasa ima veću verovatnoću (nisu važne tačne vrednosti pojedinačnih verovatnoća)

- Ukloniti (jedno od) obeležja koja su u snažnoj korelaciji (naivna pretpostavka)

- NB model je *updatable* – čim novi podaci postanu dostupni, moguće je ažurirati verovatnoće u modelu

Kada primeniti Naïve Bayes?

- Veoma brz i nezahtevan u pogledu skladištenja
- Robustan na irelevantna obeležja (imaju iste raspodele u svim klasama)
- Robustan na šum u podacima
- Ne zavisi od vrste atributa (kategorički ili kontinualni)
- Lako se ažurira kako pristižu podaci

Kada primeniti Naïve Bayes?

- Posebno se istakao u domenima klasifikacije teksta i dijagnoza bolesti

- Primeri primene:
  - Dijagnoza bolesti i donošenje odluka oko terapije

    [J. Kazmierska and J. Malicki, “Application of the na ̈ıve bayesian classifier to optimize treatment deci- sions,” Radiotherapy and Oncology, vol. 86, no. 2, pp. 211–216, 2008.]

  - Klasifikacija RNA sekvenci

    [Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, “Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy,” Applied and environmental microbiology, vol. 73, no. 16, pp. 5261–5267, 2007.]

  - Spam filtering

    [M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A bayesian approach to filtering junk e-mail,” in Learning for Text Categorization: Papers from the 1998 workshop, vol. 62, pp. 98–105, 1998.]

Kada ne treba primenjivati Naïve Bayes?

- Prilikom odabira klasifikacionog modela uvek moramo imati na umu tip podataka i tip problema

- NB je linearan klasifikator koji uvodi pretpostavku o uslovnoj nezavisnosti obeležja
  - U praksi, ova pretpostavka je često narušena, ali, i pored toga, NB često ima veoma dobre performanse

    [I. Rish, "An empirical study of the naive bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, pp. 41–46, 2001.]

- Međutim, NB može da rezultuje lošim performansama u slučaju:
  - Snažnog narušavanja pretpostavke o nezavisnosti obeležja
  - Ne-linearnih klasifikacionih problema

- U praksi se uvek preporučuje da se na konkretnom skupu podataka isproba i uporedi više klasifikacionih modela i da se uzmu u obzir predikcione performanse, ali i računska efikasnost

Dodatno čitanje

- Tom Mitchell, McGraw Hill: „Machine Learning“

  Chapter 3: Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression
  http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf