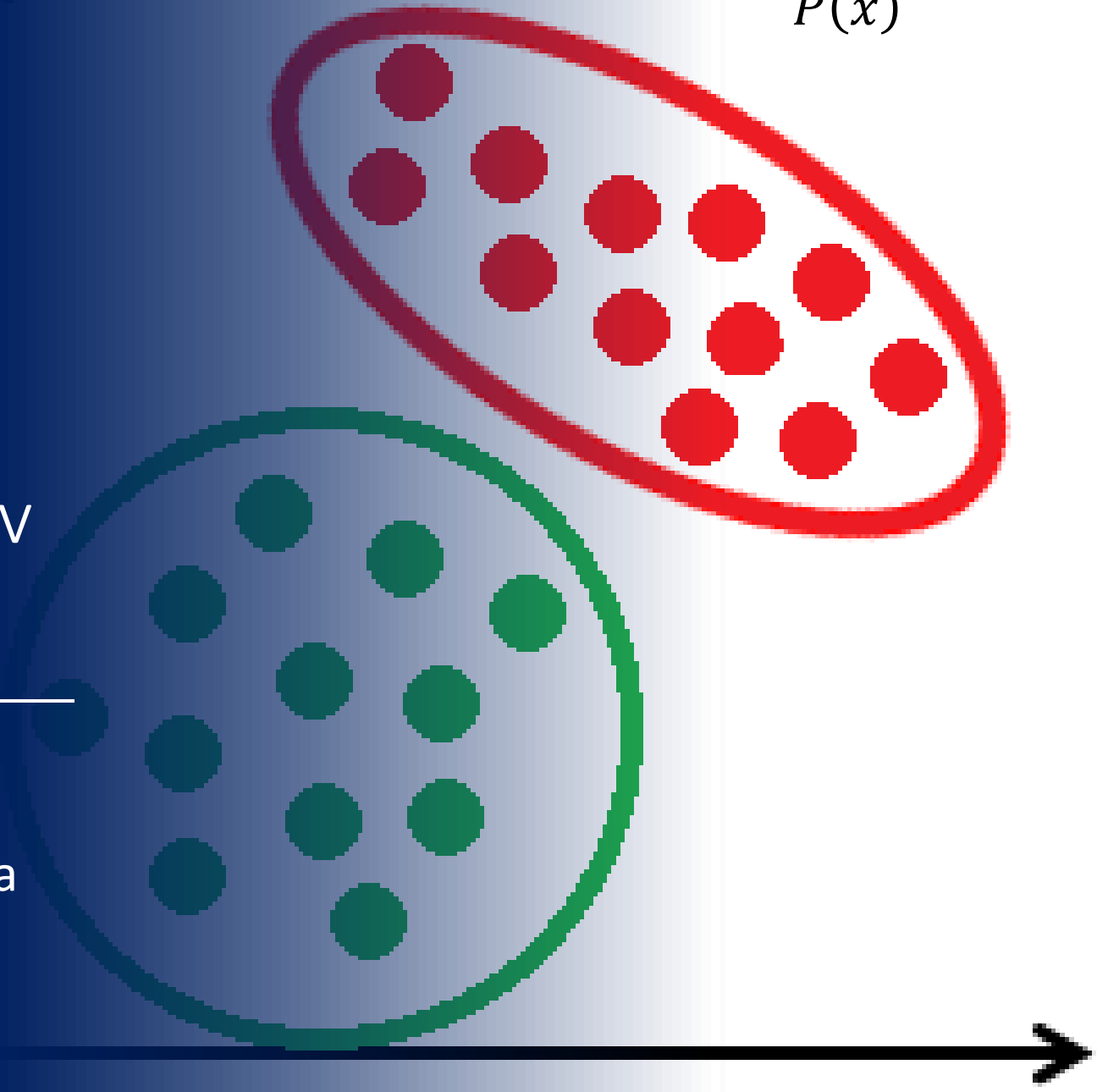
$$P(y = c|x) = \frac{P(x|y = c)P(y = c)}{P(x)}$$

Naivni Bajesov klasifikator

- Naïve Bayes
- Tipovi klasfikatora



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

Naivni Bajesov model

Diskretna obeležja

Naïve Bayes

- Spada u grupu generativnih modela
- Jednostavan linearni klasifikator
- Probabilistički model zasnovan na Bajesovoj teoremi

Bajesovo pravilo odluke

Uslovna verovatnoća

Apriorna verovatnoća

Aposteriorna verovatnoća

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

Dokaz

- Interpretacija u kontekstu klasifikacionog problema:
  - Aposteriorna verovatnoća: koja je verovatnoća da konkretan primer pripada klasi, na osnovu uočenih vrednosti obeležja?
  - Npr., koja je verovatnoća da osoba ima dijabetes ukoliko smo izmerili da ima određene vrednosti glukoze u krvi pre i posle doručka?

$$x_i = [90\,\mathrm{mg/dl}\,, 145\,\mathrm{mg/dl}]$$
$$P(diabetes|x_i)$$
$$P(\neg\ diabetes|x_i)$$

Primer 1

- Želimo da odredimo da li pacijent ima određenu formu raka. Znamo da svega 0.8% ljudi na svetu ima ovu formu raka. Postoji test krvi koji nam vraća POS i NEG rezultat. Ako osoba nema rak, dobiće NEG rezultat u 97% slučajeva. Ako osoba ima rak, dobiće POS rezultat u 98% slučajeva

$P(\text{cancer}) = 0.008$

$P(\neg\text{cancer}) = 0.992$

$P(\text{NEG}|\neg\text{cancer}) = 0.97$

$P(\text{POS}|\neg\text{cancer}) = 0.03$

$P(\text{POS}|\text{cancer}) = 0.98,$

$P(\text{NEG}|\text{cancer}) = 0.02$

Recimo da je osoba dobila rezultat POS na testu krvi. Koja je verovatnoća da ima rak?

$$P(\text{cancer}|\text{POS}) = \frac{P(\text{POS}|\text{cancer})\text{P}(\text{cancer})}{P(\text{POS})}$$

$$= \frac{0.98 \cdot 0.008}{0.98 \cdot 0.008 + 0.03 \cdot 0.992} = 0.21$$

Primer 1

$$P(\text{cancer}) = 0.008$$
$$P(\neg\text{cancer}) = 0.992$$

$$P(\text{NEG}|\neg\text{cancer}) = 0.97$$
$$P(\text{POS}|\neg\text{cancer}) = 0.03$$

$$P(\text{POS}|\text{cancer}) = 0.98,$$
$$P(\text{NEG}|\text{cancer}) = 0.02$$

$0.98 \cdot 8000$

10^6 People tested	POS	NEG
People with cancer (8000)	7 840	160
People without cancer (992 000)	29 760	962 240

$0.03 \cdot 992000$

Apriorna verovatnoća

$$\text{aposteriorna verovatnoća} = \frac{\text{uslovna verovatnoća} \cdot \text{apriorna verovatnoća}}{\text{dokaz}}$$

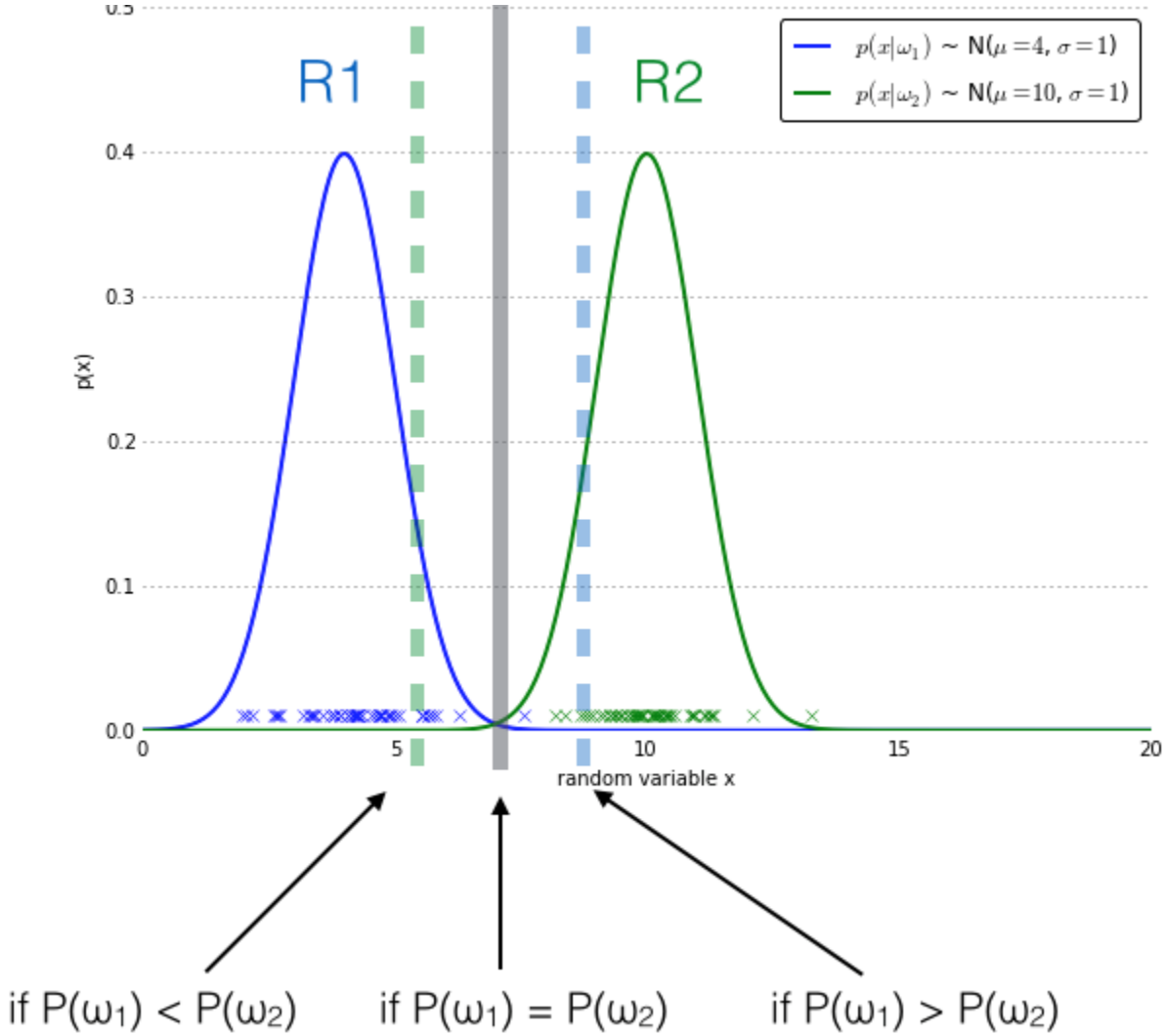
- U kontekstu klasifikacije: verovatnoća da ćemo naići na određenu klasu
  - $P(\text{cancer})$ - verovatnoća da (bilo koja) osoba ima rak
  - Ovu verovatnoću možemo dobiti u konsultaciji sa domenskim ekspertima ili estimirati iz podataka
  - *Maximum Likelihood* estimate:

$$P(y = c) = \frac{N_{y=c}}{N}$$

Broj primera iz klase *c*

Ukupan broj primera

Apriorna verovatnoća



- $x \epsilon \mathbb{R},\ y \in \{\text{plava}, \text{zelena}\}$
- Primeri *x* su izvučeni iz normalne distribucije $\sigma = 1,\ \mu_{Plava} = 4,\ \mu_{zelena} = 10$
- Ako su apriorne verovatnoće jednake $P(\text{plava}) = \text{P}(\text{zelena}) = 0.5$, granica odluke se nalazi između dve distribucije
- U suprotnom, biće pomerena ka jednoj od distribucija

Bajesovo pravilo odluke

- Ako znamo $P(x|y)$ i $P(y)$
- Možemo da damo predikciju klase pomoću Bajesove teoreme:

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

Dokaz

- Dokaz: verovatnoća da naiđemo na određenu kombinaciju vrednosti obeležja x (bez obzira kojoj klasi pripada)
- $P(x)$ možemo izračunati na sledeći način:

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

Bajesovo pravilo odluke

- Predikcija: klasa *i* sa najvećom verovatnoćom

$$\mathrm{h}_{MAP} = \arg\max_{c} P(y = c|x)$$

Maximum a Posteriory Hypothesis (MAP)

$$\mathrm{h}_{MAP} = \arg\max_{c} \frac{P(x|y = c)P(y = c)}{P(x)}$$

- Iako nam je $P(x)$ potrebno da izračunamo aposteriorne verovatnoće, sa aspekta donošenja odluke, ovo je samo faktor skaliranja:

$$\mathrm{h}_{MAP} = \arg\max_{c} P(x|y = c)P(y = c)$$

Primer 2

Zipcode	bought organic produce?	bought Sencha green tea?
88005	Yes	Yes
88001	No	No
88001	Yes	Yes
88005	No	No
88003	Yes	No
88005	No	Yes
88005	No	No
88001	No	No
88005	Yes	Yes
88003	Yes	Yes

- Koja je verovatnoća da se neko nalazi u Zipcode 88005?
  - $P(\text{Zipcode} = 88005) = \frac{5}{10} = 0.5$
- Koja je verovatnoća da se neko nalazi u Zipcode 88005 ako znamo da je kupio zeleni čaj?
  - $P(\text{Zipcode} = 88005 | y = \text{Yes}) = \frac{3}{5} = 0.6$
- Koja je verovatnoća da će neko kupiti zeleni čaj ako se nalazi u Zipcode 8805?
  - $P(y = \text{Yes} | \text{Zipcode} = 88005) = \frac{3}{5} = 0.6$
  - $P(y = \text{No} | \text{Zipcode} = 88005) = \frac{2}{5} = 0.4$

Zašto nam treba Bajesova teorema?

- U prethodnom primeru smo videli da lako možemo odrediti $P(y = \text{Yes}|\text{Zipcode} = 88005)$ i $P(y = \text{No}|\text{Zipcode} = 88005)$
- U opštem slučaju ovo ne važi. Uzmite u obzir primer testa za rak
- Lako je estimirati sledeće:
  - $P(cancer), \text{P}(\neg c\alpha\text{ncer})$ - lako dostupna statistika
  - $P(\text{POS}|\text{cancer}), \text{P}(\text{NEG}|c\alpha\text{ncer})$ - uzećemo reprezentativnu grupu pacijenata za koje znamo da imaju rak i primeniti test
  - $P(\text{POS}|\neg\text{cancer}), \text{P}(\text{NEG}|\neg c\alpha\text{ncer})$ - uzećemo reprezentativnu grupu pacijenata za koje znamo da nemaju rak i primeniti test
- Teško je direktno estimirati $P(\text{cancer}|\text{POS})$ i $P(\text{cancer}|\text{NEG})$
  - Ovo bi zahtevalo da damo test slučajno odabranoj osobi iz cele populacije
  - Za ovo bi nam trebao reprezentativan uzorak celokupne populacije. Pošto svega 0.08% populacije ima rak, trebao bi nam ekstremno velik uzorak