Neuronové sítě

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Neuronové sítě

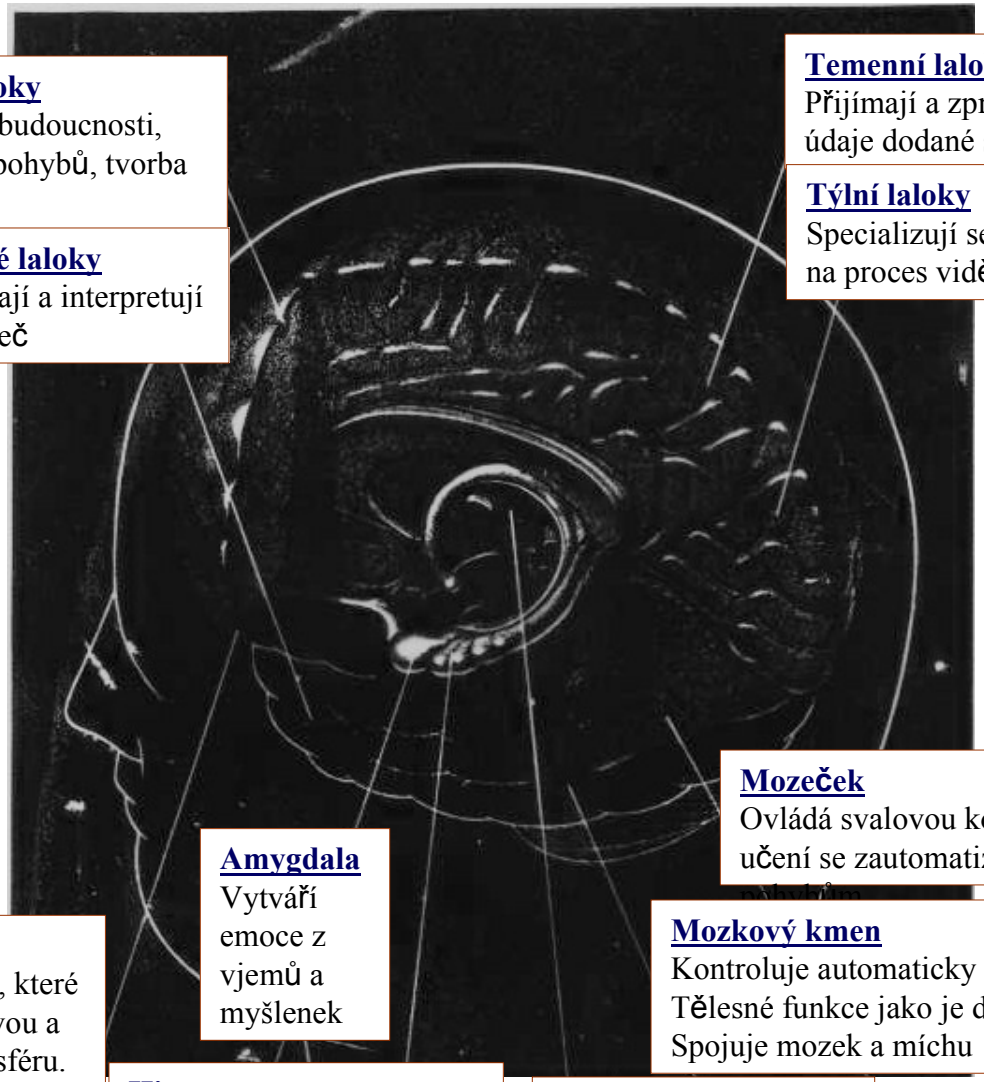
– Úvod do problematiky –

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Počítač versus mozek

- Rychlost zpracování informací
- Způsob zpracování informací
  - Seriově × paralelně
- Způsob ukládání informací
- Redundance
- Řízení



Čelní laloky
Plány do budoucnosti, kontrola pohybů, tvorba řeči

Spánkové laloky
Naslouchají a interpretují hudbu a řeč

Temenní laloky
Přijímají a zpracovávají údaje dodané smysly

Týlní laloky
Specializují se především na proces vidění

Mozeček
Ovládá svalovou koordinaci a učení se zautomatizovaným

Amygdala
Vytváří emoce z vjemů a myšlenek

Mozková kůra
Pokrývá všechny laloky, které dohromady vytvářejí levou a pravou mozkovou hemisféru. Je jen několik milimetrů silná

Mozkový kmen
Kontroluje automaticky konané Tělesné funkce jako je dýchání. Spojuje mozek a míchu

Hippocampus
Upevňuje nedávno nabyté informace, nějakým způsobem mění krátkodobou paměť v dlouhodobou

Thalamus
Přijímá smyslové informace a předává je dál do mozkové kůry

Neuronové sítě – stručně z historie

- 1943 – formální neuron (W. McCulloch, W. Pitts)
- 1949 – matematický pojem učení (D. Hebb)
- 1958 – perceptron (F. Rosenblatt)
- 1962 – Adaline a sigmoidální přenosová funkce (B. Widrow, M. Hoff)
- 1969 – Perceptrony (M. Minsky, S. Papert)
- 80. léta – další rozvoj

Neuronové sítě – stručně z historie

- od 80. let – další rozvoj:
  - Algoritmus zpětného šíření (P. Werbos, D. Rumelhart, G. Hinton, Y. Le Cun)
  - Kohonenovy mapy (T. Kohonen)
  - RBF-sítě (Radial Basis Function, J. Moody, C. Darken)
  - GNG-model (Growing Neural Gas, B. Fritzke)
  - Konvoluční neuronové sítě (Y. Le Cun)
  - SVM-stroje (Support Vector Machines, V. Vapnik)
  - ELM-sítě (Extreme Learning Machines, G.-B. Huang)

Neuronové sítě – obecný úvod

- **Současné problémy:**
  - Strategie učení – paralelizace a efektivita
  - Architektura – generalizace a robustnost
  - Konvergence a přeučení
  - Predikce
- **Použití:**
  - Dobývání znalostí – „black-box“, „white-box“
  - Shlukování a klasifikace
  - Zpracování informací – řeči, vidění, čichu, hmatu, motoriky
  - Komprese dat
  - Řešení optimalizačních úloh
  - a mnoho dalších ….

MF Dnes, září 2010

Úsporný plán: zrušit ve školách deváté třídy

Základní škola by zase mohla mít **jen osm ročníků**. Tedy v případě, že ministr školství prosadí změnu, kterou mu doporučuje nejvýraznější z jeho poradců Václav Klaus mladší. Důvod? Zkrácení doby studia i velká finanční úspora.

PRAHA Zrušte devátou třídu, ušetříte miliardy a na úrovni vzdělání se to neprojeví – nebo se dokonce zvýší. S revolučním doporučením přišel čerstvý poradce ministra školství Václav Klaus mladší, ředitel pražského gymnázia PORG.

Podle něj chybí vzdělávání na druhém stupni základních škol jasná vize: deváťáci neumějí víc, než uměly děti, které vycházely z osmých tříd.

ly. „Samozřejmě by to neslo organizační problémy, ale výsledkem by byla úspora zadarmo," míní Klaus, který svůj nápad poprvé představil v textu pro Lidové noviny.

Poradce poukazuje i na protahování délky vzdělávání. „Středoškolákům je ke dvaceti a stát na to

„Určitě si nechám udělat ekonomický propočet, nakolik by se to vyplatilo."
Josef Dobeš, ministr školství

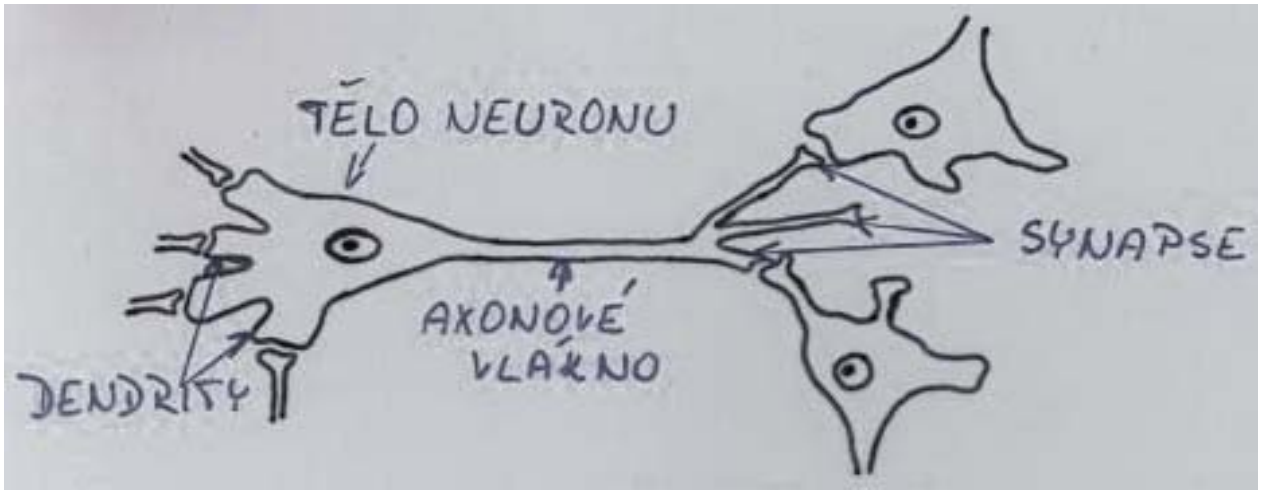
když patnáctileté děti zjistí, že je na střední školu či učiliště bez potíží vezmou, už je těžké přinutit je, aby se učily.

„Pokud má třeba dítě odklad, což je taky stále častější, je plnoleté ve druháku a má pocit, že má všechna práva. Je docela těžké donutit takového studenta k lepšímu studiu, když si sám podepisuje omluvenky a může taky jít třeba do hospody, kdy chce," popisuje důsledky dlouhých studií z druhé strany matka letošního dvacetiletého

Základní biologické poznatky (1)

- **Model neuronu**
  - základní „výpočetní jednotka“ složitějšího celku – neuronové sítě
  - biologický neuron se skládá z:

    těla (somatu), dendritů, axonu a synapsí



Základní biologické poznatky (2)

- **Tělo (soma):**
  - načítá signály předávané okolními neurony → **potenciál**
  - stanovený vnitřní potenciál neuronu vede k **excitaci (vybuzení)** neuronu
  - tělo neuronu má průměr několik μm až několik desítek μm
- **Dendrity:**
  - reprezentují vstup signálů do těla neuronu
  - délka dendritů se pohybuje okolo 2-3 mm

Základní biologické poznatky (3)

- **Axon:**
  - jediný výstup neuronu, který je však na konci bohatě rozvětvený
  - přenáší signál daný stupněm excitace k synapsím
  - délka axonového vlákna může dosahovat i přes 1 m
- **Synapse:**
  - tvoří „výstupní zařízení“ neuronů, která signál zesilují či zeslabují a předávají dalším neuronům
  - na 1 neuron připadá až 10^6 spojů s jinými neurony
- **Výstup neuronu:**
  - závisí na vstupech neuronu a jejich zpracování uvnitř těla neuronu

Základní biologické poznatky (4)

Biologické neuronové sítě:

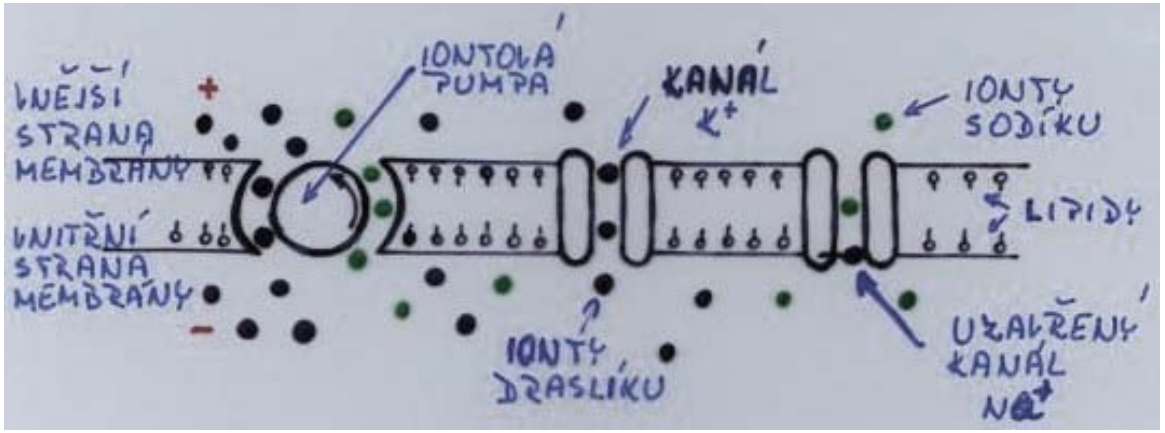
- **neurony jsou vzájemně propojeny do sítí**
  - prostřednictvím axonů, které se pomocí synapsí napojují na dendrity dalších neuronů
- **hustota neuronů:**
  - v lidském mozku dosahuje cca $70 – 80 \cdot 10^3$ / mm³
  - denně odumře cca $10 \cdot 10^3$ neuronů, které už nejsou nahrazeny
  - synapse na dendritech se ale vytvářejí během celého života
    → **vznik nových synapsí**, resp. **oživení dosud nefunkčních synapsí**

=> U Č E N Í

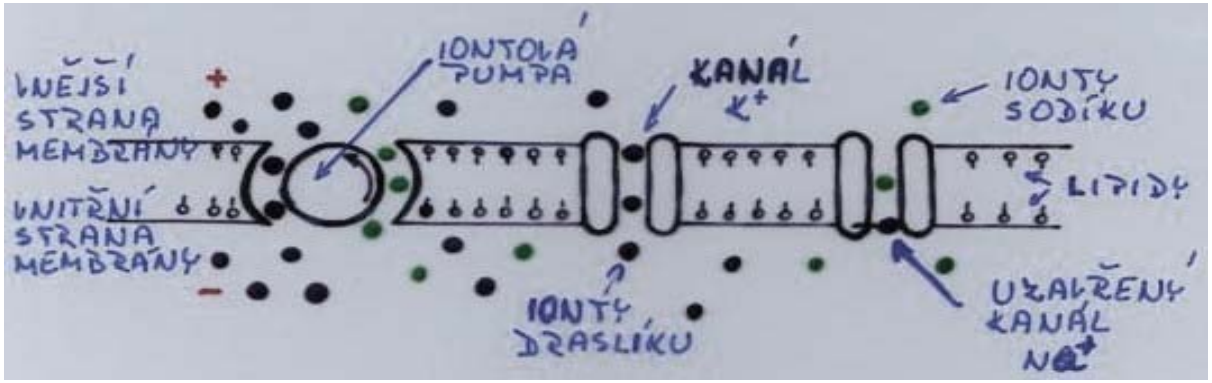
Základní biologické poznatky (5)

Povrch neuronu je pokryt membránou

- umožňuje přenášet informace
- skládá se ze dvou vrstev molekul – tzv. **lipidů**
- mezi vrstvami lipidů jsou ještě vnitromembránové proteiny, které tvoří **iontové pumpy** a **kanály**



Základní biologické poznatky (6)



- **Iontové kanály** - řídí propustnost membrány s ohledem na daný typ iontů
- **Iontové pumpy** - přenášejí přes membránu trvale ionty Na+ a K+
  → tím dochází ke stálé polarizaci membrány:
  - Vnější povrch má kladný potenciál
  - Na vnitřním povrchu je záporný potenciál
  - Rozdíl potenciálů se pohybuje kolem -70 mV

Základní biologické poznatky (7)

- **Dva typy membrán:**
  - **vodivá membrána (~ myelinový povlak)**
    - pokrývá axon
    - v určitých vzdálenostech je přerušována **Ranvierovými zářezy**
      - ty slouží k dosažení vysokých přenosových rychlostí a malého zkreslení přenášených signálů
    - bez myelinového povlaku s Ranvierovými zářezy by se signály šířily až 50-krát pomaleji
  - **transmisní membrána**
    - na rozdíl od vodivé membrány obsahuje navíc receptorové proteiny, které umožňují otvírat nebo zavírat iontové kanály

Základní biologické poznatky (8)

Funkce paměti

- **Krátkodobý paměťový mechanismus**
  - založen na cyklickém oběhu vzruchů v neuronových sítích
  - proběhne-li tato cirkulace cca 300-krát, začne docházet k fixaci informace ve střednědobé paměti – to trvá cca 30 s
- **Střednědobý paměťový mechanismus**
  - založen na změnách „vah neuronů“
  - změna váhových koeficientů v synapsi je vyvolána mnohonásobným působením téhož signálu na příslušných synaptických přechodech

Základní biologické poznatky (9)

Funkce paměti

- **Střednědobý paměťový mechanismus**
  - ve spánku přecházejí některé z takto uchovaných informací do dlouhodobých pamětí
  - informace se uchovává několik hodin a případně i dnů
- **Dlouhodobý paměťový mechanismus**
  - spočívá v kopírování informací ze střednědobé paměti do bílkovin, které jsou uvnitř neuronů – hlavně v jejich jádrech
  - některé takto uchovávané informace zůstanou v organismu celý život

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:

- $\tau / T \to 0$:

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

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

Efektivnost učícího se systému (po naučení) je největší možná

Výběr a uspořádání příznaků

Pravděpodobnost chybného rozhodnutí

×

Množství informace obsažené ve vstupních vzorech

- **Příliš velký počet příznaků:**
  - technická realizovatelnost
  - rychlost zpracování
  - nebezpečí přeučení
    - počet proměnných × počet trénovacích vzorů
  - korelace příznaků

Volba informativních příznaků

- **Výběr minimálního počtu příznaků** z předem zvolené množiny příznaků
  - nelze zaručit, že tato množina obsahuje informativní příznaky
  - volba závisí na konkrétní úloze
- **Uspořádání příznaků** v předem zvolené množině příznaků
  - podle množství nesené informace
  - využití např. u sekvenčních klasifikátorů

Karhunen-Loevovův rozvoj (1)

Vlastnosti Karhunen-Loevova rozvoje:

1. Při daném počtu členů rozvoje poskytuje ze všech rozvojů **nejmenší střední kvadratickou odchylku** od původních vzorů
2. Vzory jsou po použití disperzní matice po aproximaci nekorelované

   → **dekorelace příznaků**

Karhunen-Loevovův rozvoj (2)

3. Členy rozvoje **nepřispívají rovnoměrně k aproximaci**
   - Vliv každého z členů na přesnost aproximace se zmenšuje s jeho pořadovým číslem
     → Vliv členů s vysokými indexy bude malý a můžeme je zanedbat (~ vynechat)
4. **Velikost chyby aproximace neovlivňuje strukturu rozvoje**
   - Změna požadavků na chybu aproximace nevyžaduje přepočítávat celý rozvoj
     → Stačí jen přidat či odstranit několik posledních členů

Výhodné zejména u **sekvenčních metod klasifikace**

Karhunen-Loevovův rozvoj (3)

- Volba vhodného zobrazení $V: \boldsymbol{X}^m \rightarrow \boldsymbol{X}^p$ tak, aby vzory z $\boldsymbol{X}^p$ byly nejlepší aproximací původních vzorů z $\boldsymbol{X}^m$ ve smyslu střední kvadratické odchylky

  K vzorů z jedné třídy

  m příznaků

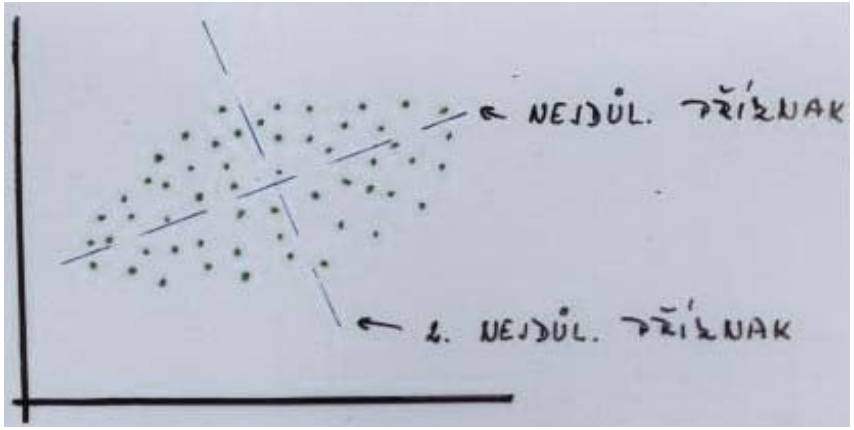
  p ortonormálních vektorů $\mathbf{e}_i$ $(1 \le i \le p)$ v $\boldsymbol{X}^m$ $(p \le m)$

  → Aproximace vektorů $\mathbf{x}_k$ z $\boldsymbol{X}^m$ $(1 \le k \le K)$ lineární kombinací vektorů $\mathbf{e}_i$:

  $$\mathbf{y}_k = \sum_{i=1}^{p} c_{ki}\, \mathbf{e}_i$$

  tak, aby kvadrát odchylky $\mathbf{x}_k$ od $\mathbf{y}_k$: $\varepsilon_k^2 = \|\mathbf{x}_k - \mathbf{y}_k\|^2$ byl minimální

Karhunen-Loevovův rozvoj (4)



$\mathbf{v} = (v_1, v_2, \ldots)^T$,

$\mathbf{x} = (x_1, x_2, \ldots)^T$

$\mathbf{y} = \mathbf{v}^T \mathbf{x} = v_1 x_1 + v_2 x_2 + \ldots$

Měřeno m příznaků, z nichž chceme získat p nejdůležitějších příznaků $(1 \leq p << m)$

Matice $V : p \times m$

$$\mathbf{V} = \begin{pmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mp} \end{pmatrix}$$

Výpočet vektoru p nejdůležitějších příznaků:

$$\mathbf{y} = \mathbf{V}^T \mathbf{x}$$

Karhunen-Loevovův rozvoj (5)

Výpočet matice V:

- vycentrovat data:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

- disperzní matice pro trénovací množinu:

$$w_{ij} = w_{ji} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \mu_i\right)\left(x_{kj} - \mu_j\right)$$

- vektory definující nejdůležitější příznaky jsou charakteristickými vektory disperzní matice

Karhunen-Loevovův rozvoj (6)

- Charakteristická čísla odpovídají rozptylu nejdůležitějších příznaků
  - prvním sloupcem matice V bude charakteristický vektor odpovídající největšímu charakteristickému číslu, …
  - další sloupce V se přestanou přidávat poté, co lze další charakteristická čísla vzhledem k jejich velikosti zanedbat

Problém:

- volba odpovídajícího počtu charakteristických čísel (p)
- nelze zaručit optimální volbu p vzhledem ke skutečnému významu jednotlivých příznaků

Karhunen-Loevovův rozvoj (7)

Modifikace:

1. **Centrované nejdůležitější příznaky**

   $\mathbf{y} = \mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu})$, kde $\boldsymbol{\mu} = (\mu_1, \ldots)$ je vektor středních hodnot

2. **Normalizované nejdůležitější příznaky**

   $\mathbf{y} = \mathbf{L}^{-1/2}\,\mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu})$, kde $\mathbf{L}$ je matice $p \times p$, prvky diagonály jsou charakteristická čísla odpovídající sloupcům $\mathbf{V}$, ostatní prvky jsou nulové

3. **Normalizace nejdůležitějších příznaků vzhledem k rozptylům**

$$w'_{ij} = \frac{w_{ij}}{\sqrt{w_{ii}\, w_{jj}}}$$

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):**

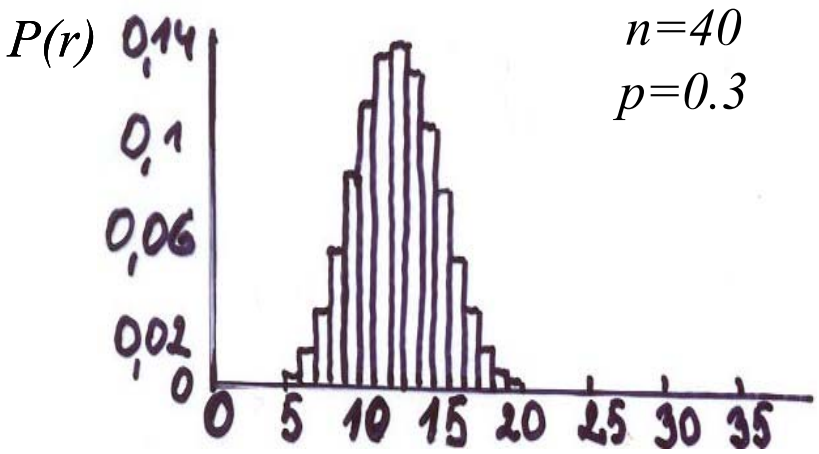
$$VAR\ (Y) \quad = \quad E\left[(Y - \mu_Y)^2 \right]$$

  - 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)

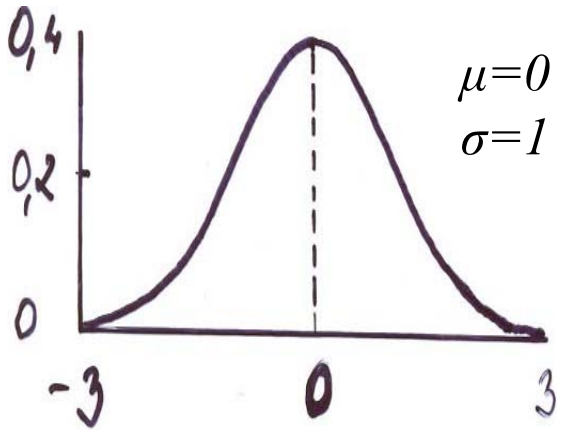
- Střední hodnota náhodné veličiny X: $E[X] = np$
- Rozptyl: $VAR(X) = np(1-p)$
- Směrodatná odchylka: $\sigma_X = \sqrt{n\ p\ (1-\ p\ )}$
- Pro dostatečně velké hodnoty n lze binomické rozložení aproximovat normálním rozložením se stejnou střední hodnotou a rozptylem
- **Doporučení:** aproximaci normálním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

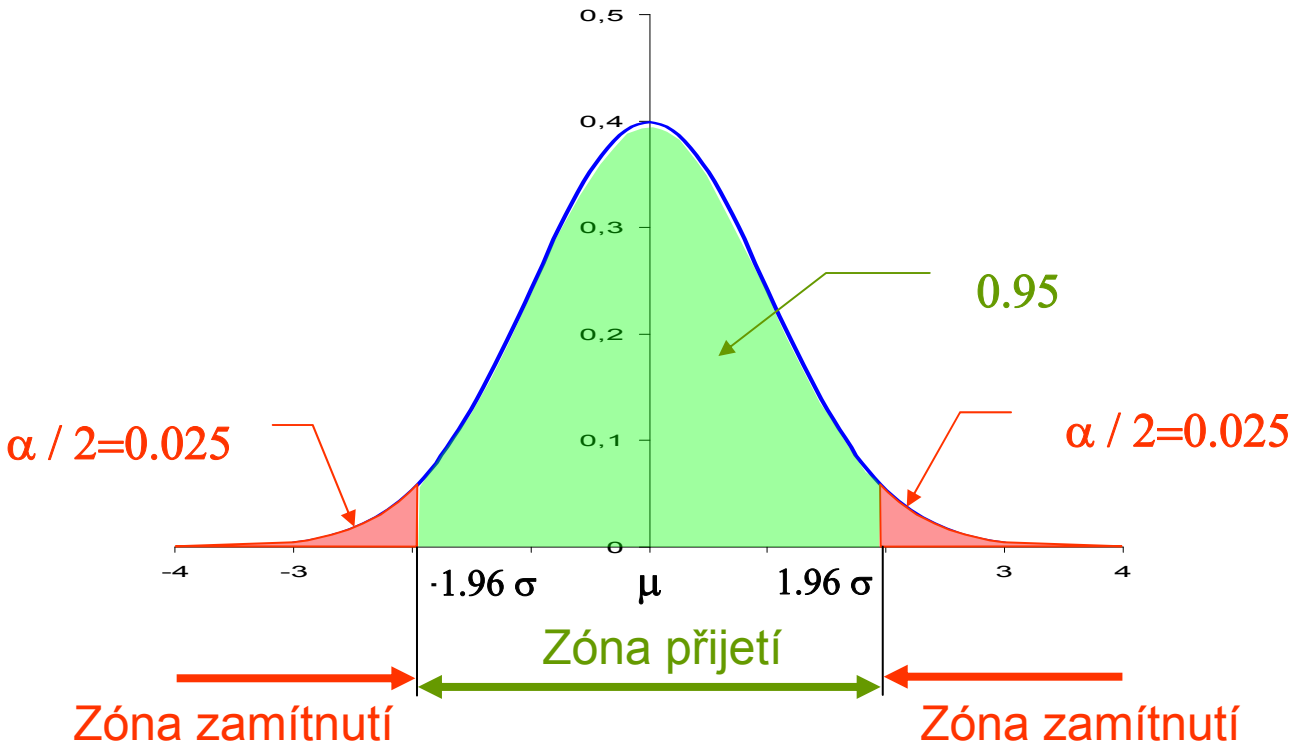
k-násobná křížová validace (4)

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

→ **Párové testy**

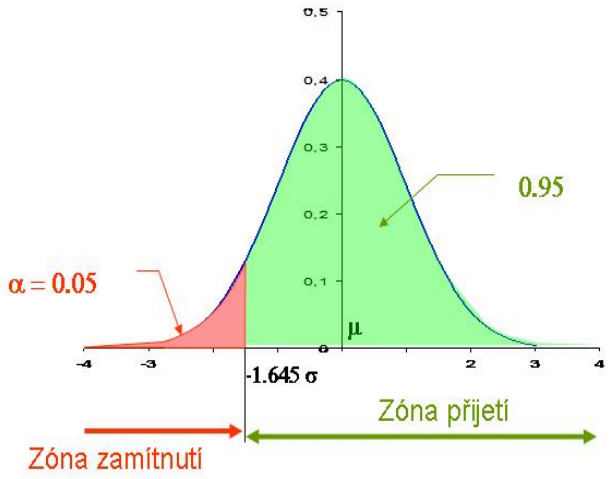
- dávají užší intervaly spolehlivosti, protože rozdíly v pozorovaných chybách vznikají kvůli rozdílům v hypotézách, ne kvůli rozdílům v datech

Oboustranný test



Jednostranný × oboustranný test

Jednostranný test



Oboustranný test