IV130 Přínosy a rizika inteligentních systémů

Neurčitost, pravděpodobnost a užitek

21. dubna 2023

Akce za nejistoty/neurčitosti

- Aktéři v reálném světě potřebují zpracovávat **nejistotu** (**neurčitost**) v důsledku jen *částečné pozorovatelnosti*, *nedeterminismu* nebo *činnosti protivníků*
- Aktér nemusí být schopný znát *přesně* stav, v němž se nachází
- Aktér může *nejistotu* zachytit mechanismem vnitřního stavu přesvědčení o světě a řešit problém ve vztahu k možným stavům světa, které tomu odpovídají, ale
  - takový aktér musí pracovat s každým možným vysvětlením stavu senzorů bez ohledu na jejich pravděpodobnost, což vede k obrovské velikosti prostoru možných stavů,
  - plány vytvářené na základě všech možností rostou nade všechny meze a také zahrnují možnosti velmi nepravděpodobné, a
  - existují i situace, kdy žádný plán nezaručuje dosažení cíle, ale aktér musí nějak konat a je třeba mechanismu porovnávajícího hodnoty plánů, u nichž není garance finálního úspěchu.
- *Neúplným plánem* je např. plán na cestu z A do B za předpokladu, že se nerozbije auto, nedojde benzin, nedojde k havárii, atd., což jako podmínky s neznámou hodnotou nedovoluje sestavit plán garantující úspěch.
- *Neurčitost* (události) formálně zachycujeme pomocí *pravděpodobnosti* (události), která kvantifikuje možné stavy světa, v nichž dochází k dané události.
- *Racionální rozhodování* závisí na relativní důležitosti různých cílů, jejich pravděpodobnosti i stupně toho, jak jich lze dosáhnout, rozhodování mezi různými možnými plány zahrnuje preference a nějakou *teorii užitku*, která preference dovoluje vyjadřovat a pracovat s nimi.

Pravděpodobnosti

- Tvrzení o pravděpodobnostech se týkají možných světů jako prvků vzorkovacího prostoru, který možné světy pokrývají úplně (vždy nějaký možný svět stavu odpovídá) a výlučně (dva možné světy nemohou platit současně).
- Pravděpodobnost je přiřazena každému možnému stavu (možnému světu) jako nezáporné číslo, s celkovou sumou rovnou 1 (resp. 100 %).
- Množiny možných světů odpovídají *událostem* (podobně jako v logice odpovídají propozicím), události bez další struktury nazýváme *atomickými*.
- Pravděpodobnost každé atomické události představuje *(úplná) sdružená distribuce*, tj. vektor pravděpodobností této události v závislosti na jiných parametrech; značeno tučně jako **P**(A).
- S událostmi formálně pracujeme jako s proměnnými nabývajícími hodnot příslušných pravděpodobností, P(A) je pravděpodobnost jevu A.
- Podmíněná (posteriorní) pravděpodobnost P(A|B) je pravděpodobnost události A za předpokladu, že nastala událost B
- Podmíněnou pravděpodobnost definujeme na základě nepodmíněné jako
  - $P(A|B) = P(A \wedge B)/P(B)$ (pro $P(B)>0$)

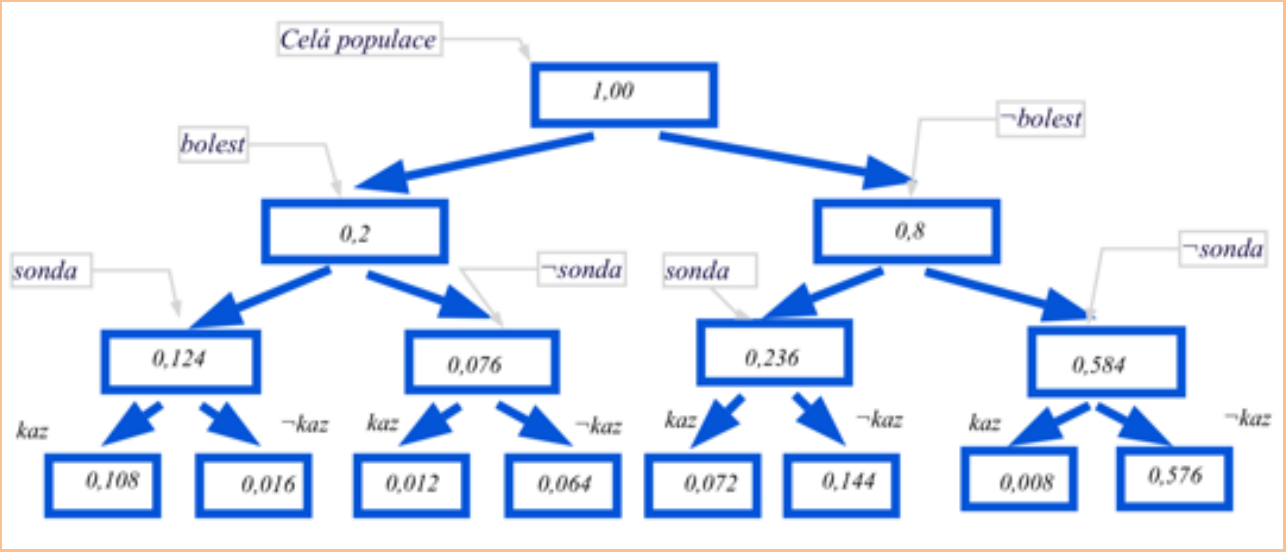
Příklad – dentistův svět

- Příznak *bolesti* (zubů), výsledek vyšetření zubařskou *sondou*, problém výskytu *kazu* v zubu
- Vazby hodnot pravděpodobností dány sdruženou distribucí

	bolest		*¬bolest*	
	sonda	*¬sonda*	*sonda*	*¬sonda*
kaz	0.108	0.012	0.072	0.008
¬kaz	0.016	0.064	0.144	0.576

Úplná sdružená distrubuce pro dentistův svět s událostmi *bolest*, *kaz*, *sonda*.

- Totéž v podobě frekvenční tabulky:



Příklad – dentistův svět

- Příznak *bolesti* (zubů), výsledek vyšetření zubařskou *sondou*, problém výskytu *kazu* v zubu
- Vazby hodnot pravděpodobností dány sdruženou distribucí

	bolest		*¬bolest*	
	sonda	*¬sonda*	*sonda*	*¬sonda*
kaz	0.108	0.012	0.072	0.008
¬kaz	0.016	0.064	0.144	0.576

Úplná sdružená distrubuce pro dentistův svět s událostmi *bolest*, *kaz*, *sonda*.

- Je zde 6 možných světů, kde platí disjunkce *kaz* v *bolest*
  - P(*kaz* v *bolest*) = 0,108+0,012+0,072+0,008+0,016+0,064 = 0,28
- Příklad výpočtu nepodmíněné (marginální) pravděpodobnosti výskytu kazu v populaci:
  - P(*kaz*) = 0,108+0,012+0,072+0,008 = 0,2
- Vysčítáním hodnost v tabulkách distribucí lze řešit zjišťování pravděpodobností příslušných proměnné bez ohledu na hodnoty jiné sady proměnných nebo zjišťování podmíněné pravděpodobnosti
- **P**(*kaz*)=**P**(*kaz*,*bolest*,*sonda*)+**P**(*kaz*,*bolest*, ¬*sonda*)+**P**(*kaz*, ¬*bolest*,*sonda*)+**P**(*kaz*, ¬*bolest*, ¬*sonda*)=<0,108;0,016>+<0,012;0,064>+<0,072;0,144>+<0,008;0,576>=<0,2;0,8>

Bayesovo pravidlo

- Pro podmíněné pravděpodobnosti platí rovnosti
  - $P(a \wedge b) = P(a|b)P(b)$ a také $P(a \wedge b) = P(b|a)P(a)$
- Odtud Bayesovo pravidlo
  - $P(b|a) = P(a|b)P(b)/P(a)$
- resp. obecněji pro distribuce
  - $\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\mathbf{P}(Y)/\mathbf{P}(X)$

- Bayesovo pravidlo lze dobře aplikovat na popis kauzálních a diagnostických vazeb:

- P(*následek*|*příčina*) = P(*příčina*|*následek*)P(*příčina*) / P(*následek*)

- P(*příčina*|*následek*) = P(*následek*|*příčina*)P(*následek*) / P(*příčina*)

- Užíváno pro diagnostiku P(*nemoc*|*symptomy*) na základě toho, že známe
  - pravděpodobnost nemoci P(*nemoc*)
  - pravděpodobnost symptomů P(*symptomy*)
  - kauzální vztah mezi nemocí a symptomy P(*symptomy*|*nemoc*)

Diagnostika v medicíně

- Příklad: Pro ženu ve stáří 40 let, které chodí na pravidený screening, je pravděpodobnost 1 percento, že má rakovinu prsu. Pokud má žena rakovinu prsu, je pravděpodobnost 80 %, že bude mít pozitivní výsledek mamografie. Pokud žena rakovinu prsu nemá, je pravděpodobnost 9.6 %, že bude mít také pozitivní mamografii. Žena v této věkové skupině dostala při pravidelném screeningu pozitivní výsledek mamografie. Jaká je pravděpodobnost, že trpí rakovinou prsu?

  $P(rp)=0,01$

  $P(m|rp)=0,8$

  $P(m|\neg rp)=9,6$

  Bayesovo pravidlo se zde užije ve tvaru

  $P(rp|m)=P(m|rp)P(rp)/(P(m|rp)P(rp)+P(m|\neg rp)P(\neg rp))$

  $P(rp|m]= (0,01x0,8)/((0,01x0,8)+(0,99x0,096))=0,078.$

Diagnostika v medicíně

- Příklad: Pro ženu ve stáří 40 let, které chodí na pravidený screening, je pravděpodobnost 1 percento, že má rakovinu prsu. Pokud má žena rakovinu prsu, je pravděpodobnost 80 %, že bude mít pozitivní výsledek mamografie. Pokud žena rakovinu prsu nemá, je pravděpodobnost 9.6 %, že bude mít také pozitivní mamografii. Žena v této věkové skupině dostala při pravidelném screeningu pozitivní výsledek mamografie. Jaká je pravděpodobnost, že trpí rakovinou prsu?

  P(rp)=0,01

  P(m|rp)=0,8

  P(m|¬rp)=9,6

  Bayesovo pravidlo se zde užije ve tvaru

  P(rp|m)=P(m|rp)P(rp)/(P(m|rp)P(rp)+P(m|¬rp)P(¬rp))

  P(rp|m]= (0,01x0,8)/((0,01x0,8)+(0,99x0,096))=0,078.

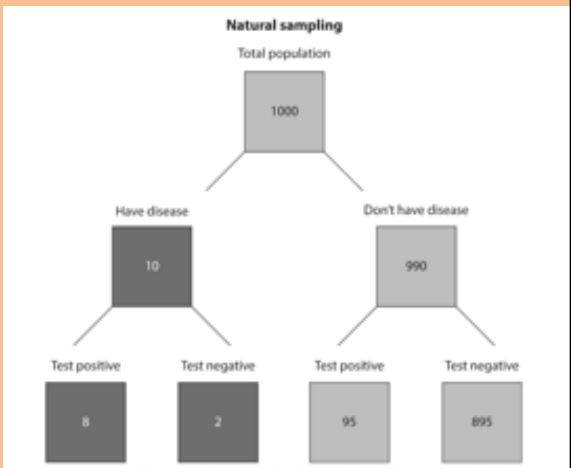


FIGURE 5.3 Visualizing Probabilities in Cancer Testing

Source: Gerd Gigerenzer and Ulrich Hoffrage, "How to improve Bayesian Reasoning without instruction: Frequency formats," *Psychological Review* 2, no. 4 (1995):684–704.

Bayesovské sítě

- Bayesovská síť je orientovaný acyklický graf, v němž je každý uzel ohodnocen kntitativní pravděpodobnostní informací
  - Uzly odpovídají náhodnostní proměnné (diskrétní nebo spojité)
  - Orientované hrany spojují rodičovský uzel s uzlem potomka
  - Každý uzel má přiřazenu pravděpodobnostní informaci s tabulkou podmíněné pravděpodobnostní distribuce **P**(X|*rodič*(X))
- Konstrukce sítě umožňuje dosáhnout kompaktnější reprezentace než u úplných sdružených distribucí.

- Užití systémů založených na Bayesovských sítích se datuje od počátku 90. let 20. století, s prvním medicinským systémem MUNIN pro diagnostiku neuromuskulárních poruch z roku 1989 a PATHFINDER pro patologii z roku 1991. V průběhu 90. let se začaly objevovat i technické/inženýrské palikace (monitoring generátorů elektřiny, 1995, moduly pro diagnostiku a opravu v Microsoft Windows, 1996 a 1998). Analýza rodokmenů/DNA se datuje od cca 2000.
- Kromě počítání marginálních pravděpodobností se Bayesovské sítě používají i pro vytváření nejpravděpodobnějších vysvětlení „Most probable explanations“)

Teorie užitku

- Pravděpodobnost se užívá k popisu toto, v co má aktér věřit na základě vjemů, které mu zprostředkovávají čidla
- *Teorie užitku* (utility) popisuje, co aktér chce

- **Princip maximálního očekávaného užitku (MEU)** stanoví, že aktéři vybírají akce, které maximalizují očekávaný užitek
- Tento princip formalizuje „nejlepší očekávané akce“, ale nestanoví postup, jak jich dosáhnout
- Obecně postup dosažení takové akce vyžaduje vnímání vstupů ze senzorů, učení, reprezentaci znalostí a učení.

- MEU lze axiomatizovat na základě uspořádání aktérových preferencí do struktrury umožňující porovnávání, zachovávající tranzitivitu, platí v něm monotonicita a spojitost a v případě složených systémů jsou rozložitelné na jednodušší pomocí zákonů pravděpodobnosti

- Pro aktérovy preference splňující axiomatický systém lze dokázat existenci užitkové funkce odpovídající preferencím a také to, že pro racionálního aktéra existuje nějaká užitková fukce takových vlastností.

Funkce užitku

- Funkce užitku v systémech týkajících se medicíny, dopravy, životního prostředí a dalších zahrnují i otázky ovlivňující život lidí.
- Absolutní preference ochrany životů je nereálná: letadla procházejí údržbou v nějakých intervalech, auta se konstruují se zřetelem na cenu produkce (nejen na to, jak chrání osádku pro nehodě), atd.
- Odhady statistické hodnoty života vycházejí z ceny souvisejících regulací na ochranu, odhady typických hodnot z USA z roku 2019 se pohybují kolem 10 milionů USD.
- „Mikromort“ jako jednotka v medicinských aplikacích a analýze rizik – riziko úmrtí v poměru 1 ku milionu: typické hodnoty ze Spojeného království jsou cca 1 mikromort na cca 340 km ujetých v autě; během životnosti auta (cca 150 000 km) jde o cca 400 mikromortů; z dat o ochotě připlácet za bezpečnostní prvky vyplývá ochota platit cca USD 60 za mikromort, atd.

Funkce užitku a konsekvencionalismus

- S ohledem na zájmy lidských uživatelů je nosným směrem úvah tzv. *konsekvencialismus*: myšlenka, že volby mají být posuzovány podle očekávaných důsledků (konsekvencí)
- Pohled uplatňovaný na lidské aktéry zahrnuje deontologická etika a etika ctnosti, které se zabývají morální povahou konání, respektive morální povahou jednotlivců, zcela odděleně od důsledků voleb, které dělají
- Teorém sociální agregace: aktér jednající jménem populace jednotlivců musí maximalizovat váženou lineární kombinaci užitků těchto jednotlivců.
- Pro konsekvencionalistické stroje je otázkou, jak hodnotit důsledky, které zasahují více lidí. Jednou přijatelnou odpovědí je dát stejnou váhu preferencím každého – jinými slovy maximalizovat součet užitků všech (Jeremy Bentham a John Stuart Mill jako autor filozofického přístupu k utilitarismu).

Porovnávání užitku

- Užitek je dobře definován vzhledem k jednomu aktérovi
- Srovnání užitků mezi jednotlivými jedinci a srovnání napříč různě velkými populacemi je problematické
- William Stacey Jevons (1871): „Citlivost jedné mysli, jak víme, může být tisíckrát větší než mysli jiné. Ale za předpokladu, že by se citlivost lišila ve stejném poměru ve všech směrech, nikdy bychom nedokázali objevit žádný hluboký rozdíl. Každá mysl je tudíž pro každou jinou mysl nevyzpytatelná a není možný žádný společný jmenovatel pocitů.“
- Kenneth Arrow, (moderní teorie společenské volby, Nobelista z roku 1972): „Budeme vycházet z pohledu, že srovnání užitku mezi různými jedinci nemá smysl a že ve skutečnosti neexistuje žádný relevantní význam ke srovnání blahobytu v tom, jak jsou užitky jednotlivců měřeny.“

Užitek přes velké populace

- Robert Nozick (1974): i kdyby bylo srovnání užitku mezi různými jedinci možné, byla by maximalizace součtu užitků stále ještě chybnou myšlenkou, protože by se dostala do konfliktu s *monstrem užitku* – osobou, jejíž zakoušení požitku a bolesti jsou mnohonásobně intenzivnější než u obyčejných lidí. Taková osoba by mohla prosazovat, že jakékoli množství zdrojů dává větší příspěvek k sumě celkového lidského štěstí, pokud je alokováno v její prospěch, nikoli ve prospěch jiných; a tedy by byl dobrý nápad i odebrat zdroje jiným ve prospěch monstra užitku.
- Henry Sidgwick (1874): správnou volbou je upravovat velikost populace až do té doby, než bude dosaženo maximálního štěstí (nikoli zvětšování populace nade všechny meze, protože by v nějaké chvíli všichni hladověli, a tudíž byli dost nešťastní).
- Derek Parfit (1984): v libovolné situaci s N velmi šťastnými lidmi existuje (podle utilitariánských principů) upřednostnitelná situace s 2^N lidmi, kteří jsou vždycky nepatrně méně šťastní. Opakování tohoto procesu dosáhne tzv. „*Odpudivého závěru*“, že nejvíce žádoucí situace je ta s obrovskou populací, v níž všichni mají život, který téměř nestojí za žití.

Teorie rozhodování

- Kombinace teorie pravděpodobnosti a teorie užitku umožňující popisovat, co má aktér dělat
- Systémy založené na zkoumání možných akcí a volbě takové, co vede k nejlepšímu výsledku, se nazývají *racionální aktéři*
- Teorie užitku ukazuje, jak racionálního aktéra popsat pomocí užitkové funkce splňující axiomatický systém – takový aktér maximalizuje očekávaný užitek
- Rozšířením bayesovských sítí lze získat tzv. *rozhodovací sítě*, v nichž je k uzlům odpovídajícím pravděpodobnostem přidáno i rozhodování a uzly vyhodnocující užitek
- Hodnota informace reprezentuje očekávané zlepšení užitku ve srovnání činnosti bez této informace – přidání činností shromažďujících takové informace před děláním rozhodnutí je další z doplňků řídicích procesů v rámci teorie rozhodování
- Existují situace, kdy odpovídající funkci užitku odpovídající lidskému uživateli nejde stanovit – rozhodování proto musí odpovídat činnosti za podmínek omezené určitosti informací (náhodnostní proměnné) a vhodné nastavení apriorních hodnot pravděpodobnosti odpovídá takovému rozhodování

Rozhodování více aktérů

- Více aktérů s jedním místem, kde se dělají rozhodnutí (ostatní aktéři plní jeho pokyny) – úlohy s více aktuátory (např. různé souběžné činnosti) nebo decentralizované plánování
- U více aktérů vykonávajících rozhodování se může jednat
  - o situaci se společným cílem všech aktérů (řešících převážně jen problém koordinace), nebo
  - mají aktéři vlastní preference, které mohou být v rozporu s ostatními (hra s nulovým součtem, atd.)
- Teorie her poskytuje různé rámce postihující charakter vykonávaných činností, zejména pak dělení na *kooperativní hry* předpokládající nějakou formu závazné dohody aktérů a její naplňování, nebo *nekooperativní hry*, v nichž nemusí jít přímo o protivníky, pouze může chybět koordinace aktérů.
- Plánování činností zahrnujících víc aktérů musí jejich činnosti brát v potaz, jak interagují s činností jiných aktérů