Dobývání znalostí

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Dobývání znalostí

– Pravděpodobnost a učení –

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Pravděpodobnost – základní pojmy (1)

Pravděpodobnost (jevu *A* z prostoru *S*):

- $P(A) \geq 0$ ($P(\{\}) = 0$)
- $P(S) = 1$
- Pro konečný počet navzájem neslučitelných jevů $A_1, A_2, \ldots, A_n$ je pravděpodobnost

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

- Pro nekonečně mnoho navzájem neslučitelných jevů $A_1, A_2, \ldots, A_n$ je pravděpodobnost

$$P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

Pravděpodobnost – základní pojmy (2)

- **Podmíněná pravděpodobnost** jevu ***B*** za předpokladu, že nastal jev ***A*** (***P(A) > 0***):

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

- **Vzájemná nezávislost** jevu ***A*** a jevu ***B:***

$$P(A \cap B) = P(A) \cdot P(B)$$

- **Vzorec pro úplnou pravděpodobnost**:

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

Pravděpodobnost – základní pojmy (3)

Bayesův vzorec pro podmíněnou pravděpodobnost:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)} \; ; \qquad P(A), P(B) > 0$$

- **Náhodná veličina:**
  - ´jméno experimentu s pravděpodobnostním výsledkem´
  - Její hodnota odpovídá výsledku experimentu
- **Rozložení pravděpodobnosti (pro náhodnou veličinu *Y*):**
  - Pravděpodobnost $P(Y = y_i)$, že Y bude mít hodnotu y_i
- **Střední hodnota (náhodné veličiny *Y*):**

$$\mu_Y = E(Y) = \sum_i y_i \, P(Y = y_i)$$

Pravděpodobnost – základní pojmy (4)

- **Rozptyl (náhodné veličiny):**

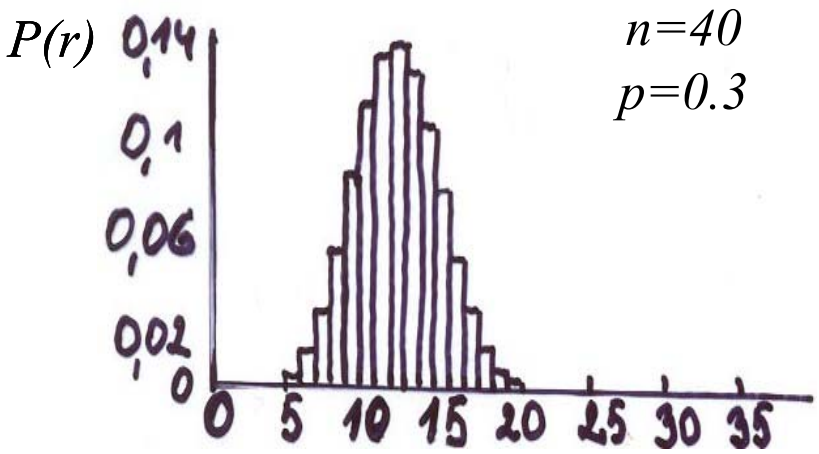
$$VAR\ (Y) \quad = \quad E\left[(Y - \mu_Y)^2 \right]$$

  - Vyjadřuje šířku (disperzi) rozložení kolem střední hodnoty
- **Směrodatná odchylka *Y*:** $\sigma_Y = \sqrt{VAR(Y)}$
- **Binomické rozložení**
  - Pravděpodobnost výskytu ***r*** ´orlů´v posloupnosti ***n*** nezávislých hodů mincí
  - Pravděpodobnost ´orla´ v jednom hodu je ***p***

Pravděpodobnost – základní pojmy (5)

Binomické rozložení

- Pravděpodobnost výskytu ***r*** ´orlů´v posloupnosti ***n*** nezávislých hodů mincí
- Pravděpodobnost ´orla´ v jednom hodu je ***p***



- **Frekvenční funkce** (rozložení pravděpodobnosti)

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Pravděpodobnost – základní pojmy (6)

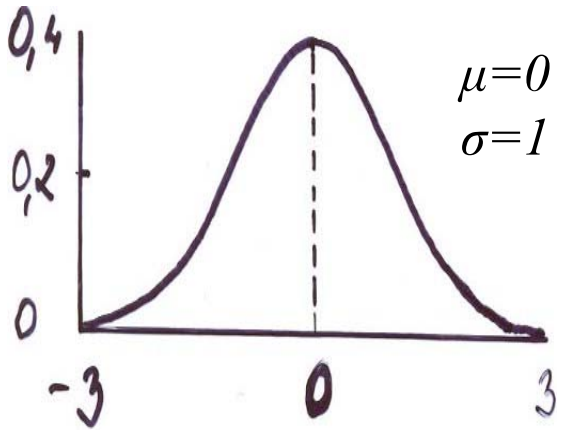
- Střední hodnota náhodné veličiny X: $E[X] = np$
- Rozptyl: $VAR(X) = np(1-p)$
- Směrodatná odchylka: $\sigma_X = \sqrt{n\ p\ (1-\ p\ )}$
- Pro dostatečně velké hodnoty n lze binomické rozložení aproximovat normálním rozložením se stejnou střední hodnotou a rozptylem
- **Doporučení:** aproximaci normálním rozložením použít jen pokud: $np(1-p) \geq 5$

Pravděpodobnost – základní pojmy (7)

Normální rozložení

- Alternativní označení – **Gaussovo rozložení**
- **Hustota normálního rozložení**

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- Pravděpodobnost, že hodnota náhodné veličiny X bude z intervalu (a, b):

$$\int_a^b p(x)\,dx$$

Pravděpodobnost – základní pojmy (8)

Normální rozložení

- Vyhovuje velkému množství přirozených jevů
- Střední hodnota náhodné veličiny X: $E[X] = \mu$
- Rozptyl: $VAR(X) = \sigma^2$
- Směrodatná odchylka: $\sigma_X = \sigma$

- **Centrální limitní věta:**

  ´Rozložení velkého počtu nezávislých náhodných veličin se stejným rozložením aproximuje normální rozložení´

Pravděpodobnost – základní pojmy (9)

- **Odhad** ~ náhodná veličina ***Y***
  - Slouží k odhadnutí parametru ***p*** z testované populace
- **Práh odhadu** ***Y*** pro parametr ***p*** : $E[Y] - p$
  - ´bezprahový´ odhad: $E[Y] = p$
- **Interval *N%* spolehlivosti** pro parametr ***p***
  - Interval, který obsahuje ***p*** s pravděpodobností ***N%***
- **Test** ~ postup, kterým rozhodujeme o správnosti statistické hypotézy ***H***
  - **Hladina významnosti** α odpovídá pravděpodobnosti zamítnutí správné hypotézy → obvykle se volí $\alpha = 0.05$

Vyhodnocování hypotéz (1)

1. **Známe správnost hypotézy pozorované na omezeném vzorku dat → jak dobrý je tento odhad správnosti na dalších vzorech?**
2. **Víme, že je jedna hypotéza na nějakém vzorku dat lepší než druhá → jaká je pravděpodobnost, že tato hypotéza bude lepší v obecném případě?**
3. **Máme k dispozici omezené množství dat → jak jich využít co možná nejlépe pro naučení hypotézy i pro odhad její správnosti a porovnat správnost dvou algoritmů učení?**

   → omezit rozdíl mezi správností pozorovanou na daném vzorku a skutečnou správností na celém rozložení dat

Vyhodnocování hypotéz (2)

Cíl: 1) Vyhodnotit, zda hypotézu použít nebo ne

2) Vyhodnocování hypotéz je součástí nejrůznějších metod učení (např. prořezávání rozhodovacích stromů)

Odhad obecné správnosti hypotézy naučené na omezeném vzorku dat:

- **Práh odhadu**: přeučení × bezprahový odhad budoucí správnosti (navzájem nezávislé trénovací a testovací množina)
- **Rozptyl odhadu**: naměřené správnost se může lišit od skutečné; větší rozptyl pro méně testovacích vzorů

Vyhodnocování hypotéz (3)

Odhad správnosti hypotézy

- Množina možných případů X, např. množina všech lidí
- Na této množině lze definovat různé cílové funkce $f: X \rightarrow \{0,1\}$, např. lidé, kteří by si letos chtěli koupit nové lyže
- Různé případy $x \in X$ se vyskytují s různou frekvencí, např. pravděpodobnost, že x pojede na hory
  - D … pravděpodobnost výskytu případů v X

Vyhodnocování hypotéz (4)

Úloha: nalézt cílovou funkci f z množiny možných hypotéz H

- k dispozici jsou trénovací vzory x spolu se správnou hodnotou cílové funkce $f(x)$, vybrané z X nezávisle s pravděpodobností D

Otázky:

- Pro hypotézu h a vzorek dat obsahující n případů vybraných náhodně s pravděpodobností D :

1. **Jak nejlépe odhadnout správnost h pro další vzory vybrané se stejnou pravděpodobností?**
2. **Jak dobrý (přesný) je tento odhad správnosti?**

Vyhodnocování hypotéz (5)

Chyba na trénovací množině $S \subset X$

~ podíl vzorů z S, které hypotéza h klasifikuje chybně

$$ERROR_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

- n … počet vzorů v S
- $\delta(f(x), h(x)) = 1$ pro $f(x) \neq h(x)$
- $\delta(f(x), h(x)) = 0$ pro $f(x) = h(x)$
- binomické rozložení $ERROR_S(h)$: $ERROR_S(h) = r / n$
  - r … počet vzorů z S, které byly hypotézou h klasifikovány chybně

Vyhodnocování hypotéz (6)

Skutečná chyba hypotézy ***h***

~ pravděpodobnost chybné klasifikace pro jeden vzor $x \in X$ vybraný s pravděpodobností D

$$ERROR_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$$

- binomické rozložení: $ERROR_D(h) = p$ ($= r/n$ … odhad pro p)
  - p … pravděpodobnost chybné klasifikace pro jeden vzor vybraný s pravděpodobností D
  - bezprahový odhad $ERROR_D(h)$ ($\sim p = r - n$)
    - Potřebná nezávislost hypotézy h a trénovací množiny S
    - Trénovací množina S obsahuje n (≥ 30) vzorků vybraných z X podle pravděpodobnosti D

Vyhodnocování hypotéz (7)

Rozptyl odhadu

- Bezprahový odhad s nejmenším rozptylem by dával nejmenší střední kvadratickou chybu mezi odhadovanou a skutečnou hodnotou parametru
- Pokud nemáme k dispozici jinou informaci, je nejpravděpodobnější hodnotou $ERROR_D(h)$ hodnota $ERROR_S(h)$
- S pravděpodobností zhruba 95% leží skutečná hodnota $ERROR_D(h)$ v intervalu

$$ERROR_S(h) \pm 1.96 \sqrt{\frac{ERROR_S(h)\,(1 - ERROR_S(h))}{n}}$$

→ v 95% experimentů bude skutečná hodnota chyby spadat do vypočteného intervalu

Vyhodnocování hypotéz (8)

Výpočet pro obecný (*N%*) interval spolehlivosti – konstanta z_N:

$$ERROR_S(h) \pm z_N \sqrt{\frac{ERROR_S(h)\,(1 - ERROR_S(h))}{n}}$$

- TABULKA: HODNOTY z_N PRO OBOUSTRANNÉ N%-NÍ INTERVALY SPOLEHLIVOSTI

N%	50%	68%	80%	90%	95%	98%	99%
z_N	0,67	1,00	1,28	1,64	1,96	2,33	2,58

- Větší interval pro vyšší pravděpodobnost
- Dobrá aproximace pro $n \geq 30$, resp.

$$n \cdot ERROR_S(h)\,(1 - ERROR_S(h)) \geq 5$$

Vyhodnocování hypotéz (9)

Obecný postup pro odvození intervalu spolehlivosti:

1. **Identifikace parametru** p, který je třeba odhadnout ($ERROR_D(h)$)
2. **Definice odhadu** Y (např. $ERROR_S(h)$)

   – je vhodné volit bezprahový odhad s minimálním rozptylem
3. **Určit pravděpodobnostní rozložení** D_Y pro odhad Y včetně střední hodnoty a rozptylu
4. **Určit** $N\%$ **-ní interval spolehlivosti**

   – nalézt meze L a U tak, aby $N\%$ případů vybraných s pravděpodobností D_Y padlo mezi L a U

Vyhodnocování hypotéz (10)

Porovnání dvou hypotéz:

- Diskrétní hodnoty cílové funkce
- Hypotéza h_1 byla testována na množině S_1 n_1 náhodně zvolených vzorů
- Hypotéza h_2 byla testována na množině S_2 n_2 náhodně zvolených vzorů
- **Chceme odhadnout rozdíl d mezi skutečnými chybami těchto dvou hypotéz:**

$$d = ERROR_D(h_1) - ERROR_D(h_2)$$

Vyhodnocování hypotéz (11)

→ Odhad $\hat{d}$ ~ rozdíl chyb na testovaných datech:

$$\hat{d} \equiv ERROR_{S_1}(h_1) - ERROR_{S_2}(h_2)$$

$\hat{d}$ je bezprahový odhad

- Normální rozložení s $E\left[\hat{d}\right] = d$ a rozptylem $\sigma_{\hat{d}}^2$

$$\sigma_{\hat{d}}^2 \approx \frac{ERROR_{S_1}(h_1)(1-ERROR_{S_1}(h_1))}{n_1} + \frac{ERROR_{S_2}(h_2)(1-ERROR_{S_2}(h_2))}{n_2}$$

- ***N%*** -ní interval spolehlivosti:

$$\hat{d} \pm z_N \sqrt{\frac{ERROR_{S_1}(h_1)(1-ERROR_{S_1}(h_1))}{n_1} + \frac{ERROR_{S_2}(h_2)(1-ERROR_{S_2}(h_2))}{n_2}}$$

Vyhodnocování hypotéz (12)

Porovnání algoritmů učení:

- test pro porovnání algoritmu učení L_A a L_B
- statistická významnost pozorovaného rozdílu mezi algoritmy

→ chceme určit, který z algoritmů L_A a L_B je lepší pro učení hledané funkce f

- Uvažovat průměrnou správnost obou algoritmů na všech možných trénovacích množinách velikosti n, které lze vytvořit pro rozložení D

Vyhodnocování hypotéz (13)

Porovnání algoritmů učení:

→ **odhad střední hodnoty rozdílu chyb**

$$\underset{S \subset D}{E}\left[ERROR_D(L_A(S)) - ERROR_D(L_B(S))\right]$$

$L(S)$ … hypotéza získaná pomocí algoritmu učení ***L*** na trénovací množině ***S***

$S \subset D$ … střední hodnota se počítá přes vzorky vybrané podle rozložení ***D***

→ **v praxi je pro porovnání metod učení k dispozici omezené množství trénovacích dat** D_0

Vyhodnocování hypotéz (14)

- Rozdělit množinu D_0 na trénovací množinu S_0 a testovací množinu T_0, které jsou navzájem disjunktní
  - Trénovací vzory se použijí při učení L_A a L_B
  - Testovací vzory se použijí k vyhodnocení správnosti naučených hypotéz:

$$ERROR_{T_0}(L_A(S_0)) - ERROR_{T_0}(L_B(S_0))$$

  - Chybu $ERROR_D(h)$ aproximuje chyba $ERROR_{T_0}(h)$
  - Chyba se měří pro jednu trénovací množinu S_0 (a nikoliv jako střední hodnota rozdílu přes všechny možné vzorky S vybrané podle rozložení D)

k-násobná křížová validace (1)

1. Rozděl trénovací data D_0 do k navzájem disjunktních podmnožin $T_1, T_2, \ldots, T_k$ stejné velikosti (≥ 30).

2. **FOR i:=1 TO k DO**

   použij T_i jako testovací množinu, zbylá data použij k vytvoření trénovací množiny S_i

   $S_i \leftarrow \{ D_0 \setminus T_i \}$

   $h_A \leftarrow L_A(S_i)$

   $h_B \leftarrow L_B(S_i)$

   $\delta_i \leftarrow ERROR_{T_i}(h_A) - ERROR_{T_i}(h_B)$

3. Vrať hodnotu $\overline{\delta}$, kde: $\overline{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$

k-násobná křížová validace (2)

N % **-** ní interval spolehlivosti: $\bar{\delta} \pm t_{N,k=1}\, s_{\bar{\delta}}$

$s_{\bar{\delta}}$ … odhad směrodatné odchylky:

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} \left(\delta_i - \bar{\delta}\right)^2}$$

$t_{N,k-1}$ … konstanta (hodnoty $t_{N,v}$ pro oboustranné intervaly spolehlivosti pro $v \to \infty$ se $t_{N,v}$ blíží z_N)

N …… požadovaná úroveň spolehlivosti

v ……. počet stupňů volnosti (počet navzájem nezávislých náhodných událostí uvažovaných při výpočtu $\bar{\delta}$)

k-násobná křížová validace (3)

	ÚROVEŇ SPOLEHLIVOSTI N			
	90 %	95 %	98 %	99 %
$\nu = 2$	2,92	4,30	6,96	9,92
$\nu = 5$	2,02	2,57	3,36	4,03
$\nu = 10$	1,81	2,23	2,76	3,17
$\nu = 20$	1,72	2,09	2,53	2,84
$\nu = 30$	1,70	2,04	2,46	2,75
$\nu = 120$	1,66	1,98	2,36	2,62
$\nu = \infty$	1,64	1,96	2,33	2,58

N … požadovaná úroveň spolehlivosti

ν … počet stupňů volnosti

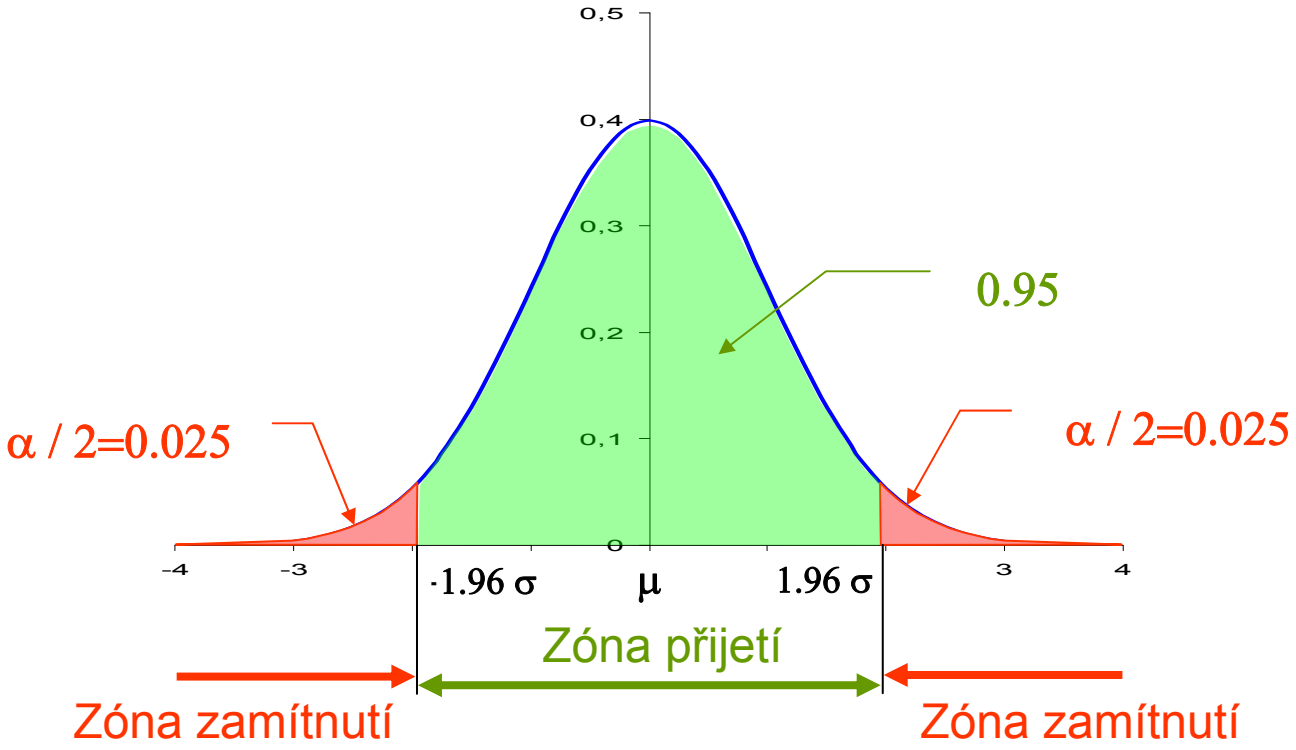
k-násobná křížová validace (4)

- **Testování je třeba provádět na identických testovacích množinách!**
  - na rozdíl od porovnávání hypotéz, které vyžaduje nezávislé množiny

→ **Párové testy**

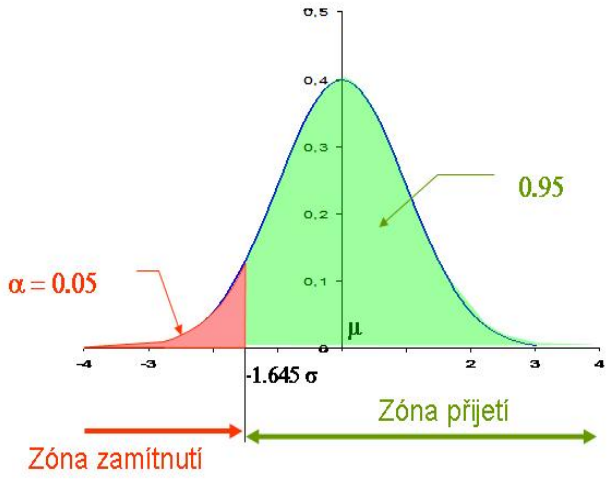
- dávají užší intervaly spolehlivosti, protože rozdíly v pozorovaných chybách vznikají kvůli rozdílům v hypotézách, ne kvůli rozdílům v datech

Oboustranný test

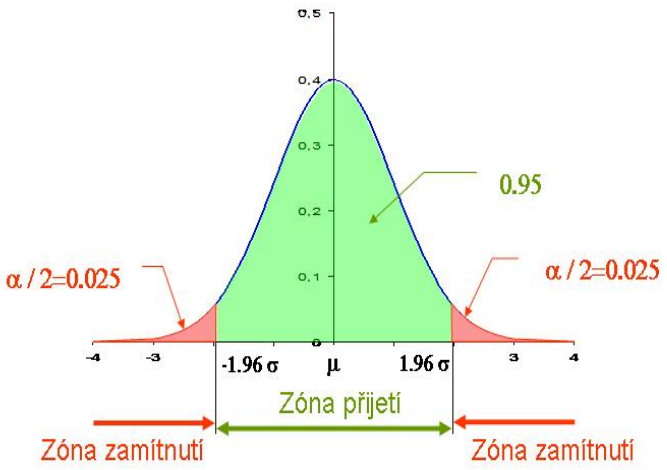


Jednostranný × oboustranný test

Jednostranný test



Oboustranný test



Adaptace a učení

Adaptace:

- schopnost přizpůsobit se změnám okolního prostředí

Adaptivní proces: proces přizpůsobení

- každá adaptace představuje pro systém jistou ztrátu (materiál, energie, …)
- živé organismy jsou schopné tyto ztráty při mnohonásobném opakování adaptace na určitou změnu prostředí zmenšovat

U Č E N Í :

- minimalizace ztrát vynaložených na adaptaci
- výsledek mnohonásobného opakování adaptace

Adaptace a učení: formalismus (1)

- **Projev prostředí: $\boldsymbol{x}$**
- **Příznakový popis předmětů:**
  - výběr n elementárních vlastností – příznaků $x_1, \ldots, x_n$
  - $\boldsymbol{x} = (x_1, \ldots, x_n)$
- **Informace o požadovaném chování systému (reakci) na projev prostředí: Ω**
- Systém reaguje na libovolný projev prostředí $\boldsymbol{x}$ a informaci Ω tak, že na výstupu vydá jeden ze symbolů $\omega_r ; r = 1, \ldots, R$

Adaptace a učení: formalismus (2)

- Každé přiřazení $[\boldsymbol{x}, \Omega] \rightarrow \omega_r$ doprovází jistá ztráta daná funkcí $Q(\boldsymbol{x}, \Omega, \omega_r)$ za časovou jednotku
- **Cíl systému:**
  - najít pro každé $\boldsymbol{x}$ a Ω takové přiřazení

$$[\boldsymbol{x}, \Omega] \rightarrow \omega_r ,$$

pro které je **ztráta minimální:**

$$Q(x, \Omega, \omega_r) = \min_{\omega} Q(x, \Omega, \omega)$$

Adaptivní systémy (1)

Adaptivní systém

~ systém se dvěma vstupy a jedním výstupem určený:

1) Množinou X **projevů prostředí** $\boldsymbol{x}$
2) Množinou O_1 **informací o požadovaném chování** Ω
3) Množinou O_2 **výstupních symbolů** ω
4) Množinou D **rozhodovacích pravidel** $\omega = d(\boldsymbol{x}, q)$
5) **Ztrátou** $Q(\boldsymbol{x}, \Omega, q)$

Pro každou dvojici $[\boldsymbol{x}, \Omega]$ hledá takový parametr q^*,
při kterém platí: $\boldsymbol{Q(x, \Omega, q^*) = \min_{q} Q(x, \Omega, q)}$

Adaptivní systémy (2)

- Počáteční přiřazení $[\boldsymbol{x}, \Omega] \rightarrow \omega_s$
- Setrvá-li systém po dobu T na počátečním přiřazení, utrpí celkovou ztrátu $T\,Q(\boldsymbol{x}, \Omega, \omega_s)$
- Je-li systém schopen měnit své chování na základě průběžného vyhodnocování ztráty, nalezne **po určité době τ potřebné k vyhodnocení** ω_r, pro které je ztráta minimální

Adaptivní systémy (3)

Celková ztráta za dobu T :

$$\tau\, Q(\boldsymbol{x}, \Omega, \omega_s) + (T - \tau)\, Q(\boldsymbol{x}, \Omega, \omega_r)$$

- je větší než nejmenší možná celková ztráta $T\, Q(\boldsymbol{x}, \Omega, \omega_r)$
- je menší než celková ztráta systému, který nemůže měnit své rozhodnutí, $T\, Q(\boldsymbol{x}, \Omega, \omega_s)$

$$T\, Q(x, \Omega, \omega_r) < \tau\, Q(x, \Omega, \omega_s) + (T - \tau)\, Q(x, \Omega, \omega_r) < T\, Q(x, \Omega, \omega_s)$$

Učící se systémy (1)

Uložení výsledku adaptace do paměti:

- Odstranění doby τ potřebné k nalezení minima ztráty při opakovaném výskytu příslušného projevu prostředí
- Dále nebude třeba vyčíslovat ztráty

  → po naučení není nutná informace Ω o požadovaném chování

Celková ztráta učícího se systému po naučení:

$$T\,Q(\mathbf{x}, \Omega, \omega_r)$$

- je menší než celková ztráta adaptivního systému

Učící se systémy (2)

Učící se systém

~ systém se dvěma vstupy a jedním výstupem určený:

1) Množinou X **projevů prostředí** $\boldsymbol{x}$
2) Množinou O_1 **informací o požadovaném chování** Ω
3) Množinou O_2 **výstupních symbolů** ω
4) Množinou D **rozhodovacích pravidel** $\omega = d\,(\boldsymbol{x}, q)$
5) **Požadovaným chováním** $\Omega = T(\boldsymbol{x})$
6) **Střední ztrátou** $J(q)$ vyčíslenou na $X \times O_1$

Učící se systémy (3)

Učící se systém

- po postupném předložení dvojic z posloupnosti

  $\{ [x_k, \Omega_k] \}; 1 \leq k \leq \infty$, kde $\Omega = T_k (x_k)$,

  nalezne takový parametr q^*, při kterém platí:

  $$J(q^*) = \min_q J(q)$$

- **Sekvenčnost** ~ postupné předkládání dvojic $[x_k, \Omega_k]$
- **Induktivnost** ~ nalézt po prozkoumání spočetně mnoha $[x_k, \Omega_k]$ parametr q^*, který minimalizuje střední ztrátu přes celou X

Efektivnost adaptace a učení

Efektivnost adaptivního systému je tím větší, čím kratší je doba adaptace τ a čím delší jsou časové intervaly T během kterých nedochází ke změnám prostředí:

- $\tau / T \rightarrow 0$:

  Efektivnost AS je porovnatelná s efektivností učícího se systému po naučení
- $\tau / T \rightarrow 1 \quad (\tau / T < 1)$:

  AS je zhruba stejně efektivní jako neadaptivní systém
- $\tau / T \geq 1$: K adaptaci nedochází

Efektivnost učícího se systému (po naučení) je největší možná