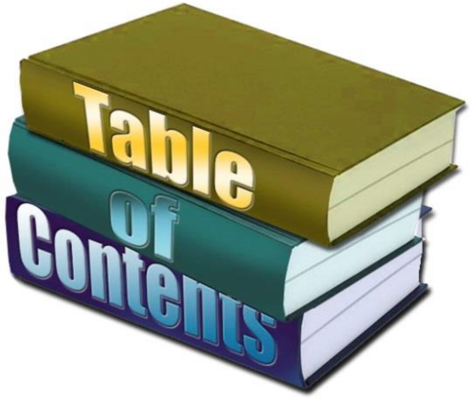
OSNOVE UMETNE INTELIGENCE 2022/23

diskretizacija, manjkajoči atributi
naivni Bayesov klasifikator
nomogrami

Pridobljeno znanje s prejšnjih predavanj

- **strojno učenje**
  - prostor hipotez odločitvenih dreves
  - problem pretiranega prileganja učnim podatkom
    - učenje šuma
    - prevelika drevesa, slaba razumljivost
    - slaba splošnost
  - **strategije rezanja**: vnaprej, nazaj
  - ocenjevanje verjetnosti z bolj stabilnimi ocenami
  - rezanje nazaj:
    - REP: **uporaba rezalne množice**, odrežemo poddrevesa, ki povečujejo število napak glede na število napak v njihovem korenu
    - MEP: **uporaba m-ocene verjetnosti**, ki kombinira apriorno in empirično verjetnost, ocenimo statično in dinamično napako
  - ocenjevanje učenja (prečno preverjanje)

Pregled



- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami

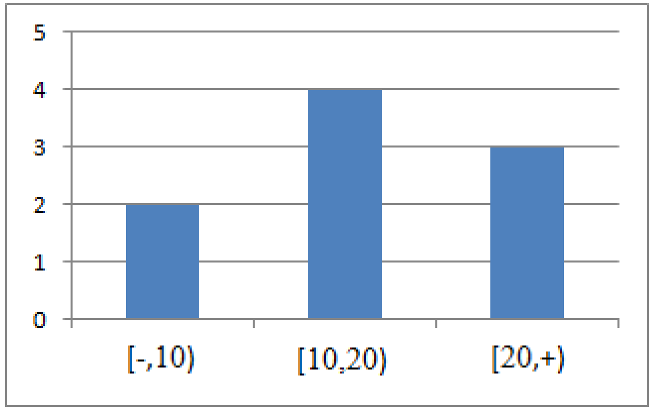
Različne vrste atributov

- vrste atributov:
  - **diskretni**:
    - **nominalni** ({true, false}, {M, F}, {French, Thai, Burger, Italian})
    - **ordinalni** ({None, Some, Full}, {Low, Med, High})
  - **zvezni/numerični** ($\in \mathbb{R}$)

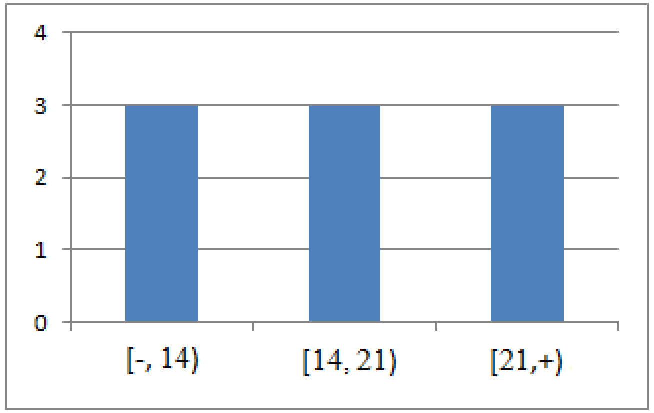
- **Kako obravnavati zvezne/numerične atribute?**
  - Običajno izvedemo diskretizacijo v dva (binarizacija) ali več diskretnih intervalov.
  - Različni pristopi k diskretizaciji:
    - intervali **enake širine** (equal-width)
    - intervali z **enako frekvenco** primerov (equal-frequency)
    - intervali, ki **maksimizirajo informacijski prispevek**

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)

Equal width



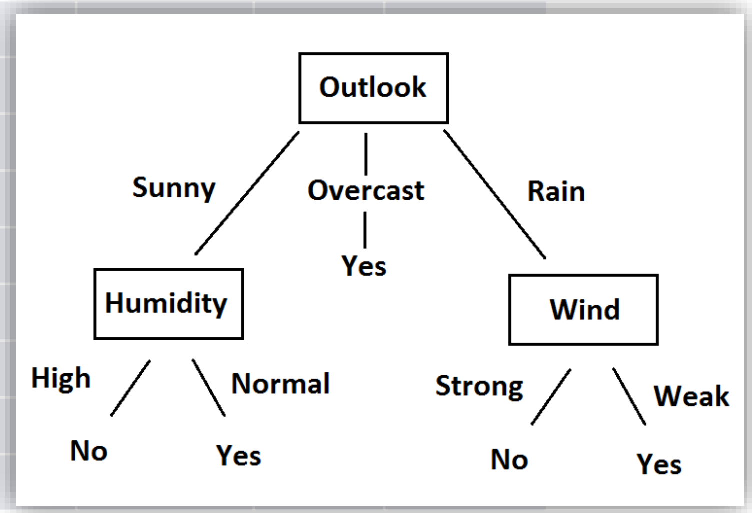
Equal frequency



Diskretizacija z maksimizacijo informacijskega prispevka

- numerični atributi namesto diskretnih?

Day	Outlook	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool			
7	Overcast	Cool			
8	Sunny	Mild			
9	Sunny	Cool			
10	Rain	Mild			
11	Sunny	Mild			
12	Overcast	Mild			
13	Overcast	Hot			
14	Rain	Mild			



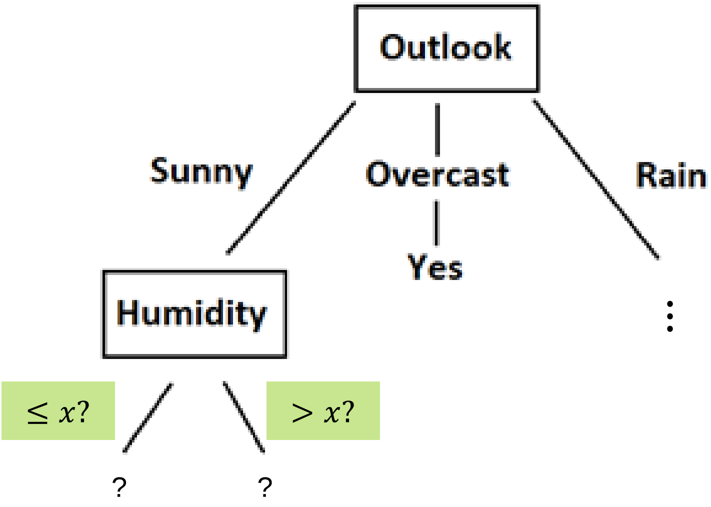
???

Day	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	86	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	91	Strong	No

https://www.slideshare.net/prithwis/lecture02-data-mining-analytics
https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/
https://www.chegg.com/homework-help/questions-and-answers/following-table-gives-data-set-deciding-whether-play-cancel-ball-game-depending-weather-co-q27982386

Diskretizacija z maksimizacijo informacijskega prispevka

- kako diskretizirati atribut Humidity?



Day	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
11	Sunny	75	70	Strong	Yes

- primeri možni vrednosti za mejo x: 70, 85, 90

- entropije ob različnih delitvah:
  - $x = 70$ → levo [2,0], desno [0,3]
    $H_{res} = \frac{2}{5} \cdot 0 + \frac{3}{5} \cdot 0 = 0$

    najnižja residualna entropija (najvišji informacijski prispevek)
  - $x = 85$ → levo [2,1], desno [0,2]
    $H_{res} = \frac{3}{5} \cdot \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{5} \cdot 0 = 0{,}551$
  - $x = 90$ → levo [2,2], desno [0,1]
    $H_{res} = \frac{4}{5} \cdot 1 + \frac{1}{5} \cdot 0 = 0{,}8$

- izberemo:
  - LOW: ≤ 70
  - HIGH: > 70

https://www.slideshare.net/prithwis/lecture02-data-mining-analytics
https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/

Obravnava manjkajočih atributov

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	???	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

- različni načini **obravnave manjkajočih** atributov, glede na:
    - ali učni algoritem lahko poišče hipotezo **tudi, če manjkajo** vrednosti atributov?
    - ali imamo **dovolj** učnih primerov na razpolago?
    - ali je primerno **nadomeščanje** manjkajočih vrednosti?

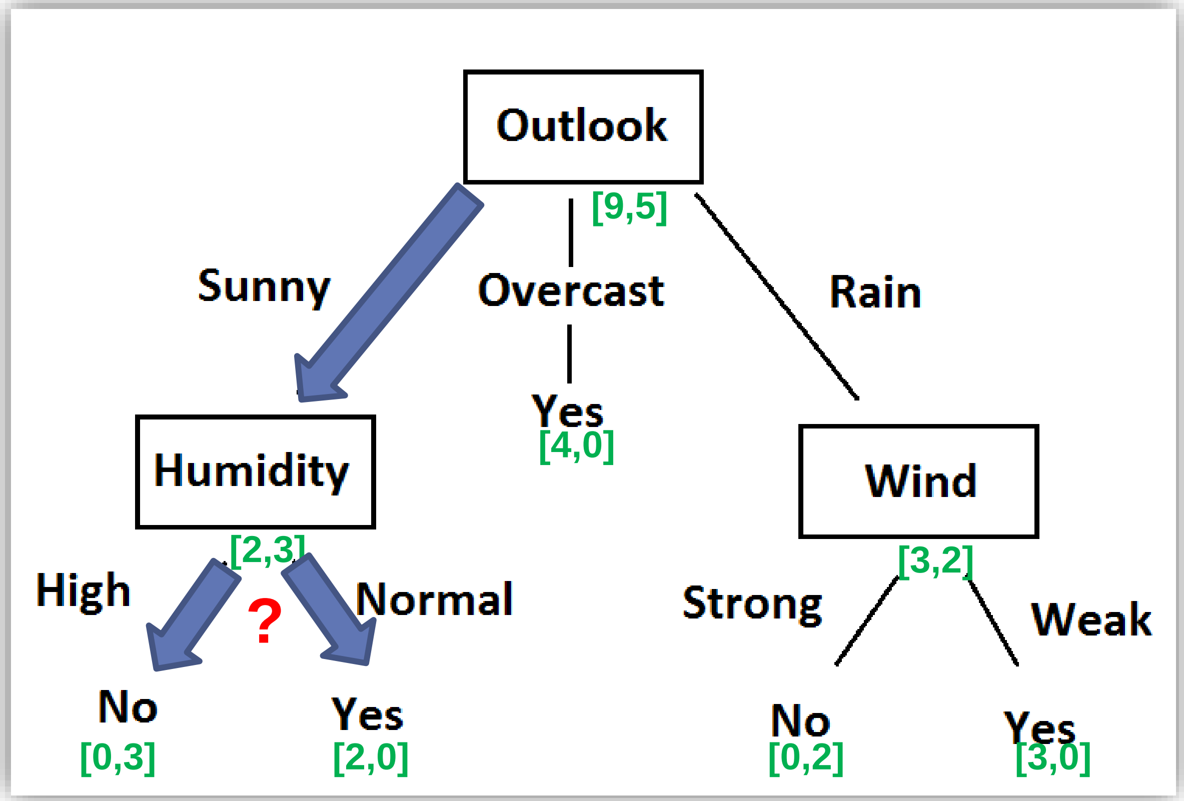
- **pristopi pri učenju**:
    - učimo z manjkajočimi vrednostmi
    - ignoriranje celih učnih primerov z neznanimi vrednostmi
    - uporaba posebne vrednosti NA/UNKNOWN?
    - nadomestiti manjkajočo vrednost (povprečna, najbolj pogosta, najbolj pogosta glede na razred, naključna, napovedana)
    - primer obravnavamo verjetnostno glede na vse možne vrednosti atributa

- **pristopi pri napovedovanju**:
    - verjetnostna klasifikacija glede na vse možne vrednosti atributa (npr. pri drevesu: klasificiramo v več listov)

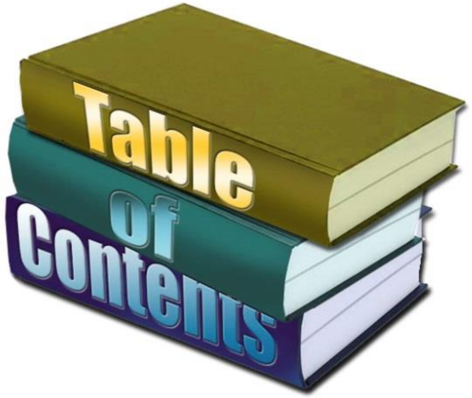
Obravnava manjkajočih atributov

- Kako klasificirati primer
  Outlook=Sunny,
  Temperature=Mild,
  Humidity=???,
  Wind=Weak ?

- verjetnostna klasifikacija glede na relativno frekvenco:
  - P(razred = No | primer) = 3/5
  - P(razred = Yes | primer) = 2/5



Pregled



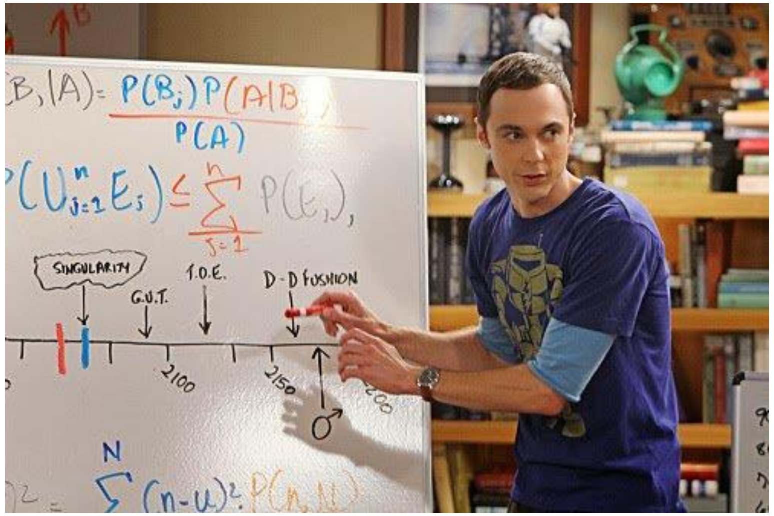
- strojno učenje
    - uvod v strojno učenje
    - učenje odločitvenih dreves
    - učenje dreves iz šumnih podatkov (rezanje dreves)
    - ocenjevanje učenja
    - diskretizacija atributov, obravnava manjkajočih vrednosti
    - naivni Bayesov klasifikator
    - nomogrami

Naivni Bayesov klasifikator

- Thomas Bayes, 1702 – 1761
- opomnik iz teorije o verjetnosti:

$$P(AB) = P(A|B) \cdot P(B)$$
$$P(AB) = P(B|A) \cdot P(A)$$
$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$
$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

Bayesovo pravilo



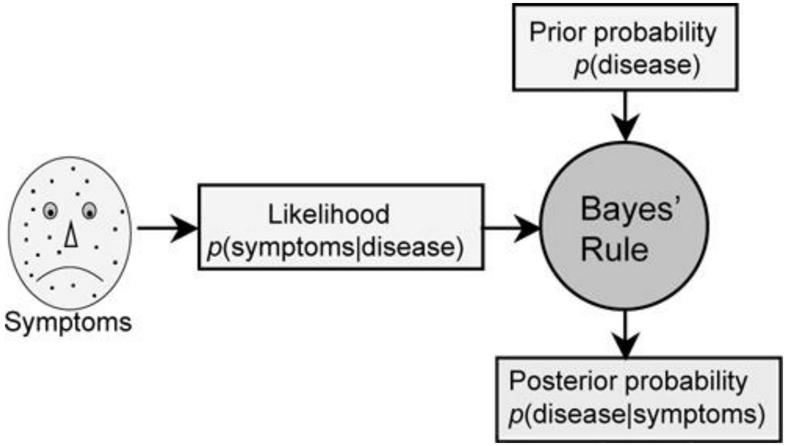
Naivni Bayesov klasifikator

- zdravniki razpolagajo z **vzročno in statistično informacijo** o simptomih (opažanja) in diagnozah (hipoteza):
  - verjetnost **izraženih simptomov pri neki bolezni** - $P(opažanje|hipoteza)$
  - verjetnost **določene bolezni** - $P(hipoteza)$
  - verjetnost **določenega simptoma** - $P(opažanje)$

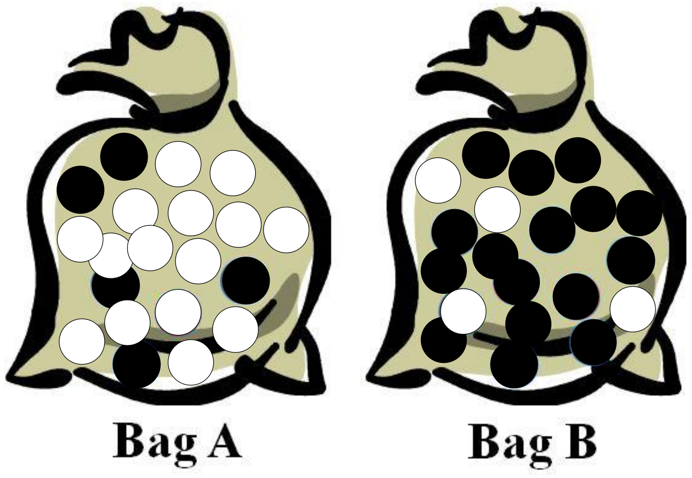
- aplikacija v medicini:

$$P(hipoteza|opažanje) = \frac{P(opažanje|hipoteza) \cdot P(hipoteza)}{P(opažanje)}$$

- Bayesovo pravilo nam izraža **diagnostično pogojno verjetnost** $P(hipoteza|opažanje)$ na podlagi **vzročne pogojne verjetnosti** $P(opažanje|hipoteza)$



Vaja – opomnik iz verjetnosti



- dve vrsti vrečk s frnikulami:
  - 4 vrečke tipa A (vsaka 5 črnih, 15 belih frnikul)
  - 1 vrečka tipa B (16 črnih, 4 bele frnikule)

- primeri vprašanj, zapisanih kot verjetnosti:
  - Kakšna je verjetnost, da je to vrečka tipa B?
    $P(B) = ?$
  - Kakšna je verjetnost, da naključno izberemo črno frnikulo, če izbiramo iz vrečke tipa B?
    $P(\check{C}|B) = ?$
  - Naključno izberemo eno izmed vrečk in iz nje naključno izberemo frnikulo. Kakšna je verjetnost, da smo izbrali črno frnikulo iz vrečke tipa B?
    $P(B\check{C}) = P(B) \cdot P(\check{C}|B) = ?$
  - Naključno izberemo eno izmed vrečk in iz nje naključno izberemo frnikulo. Kakšna je verjetnost, da smo izbrali črno frnikulo?
    $P(\check{C}) = P(B) \cdot P(\check{C}|B) + P(A) \cdot P(\check{C}|A)$

Vaja

- Ena vrečka ima poškodovan ovoj tako, da se skozi njega vidi črna frnikula. Kakšna je verjetnost, da je to vrečka tipa B?
  $P(B|\check{C}) = ?$
- B = hipoteza, Č = evidenca, opažanje
- verjetnost $P(B|\check{C})$ lahko določimo iz drugih bolj očitnih verjetnosti z Bayesovo formulo:

$$P(B|\check{C}) = \frac{P(B) \cdot P(\check{C}|B)}{P(\check{C})}$$

- $P(B) = \frac{1}{5} = 0{,}2$
  $P(\check{C}|B) = \frac{16}{20} = 0{,}8$
  $P(\check{C}) = \frac{4 \cdot 5 + 1 \cdot 16}{5 \cdot 20} = 0{,}360$
- $P(B|\check{C}) = \frac{0{,}2 \cdot 0{,}8}{0{,}360} = 0{,}444$

dve vrsti vrečk s frnikulami:

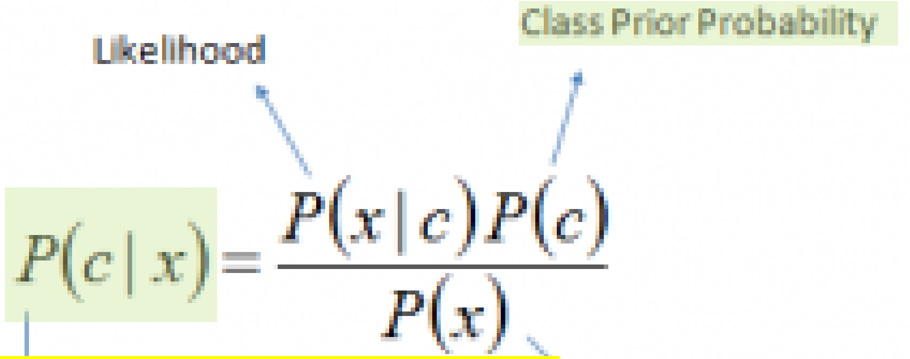
- 4 vrečke tipa A (vsaka 5 črnih, 15 belih frnikul)
- 1 vrečka tipa B (16 črnih, 4 bele frnikule)

Naivni Bayes v strojnem učenju

- evidenca → atributi
  hipoteza → razred
- zanima nas, kakšna je verjetnost razreda C pri podanih vrednostih atributov $A_1 = X_1, A_2 = X_2, \dots, A_n = X_n$:

$$P(C|X_1X_2 \dots X_n) = \frac{P(C) \cdot P(X_1X_2 \dots X_n|C)}{P(X_1X_2 \dots X_n)}$$

apriorna verjetnost razreda



aposteriorna verjetnost napovedi

Naivni Bayes v strojnem učenju

$$P(hipoteza|evidenca) = \frac{P(hipoteza) \cdot P(evidenca|hipoteza)}{P(evidenca)} \Rightarrow P(C|X_1X_2 \dots X_n) = \frac{P(C) \cdot P(X_1X_2 \dots X_n|C)}{P(X_1X_2 \dots X_n)}$$

- pravilo za produkt verjetnosti: $P(AB) = P(A) \cdot P(B|A)$
- pravilo za produkt verjetnosti s pogojno verjetnostjo: $P(AB|C) = P(A|C) \cdot P(B|AC)$

- $P(X_1X_2 \dots X_n|C) = P(X_1|C) \cdot P(X_2 \dots X_n|X_1C) =$
  $= P(X_1|C) \cdot P(X_2|X_1C) \cdot P(X_3 \dots X_n|X_1X_2C) =$
  $= P(X_1|C) \cdot P(X_2|X_1C) \cdot P(X_3|X_1X_2C) \cdot \dots \cdot P(X_n|X_1X_2 \dots X_{n-1}C)$
- $P(X_1X_2 \dots X_n) = P(X_1|X_2 \dots X_n) \cdot P(X_2|X_3 \dots X_n) \cdot \dots \cdot P(X_{n-1}|X_n) \cdot P(X_n)$

verižno pravilo

- potrebujemo veliko število pogojnih verjetnosti, katerih poznavanje je v praksi težavno
- število kombinacij pogojnih verjetnosti je glede možne vrednosti $X_1X_2 \dots X_n$ eksponentno
- praktična rešitev: **naivni Bayesov klasifikator**

Naivni Bayes v strojnem učenju

- predpostavimo, da so atributi med seboj **verjetnostno neodvisni** in poenostavimo:

$$P(X_1X_2 \dots X_n|C) = P(X_1|C) \cdot P(X_2|X_1C) \cdot \dots \cdot P(X_n|X_1X_2 \dots X_{n-1}C)$$
$$P(X_1X_2 \dots X_n) = P(X_1|X_2 \dots X_n) \cdot P(X_2|X_3 \dots X_n) \cdot \dots \cdot P(X_{n-1}|X_n) \cdot P(X_n)$$

$$\Downarrow$$

$$P(X_1X_2 \dots X_n|C) \approx P(X_1|C) \cdot P(X_2|C) \cdot \dots \cdot P(X_n|C)$$
$$P(X_1X_2 \dots X_n) \approx P(X_1) \cdot P(X_2) \cdot \dots \cdot P(X_{n-1}) \cdot P(X_n)$$

- približki so dobri, **če so atributi med seboj dovolj neodvisni**
- velja torej:

$$P(C|X_1X_2 \dots X_n) = \frac{P(C) \cdot P(X_1X_2 \dots X_n|C)}{P(X_1X_2 \dots X_n)} \approx \frac{P(C) \cdot \prod_i P(X_i|C)}{\prod_i P(X_i)}$$

konstanten člen, ki je neodvisen od ciljne spremenljivke (če opazujemo samo relativne velikosti napovedi različnih razredov, ga lahko izpustimo)

Naivni Bayes v strojnem učenju

- Bayesov klasifikator: primer **klasificiramo v razred, ki je najbolj verjeten**:

$$h(C_k|X_1X_2 \ldots X_n) = \operatorname{argmax}_k P(C_k) \cdot \prod_{i=1}^{n} P(X_i|C_k)$$

(to ni več verjetnost)

- **učenje**: ocenimo verjetnosti $P(C_k)$ in $P(X_i|C_k)$ za vse razrede C_k in vrednosti atributov X_i
- **napovedovanje**: uporabimo zgornjo enačbo za napovedovanje razreda novim primerom
- *opomba: s poenostavitvijo formule in izpustitvijo imenovalca izgubimo verjetnostno interpretacijo (verjetnosti razredov se ne seštejejo več v 1). Problem rešujemo npr. z normalizacijo rezultatov.*

Primer

- Zajeli smo podatke za 1000 sadežev, ki so lahko bodisi: *banana*, *pomaranča* ali *drugi sadež* (= vrednosti **razreda**). Za vsakega izmed sadežov smo izmerili, ali je *podolgovat*, *sladek* in *rumen* (= **atributi**). Meritve smo zapisali v tabelo:

sadež	**podolgovat**		**sladek**		**rumen**		**skupaj**
	da	**ne**	**da**	**ne**	**da**	**ne**	
banana	400	100	350	150	450	50	500
pomaranča	0	300	150	150	300	0	300
drugo	100	100	150	50	50	150	200
	500	500	650	350	800	200	1000

- iz tabele lahko razberemo različne verjetnosti, npr.:
  - verjetnosti razredov: $P(banana) = \frac{500}{1000} = 0{,}5$, $P(pomaranča) = 0{,}3$, $P(drugo) = 0{,}2$
  - pogojne verjetnosti: $P(dolg|banana) = \frac{4}{5} = 0{,}8$

    $P(sladek|banana) = 0{,}7$

    $P(rumen|banana) = 0{,}9$

A simple explanation of Naive Bayes Classification, https://stackoverflow.com/questions/10059594/a-simple-explanation-of-naive-bayes-classification

Primer

sadež	podolgovat		sladek		rumen		skupaj
	da	ne	da	ne	da	ne	
banana	400	100	350	150	450	50	500
pomaranča	0	300	150	150	300	0	300
drugo	100	100	150	50	50	150	200
	500	500	650	350	800	200	1000

- Imamo sadež, ki ni podolgovat, ni sladek, je pa rumen. Kateri sadež je to?

$$P(banana|neP, neS, daR)$$
$$\approx P(banana) \cdot P(neP|banana) \cdot P(neS|banana) \cdot P(daR|banana) =$$
$$= \frac{500}{1000} \cdot \frac{100}{500} \cdot \frac{150}{500} \cdot \frac{450}{500} = 0{,}5 \cdot 0{,}2 \cdot 0{,}3 \cdot 0{,}9 = 0{,}027$$

$$P(pomaranča|neP, neS, daR) \approx 0{,}3 \cdot 1 \cdot 0{,}5 \cdot 1 = 0{,}15 \longleftarrow$$

ta sadež je najverjetneje pomaranča

$$P(drugo|neP, neS, daR) \approx 0{,}2 \cdot 0{,}5 \cdot 0{,}25 \cdot 0{,}25 = 0{,}00625$$

Izpitna naloga

- 2. izpitni rok, 15. 2. 2018 (prilagojena naloga)

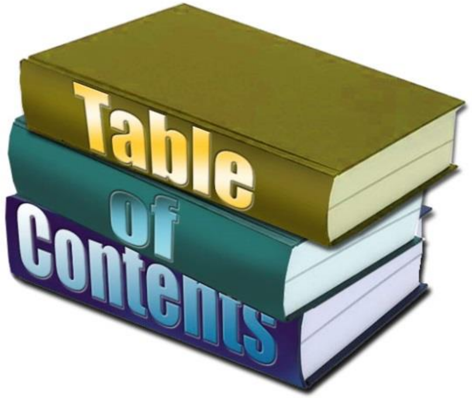
2. NALOGA (25t):

Podana je učna množica primerov, ki je prikazana v tabeli (*vreme* in *pritisk* sta atributa, *glavobol* pa je razred). Naloge:

V kateri razred bi naivni Bayesov klasifikator klasificiral učni primer z vrednostmi atributov *vreme=deževno, pritisk=srednji* (verjetnosti računamo z relativno frekvenco)?

vreme	pritisk	glavobol
sončno	nizek	ne
sončno	nizek	ne
sončno	srednji	da
sončno	visok	ne
sončno	nizek	ne
sončno	nizek	da
deževno	srednji	ne
deževno	srednji	da
deževno	visok	da

Pregled

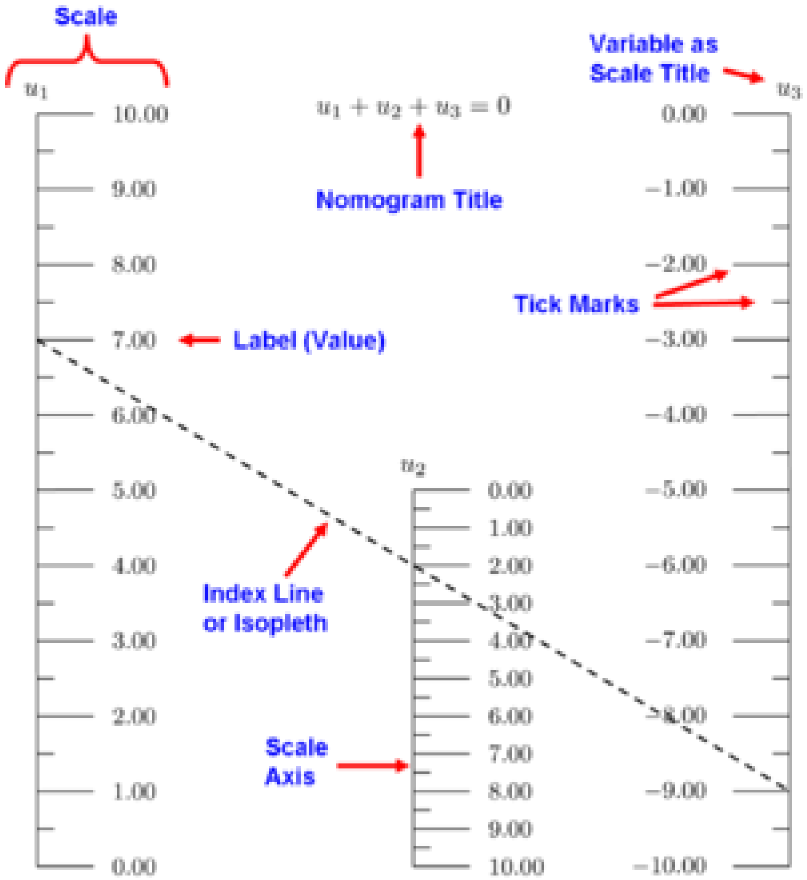


- strojno učenje
  - uvod v strojno učenje
  - učenje odločitvenih dreves
  - učenje dreves iz šumnih podatkov (rezanje dreves)
  - ocenjevanje učenja
  - diskretizacija atributov, obravnava manjkajočih vrednosti
  - naivni Bayesov klasifikator
  - nomogrami

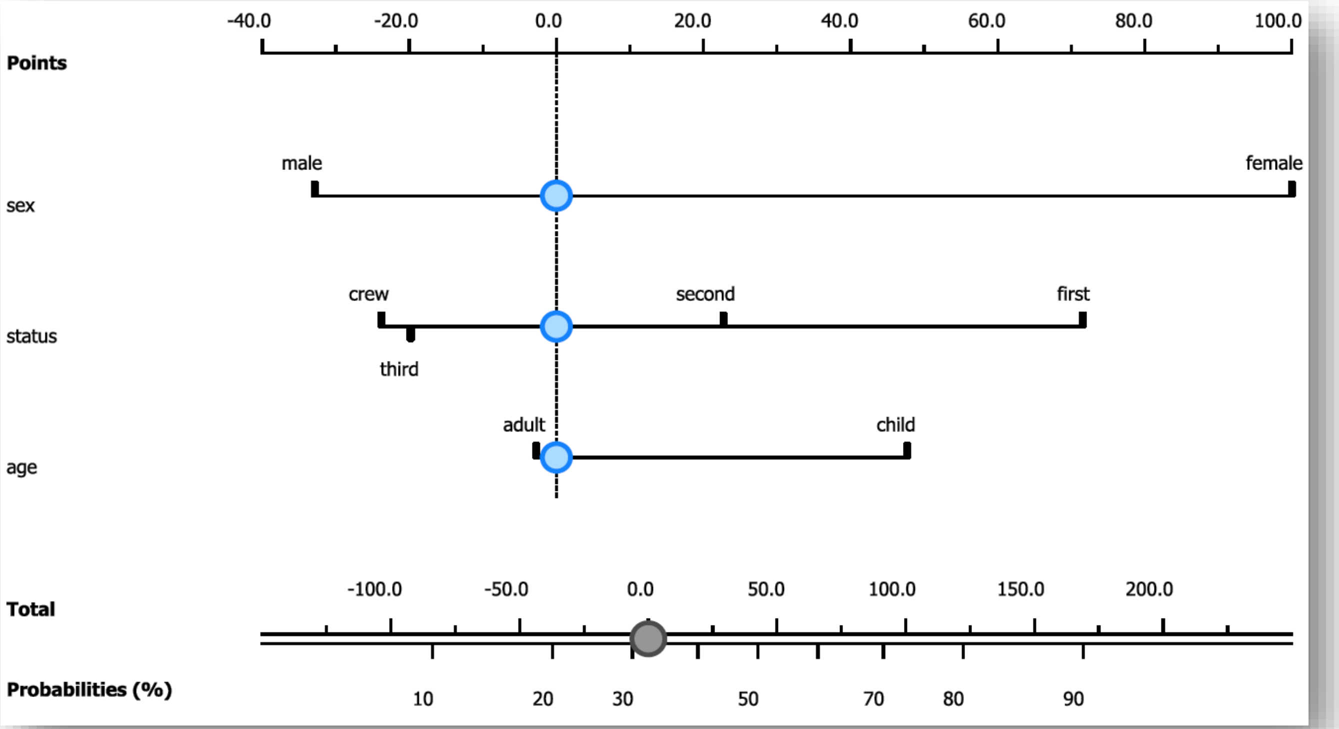
Nomogrami

- pristop za **vizualizacijo** naivnega Bayesovega modela

- "nomogram":
  - je *grafična upodobitev numeričnih odnosov med spremenljivkami*
  - omogoča uporabniku grafično pridobiti rezultat brez računanja
- uporaba:
  - matematika (iskanje vrednosti funkcij)
  - zdravniki v medicini (napovedovanje bolezni – npr. infarkta ali raka na podlagi vhodnih atributov)
- prikazuje:
  - pomembnost posameznih **vrednosti** vsakega atributa na ciljni razred
  - pomembnost posameznih **atributov** na ciljni razred
  - **vizualno razlago** napovedanih verjetnosti (brez kalkulatorja)



Primer - ideja



- vsaka **vrednost atributa doprinaša določeno število točk** k ciljnemu razredu
- točke vseh vrednosti atributov seštejemo v **skupno vsoto točk**, ki je povezana z verjetnostjo ciljnega razreda
- **razpon točk vsakega atributa** govori o pomembnosti atributa za napovedovanje ciljnega razreda (zgoraj urejeni od najbolj do najmanj pomembnega)

Izračun nomograma

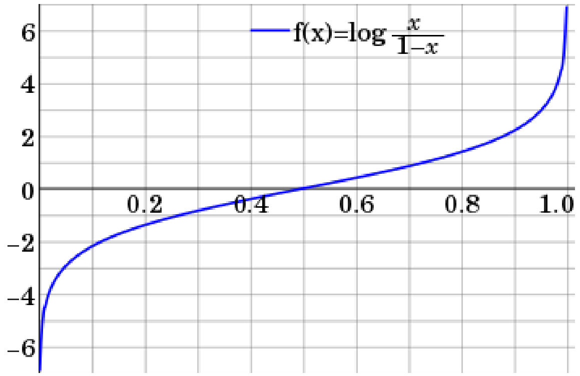
- verjetje razreda pri naivnem Bayesu:

$$h(C|X_1X_2\dots X_n) = P(C)\cdot\prod_{i=1}^{n} P(X_i|C)$$

- na zgornjem pravilu uporabimo logistično funkcijo (verjetnosti z intervala [0,1] preslikamo na interval $(-\infty,\infty)$, uporabimo logaritme)

$$\text{logit}\, P = \log\frac{P}{1-P}$$

logistična funkcija



- $\text{logit}\, h(C|X_1X_2\dots X_n) =$

$$\log\frac{P(C)\cdot\prod_{i=1}^{n}P(X_i|C)}{1-P(C)\cdot\prod_{i=1}^{n}P(X_i|C)} = \log\frac{P(C)\cdot\prod_{i=1}^{n}P(X_i|C)}{P(\bar{C})\cdot\prod_{i=1}^{n}P(X_i|\bar{C})} = log\frac{P(C)}{P(\bar{C})} + \log\frac{\prod_{i=1}^{n}P(X_i|C)}{\prod_{i=1}^{n}P(X_i|\bar{C})}$$

$$= logit\, P(C) + \sum_i \log\frac{P(X_i|C)}{P(X_i|\bar{C})}$$

samo ta člen je odvisen od vrednosti atributov X_i,
to količino imenujemo razmerje verjetja
(odds ratio)

Izračun nomograma

to količino imenujemo razmerje verjetja (odds ratio)

- $\text{logit}\, P(C) + \sum_i \log \frac{P(X_i|C)}{P(X_i|\bar{C})} = \text{logit}\, P(C) + \sum_i \log OR(X_i)$
- od zgornjih členov je edino razmerje verjetja (drugi člen) **odvisno od vrednosti atributov $\boldsymbol{X_i}$**, torej ga lahko uporabimo za "točkovanje" doprinosa atributa:

$$točke(C|X_i) = \log OR(X_i) = \log \frac{P(X_i|C)}{P(X_i|\bar{C})}$$

- **skupno število točk** za verjetnost celotnega primera (upoštevamo vse attribute):

$$točke(C|X_1 X_2 \dots X_n) = \sum_i \log OR(X_i) = \sum_i \log \frac{P(X_i|C)}{P(X_i|\bar{C})}$$

- Kako izračunati **izraz znotraj logaritma $\boldsymbol{OR(X_i)}$**? Po Bayesovem pravilu (znova) velja:

$$\frac{P(X_i|C)}{P(X_i|\bar{C})} = \frac{\frac{P(C|X_i) \cdot P(X_i)}{P(C)}}{\frac{P(\bar{C}|X_i) \cdot P(X_i)}{P(\bar{C})}} = \frac{\frac{P(C|X_i)}{P(C)}}{\frac{P(\bar{C}|X_i)}{P(\bar{C})}} = \frac{\frac{P(C|X_i)}{P(\bar{C}|X_i)}}{\frac{P(C)}{P(\bar{C})}}$$

Primer

- učna množica titanic, 2201 učnih primerov (711 preživelih – razred YES, 1490 umrlih – razred NO)

atribut	vrednost	razred = YES	razred = NO
status	first	203	122
	second	118	167
	third	178	528
	crew	212	673
age	adult	654	1438
	child	57	52
sex	male	367	1364
	female	344	126
razred		711	1490

- Kako konstruiramo nomogram?
- Kako lahko vizualiziramo odločitev za **odraslega moškega**, ki je potoval v **drugem** razredu?

primer iz Možina, Martin, et al. "Nomograms for visualization of naive Bayesian classifier." European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 2004.

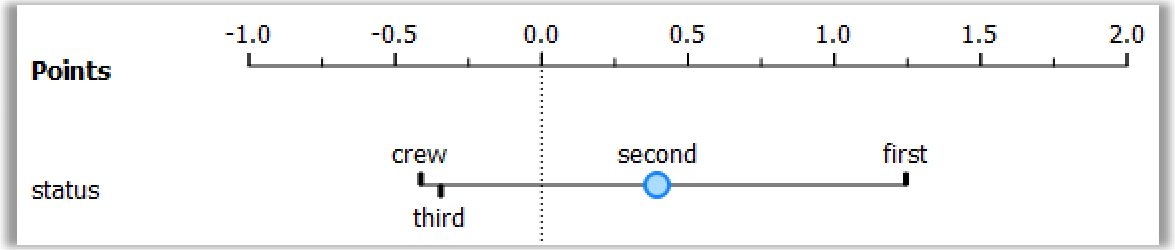
Konstrukcija nomograma

$$točke(yes|status = first) = \log\frac{\frac{P(yes|first)}{P(no|first)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{203}{122}}{\frac{711}{1490}} = \log\frac{1{,}66}{0{,}48} = 1{,}25$$

$$točke(yes|status = second) = \log\frac{\frac{P(yes|second)}{P(no|second)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{118}{167}}{\frac{711}{1490}} = \log\frac{0{,}71}{0{,}48} = 0{,}39$$

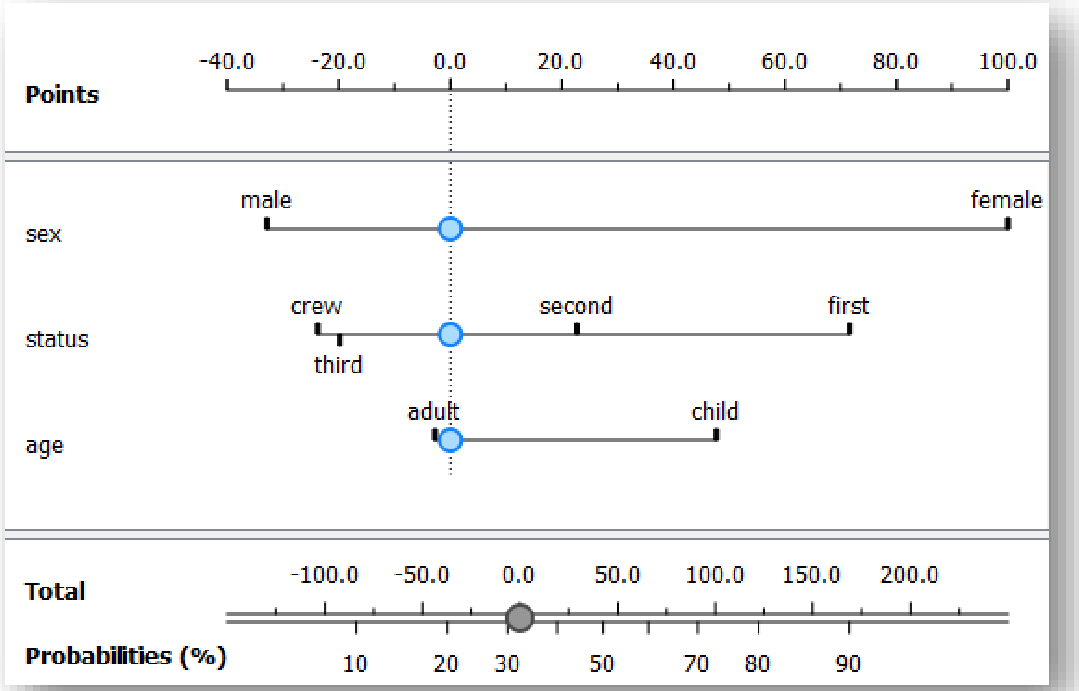
$$točke(yes|status = third) = \log\frac{\frac{P(yes|third)}{P(no|third)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{178}{528}}{\frac{711}{1490}} = \log\frac{0{,}34}{0{,}48} = -0{,}35$$

$$točke(yes|status = crew) = \log\frac{\frac{P(yes|crew)}{P(no|crew)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{212}{673}}{\frac{711}{1490}} = \log\frac{0{,}32}{0{,}48} = -0{,}42$$



Konstrukcija nomograma

- osi ostalih atributov **poravnamo glede na ničelno vrednost** prispevka atributa
- prikažemo lahko tudi **skupno skalo** za celotno napoved (vsoto točk)
- točke posameznih vrednosti atributov (log OR) lahko skaliramo v skalo točk, kjer s 100 točkami predstavimo prispevek največje vrednosti atributa



dejanske ali normalizirane točke atributa

skupna vsota točk

preslikava v verjetnosti

- skupne točke lahko preslikamo nazaj v verjetnosti

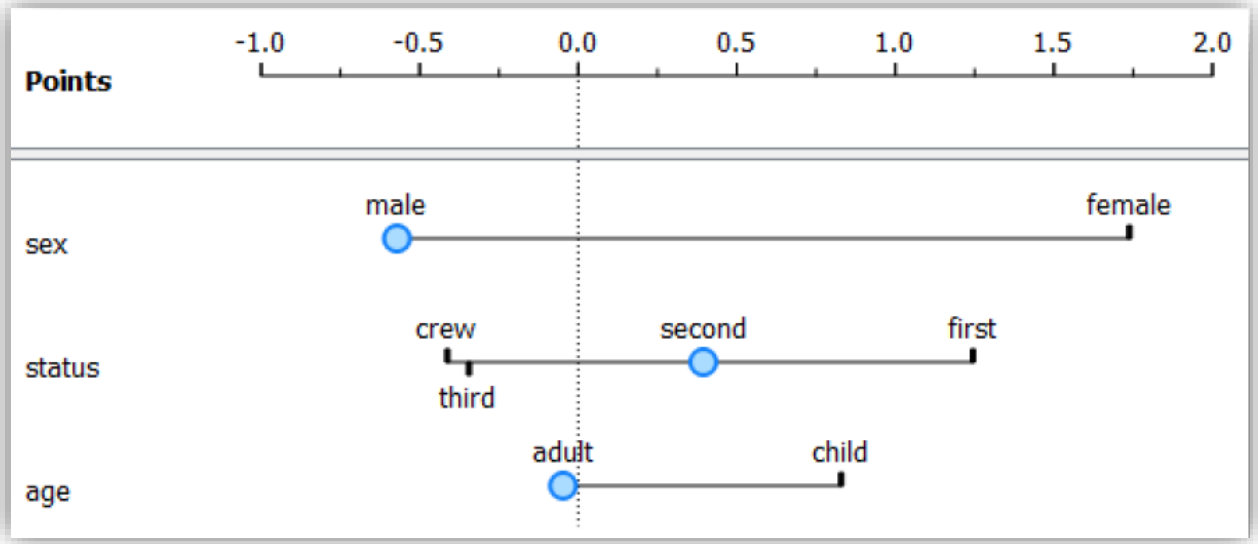
Neobvezna dodatna literatura: Možina, Martin, et al. "Nomograms for visualization of naive Bayesian classifier." European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 2004.

Primer

- Kako lahko pojasnimo odločitev, da je **odrasli moški**, ki je potoval v **drugem** razredu, **preživel**?
- $točke(yes|age = adult) = \log\frac{\frac{P(yes|adult)}{P(no|adult)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{654}{1438}}{\frac{711}{1490}} = \log\frac{0{,}45}{0{,}48} = -0{,}05$

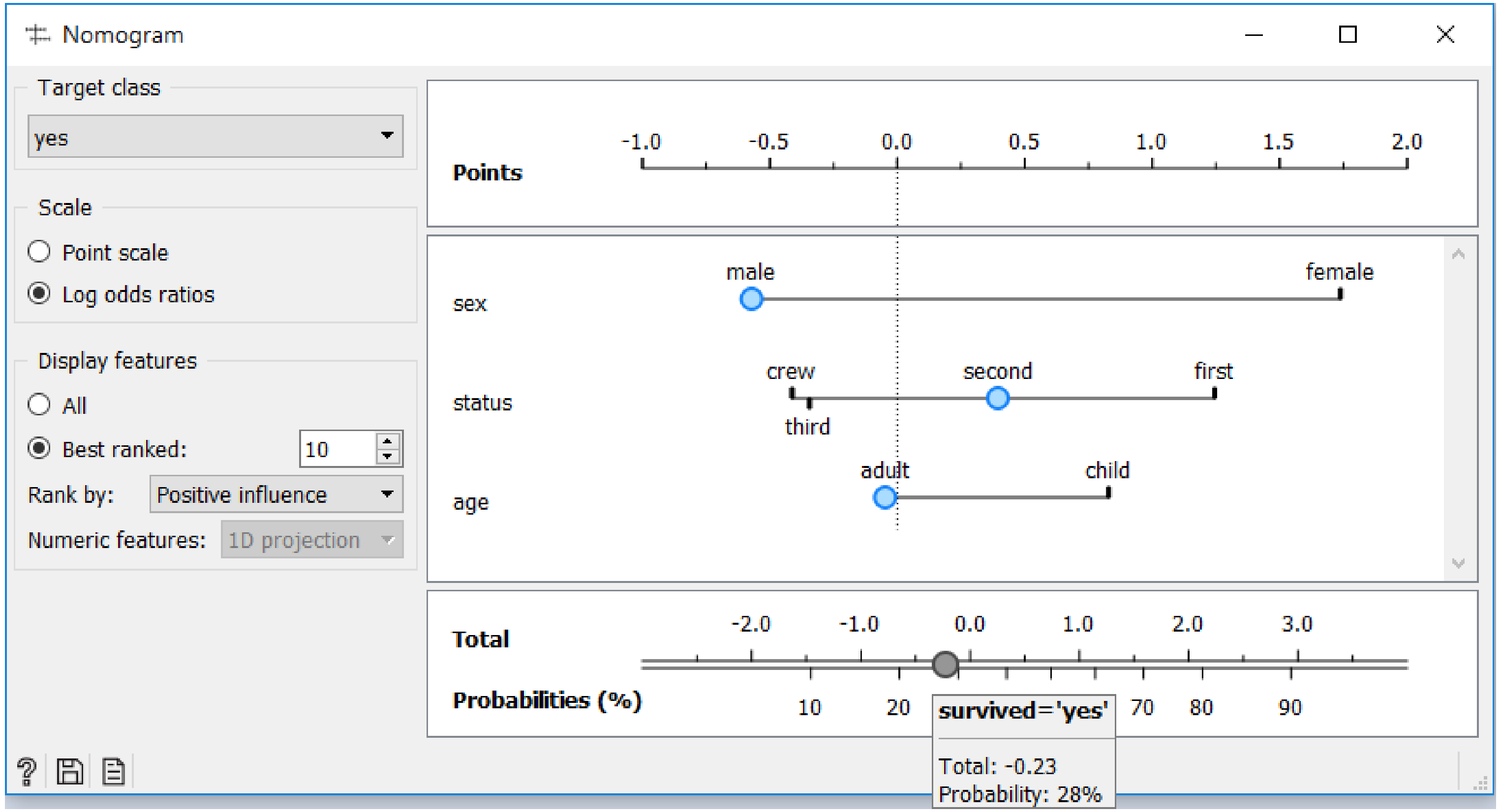
  $točke(yes|sex = male) = \log\frac{\frac{P(yes|male)}{P(no|male)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{367}{1364}}{\frac{711}{1490}} = \log\frac{0{,}27}{0{,}48} = -0{,}57$

  $točke(yes|status = second) = \log\frac{\frac{P(yes|second)}{P(no|second)}}{\frac{P(yes)}{P(no)}} = \log\frac{\frac{118}{167}}{\frac{711}{1490}} = \log\frac{0{,}71}{0{,}48} = 0{,}39$



Primer

- Kako lahko pojasnimo odločitev, da je **odrasli moški**, ki je potoval v **drugem** razredu, **preživel**?
- $točke(yes|adult, male, second) = -0{,}05 - 0{,}57 + 0{,}39 = -0{,}23$



Izpitna naloga

- 3. izpitni rok, 2. 9. 2019

3. NALOGA (10t):

Podana je učna množica podatkov o pacientih, ki so zboleli za srčno boleznijo. Ta vsebuje dva atributa:

- **družina**: če je za sorodno boleznijo zbolel tudi še kak ožji družinski član (vrednosti: da/ne)
- **spol**: spol pacienta (vrednosti: moški/ženska).

  Vsak učni primer je označen z razredom **bolezen**, ki pove, ali je pacient zbolel (vrednosti: da/ne).

Povzetek števila učnih primerov s posameznimi vrednostmi atributov je podan z naslednjo tabelo:

		bolezen	
		DA	NE
družina	da	200	150
	ne	120	110
spol	moški	140	160
	ženska	180	100

b.) Nariši nomogram za verjetnostno razlago modela za klasificiranje v razred bolezen=DA. Nomogram naj prikazuje:

- **skalo za prispevke atributa, ki naj prikazuje izračunano število točk (skaliranje ni potrebno),**
- **os za vsak atribut z jasno označenimi prispevki njegovih posameznih vrednosti.**

Pri izračunu uporabljaj naravni logaritem.

c) Pacient s kakšnimi lastnostmi ima največjo verjetnost, da ostane zdrav (pojasni odgovor)?

k najbližjih sosedov, regresija