# Pravidlo *k*-nejbližších sousedů

Jana Štanclová

jana.stanclova@ruk.cuni.cz

LS 2/0 Zk

www.cuni.cz/~stancloj/

# Pravidlo *k*-nejbližších sousedů

- obsah
  - pravidlo nejbližšího souseda
    - vlastnosti
  - pravidlo k-nejbližších sousedů
    - zlepšení výpočetní složitosti
    - metriky

# Pravidlo nejbližšího souseda

# Pravidlo nejbližšího souseda

- pravidlo nejbližšího souseda
  - vstup
    - $D^n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ ... „označkované prototypy“
    - $\boldsymbol{x}$ ... neznámý vzor → klasifikován
  - výstup
    - $\boldsymbol{x}'$ ... prototyp z $D^n$ nejblížší k $\boldsymbol{x}$ → klasifikace $\boldsymbol{x}$ do třídy, kam patří $\boldsymbol{x}'$

- vlastnosti
  - suboptimální
  - typicky
    - větší chyba pravděpodobnosti než minimální možná chyba
  - lze ukázat
    - pro $n \to \infty$

      pravděpodobnost_chyby ≤ 2 * Bayesovská_pravděpodobnost_chyby

# Pravidlo nejbližšího souseda – proč funguje?

- pravidlo nejbližšího souseda
  - $\theta'$... značka třídy, kam patří prototyp $\boldsymbol{x'}$
  - pravděpodobnost $\theta'=\omega_i$
    → aposteriorní pravděpodobnost $P(\omega_i / \boldsymbol{x'})$
  - velký počet vzorů
    - lze předpokládat, že $\boldsymbol{x'}$ dostatečně blízko k $\boldsymbol{x}$
      → $P(\omega_i / \boldsymbol{x'}) \approx P(\omega_i / \boldsymbol{x})$

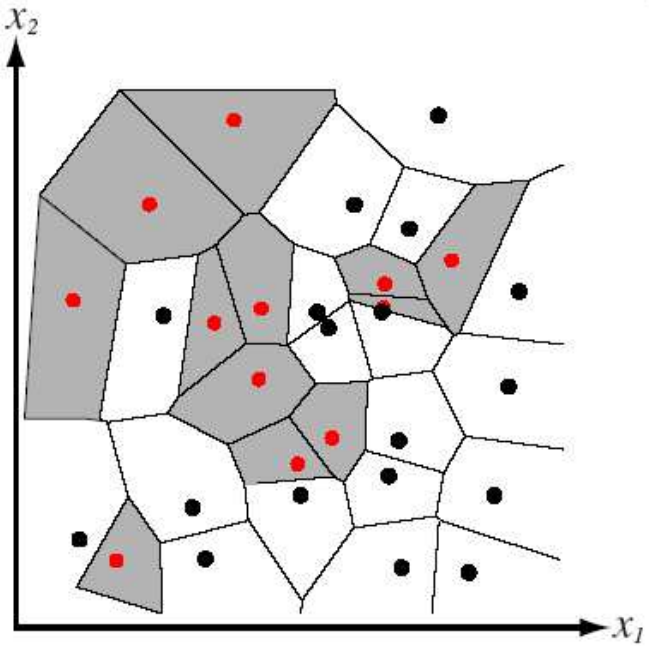
- rozdělení příznakového prostoru do oblastí
  - oblast $R_i = \{ \boldsymbol{x} ; d(\boldsymbol{x},\boldsymbol{x}_i) \leq d(\boldsymbol{x},\boldsymbol{x}_k), i \neq k \}$
  - body v oblasti označkovány třídou „jejich“ trénovacího vzoru
    → ??

# Pravidlo nejbližšího souseda – proč funguje?

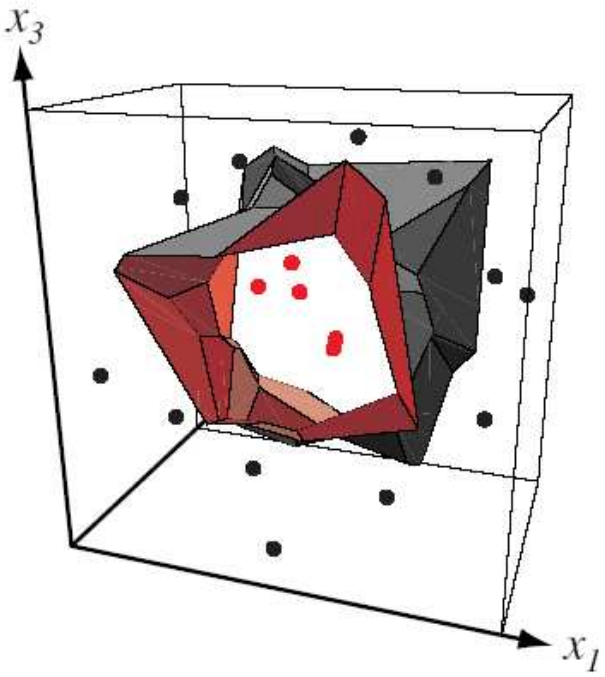
- pravidlo nejbližšího souseda
  - $\theta'$... značka třídy, kam patří prototyp $\boldsymbol{x'}$
  - pravděpodobnost $\theta'=\omega_i$
    → aposteriorní pravděpodobnost $P(\omega_i / \boldsymbol{x'})$
  - velký počet vzorů
    - lze předpokládat, že $\boldsymbol{x'}$ dostatečně blízko k $\boldsymbol{x}$
      → $P(\omega_i / \boldsymbol{x'}) \approx P(\omega_i / \boldsymbol{x})$

- rozdělení příznakového prostoru do oblastí
  - oblast $R_i = \{ \boldsymbol{x} ; d(\boldsymbol{x},\boldsymbol{x}_i) \leq d(\boldsymbol{x},\boldsymbol{x}_k), i \neq k \}$
  - body v oblasti označkovány třídou „jejich“ trénovacího vzoru
    → Voronoiova mozaika

# Pravidlo nejbližšího souseda – Voronoiova mozaika




*n*-úhelníkové oblasti




3D-krystaly

# Pravidlo nejbližšího souseda – vlastnosti 1

- vlastnosti pravidla
  - $\omega_i$ ... nejpravděpodobnější třída $\boldsymbol{x}$

$$P(\omega_i \mid \mathbf{x}) = \max_k P(\omega_k \mid \mathbf{x})$$

  - $P(\omega_i \mid \boldsymbol{x}) \sim 1$
    - pravidlo nejbližšího souseda ≈ Bayesovský výběr
    - minimum chybné pravděpodobnosti malé → pravděpodobnost chyby při metodě nejbližšího souseda také malá
  - $P(\omega_i \mid \boldsymbol{x}) \sim 1/c$
    - pravidlo nejbližšího souseda ≠ Bayesovský výběr
    - pravděpodobnosti chyby přibližně stejná u obou metod

# Pravidlo nejbližšího souseda – vlastnosti 2

- $n \to \infty$
  - značení

    $$P = \lim_{n \to \infty} P_n$$

    průměrná pravděpodobnost chyby na $n$ trénovacích vzorech

  - lze dokázat

    $$P^* \le P \le P^* \left( 2 - \frac{c}{c-1} P^* \right) \le 2P^*$$

    Bayesovská chyba

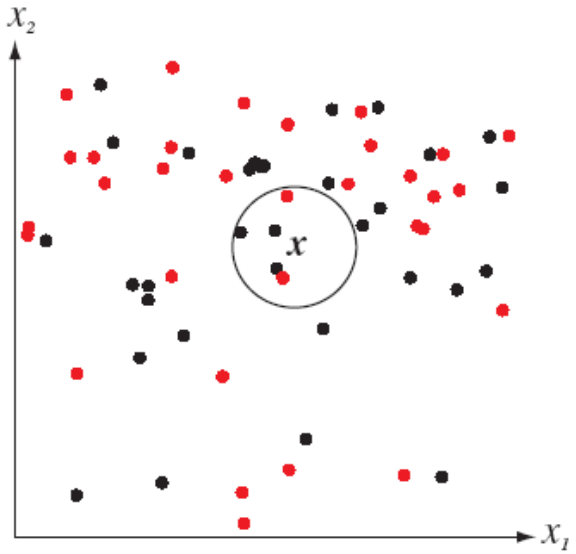
    počet tříd

- $n < \infty$
  - obecný případ → není znám žádný vztah
    - konvergence může být libovolně pomalá
    - chyba $P_n$ se nemusí monotónně snižovat
  - obecně nelze analyticky spočítat
    - nutné další předpoklady o pravděpodobnostní struktuře problému

# Pravidlo *k*-nejbližších sousedů

# Pravidlo *k*-nejbližších sousedů

- zobecnění metody
  - vzor ***x*** klasifikován do třídy nejvíce zastoupené mezi *k* nejbližšími trénovacími vzory
    - nalezení *k* nejbližších sousedů (trénovacích vzorů)
    - zjištění tříd nejbližších sousedů
    - vítězství třídy s „největším počtem hlasů“



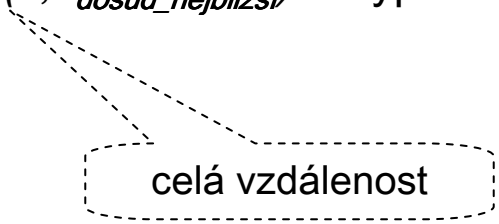

*k=5*

# Zlepšení výpočetní složitosti nejbližšího souseda

- literatura
  - mnoho analýz ohledně výpočetní složitosti (pro *dim=1* a *dim=2*)
- 2 základní techniky
  - 1. metoda částečné vzdálenosti
  - 2. eliminace zbytečných prototypů

# Metoda částečné vzdálenosti – 1

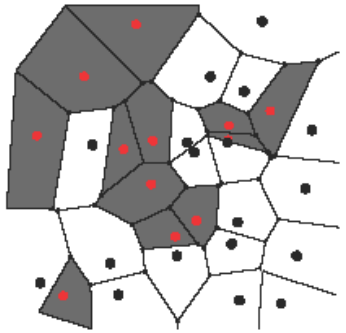
- vstup
  - *n* trénovacích vzorů (prototypů) dimenze *d*
  - konstanta *r (r<d)*
- výpočet
  - částečná vzdálenost $d_r(\mathbf{x},\mathbf{x'})$ .... prvních *r* příznaků
  - $d_r(\mathbf{x},\mathbf{x'}) > d(\mathbf{x},\mathbf{x}_{dosud\_nejbližší})$ → výpočet končí



# Eliminace zbytečných prototypů – 2

- metoda
  - eliminace prototypů, v jejichž okolí jsou jen prototypy ze stejné třídy

→ rozhodovací hranice i celková chyba se nezmění



- algoritmus

```
1. begin
2.   inicializace j ← 0, D ← trénovací_data, n ← počet_prototypů
3.   vytvoření úplného Voronoiova diagramu z D
4.   do j ← j+1
5.      nalezení Voronoiovy sousedy pro prototyp xj'
6.     if (některý soused z jiné třídy než xj') then označení xj'
7.   until j=n
8.   odstranění všech neoznačených prototypů
9.   vytvoření Voronoiova diagramu ze zbylých (označených) prototypů
10.end
```

# Eliminace zbytečných prototypů – 2

- prototyp zůstává
  - přispívá-li k rozhodovací hranici → aspoň jeden z jeho sousedů patří k jiné třídě

- vlastnosti
  - negarantuje minimální množinu prototypů (–)
  - sníží výpočetní složitost (+)
    - bez změny přesnosti výpočtu
  - nelze dodatečně přidávat prototypy do „vyčištěného“ modelu (–)
    - k vyčištení je potřeba znalost všech trénovacích dat

# Metriky

# Metriky pro metody nejbližších sousedů – 1

- metrika
  - nezápornost: $D(\mathbf{a},\mathbf{b}) \geq 0$
  - reflexivita: $D(\mathbf{a},\mathbf{b}) = 0$ jen když $\mathbf{a} = \mathbf{b}$
  - symetrie: $D(\mathbf{a},\mathbf{b}) = D(\mathbf{b},\mathbf{a})$
  - trojúhelníková nerovnost: $D(\mathbf{a},\mathbf{b}) + D(\mathbf{b},\mathbf{c}) \geq D(\mathbf{a},\mathbf{c})$

- nejběžnější metriky
  - Eukleidovská metrika

$$D(\mathbf{a},\mathbf{b}) = \left( \sum_{k=1}^{d} (a_k - b_k)^2 \right)^{1/2}$$

  - Minkowského metrika

$$L_k(\mathbf{a},\mathbf{b}) = \left( \sum_{i=1}^{d} |a_i - b_i|^k \right)^{1/k}$$

# Metriky pro metody nejbližších sousedů – 2

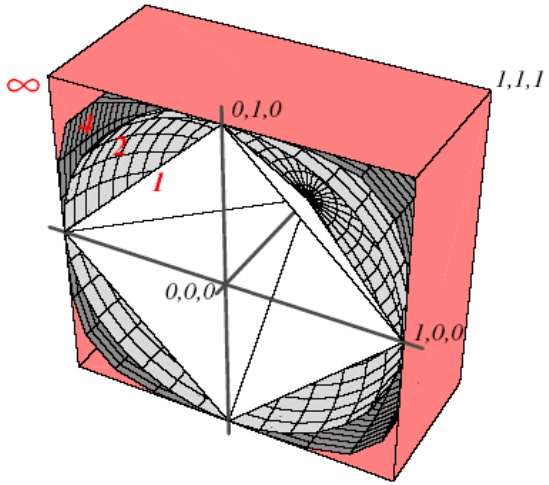
- Minkowského metrika $L_k(\mathbf{a},\mathbf{b}) = \left( \sum_{i=1}^{d} |a_i - b_i|^k \right)^{1/k}$
  - *k=1* ... Manhatannovská vzdálenost
  - *k=2* ... Eukleidovská vzdálenost
  - *k=∞* ... maximum z projekcí bodů na jednotlivé souřadnicové osy

  - jak vypadá množina bodů ve vzdálenosti 1 od počátku pro *k=1, k=2, k=∞* ?

# Metriky pro metody nejbližších sousedů – 2

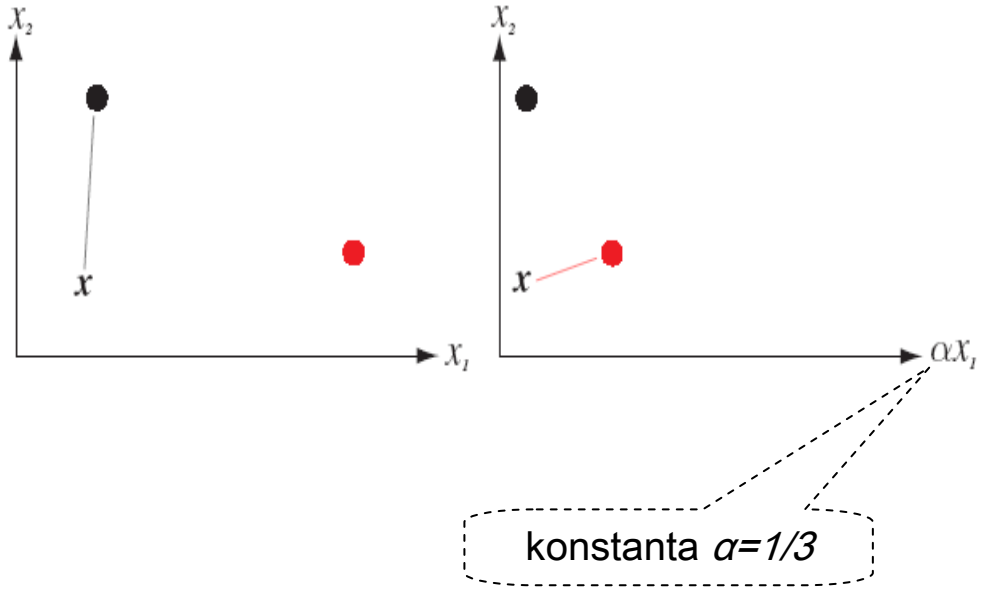
- Minkowského metrika $L_k(\mathbf{a},\mathbf{b}) = \left( \sum_{i=1}^{d} |a_i - b_i|^k \right)^{1/k}$
  - *k=1* ... Manhatannovská vzdálenost
  - *k=2* ... Eukleidovská vzdálenost
  - *k=∞* ... maximum z projekcí bodů na jednotlivé souřadnicové osy

  - jak vypadá množina bodů ve vzdálenosti 1 od počátku pro *k=1, k=2, k=∞* ?



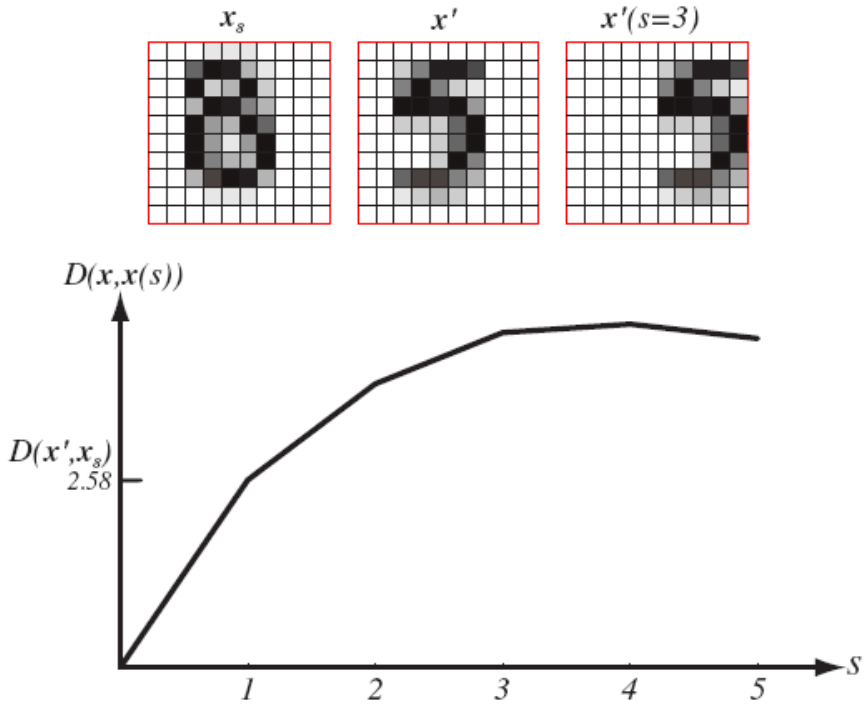
# Metriky pro metody nejbližších sousedů – 3

- ne vždy se hodí Eukleidovská metrika
  - např. vynásobení každé souřadnicové osy konstantou
    → odlišné Eukleidovské vzdálenosti v transformovaném prostoru
    → vliv na nalezení nejbližšího souseda



# Metriky pro metody nejbližších sousedů – 4

- Eukleidovská metrika – invariance vůči posunutí




$L_2(\mathbf{x}',\mathbf{x}'(s)) >> L_2(\mathbf{x}',\mathbf{x}_8)$

→ pravidlo nejbližšího souseda s Eukleid. vzdáleností vede k velkým chybám

# Tangentová metrika – konstrukce klasifikátoru

- tangentová metrika
  - výpočetně náročná (–)
  - invariantní vůči základním transformacím (+)
  - *r* transformací
    - např. horizontální a verikální posun, rotace, škálování, čárové ztenčení, ....

- konstrukce klasifikátoru
  - pro každý prototyp $\boldsymbol{x'}$
    1. aplikace jednotlivých transformací $F_i(\boldsymbol{x'}, \alpha_i)$
    2. vytvoření tangentového vektoru $TV_i$ pro každou transformaci $i$

       $$TV_i = F_i(\boldsymbol{x'}, \alpha_i) - \boldsymbol{x'}$$
    3. uspořádání tangentových vektorů $TV_i$ do matice $T(\boldsymbol{x'})$

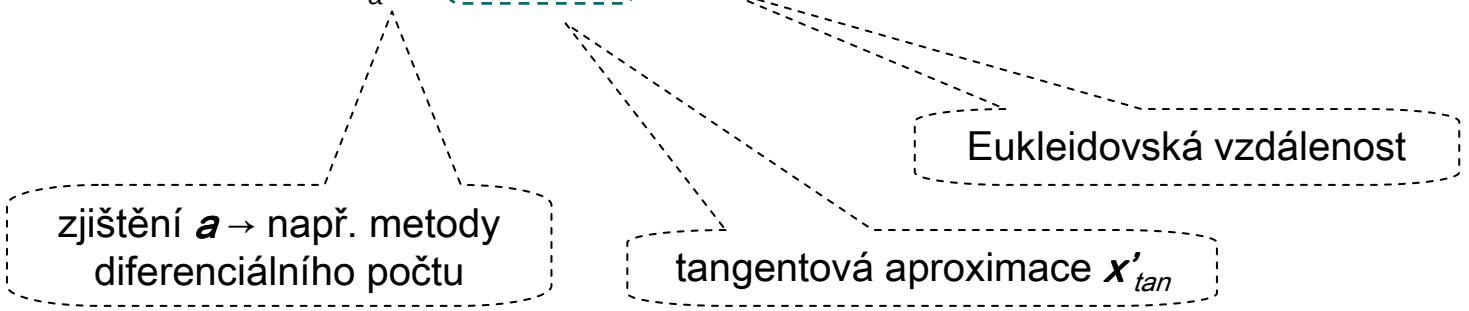
$$T(\mathbf{x}') = \begin{pmatrix} \mathbf{TV}_1 \\ \vdots \\ \mathbf{TV}_r \end{pmatrix}$$

tangentové vektory

# Tangentová metrika – klasifikace

- klasifikace neznámého vektoru ***x***
  - 1. spočtení tangentové vzdálenosti pro každý prototyp ***x'***

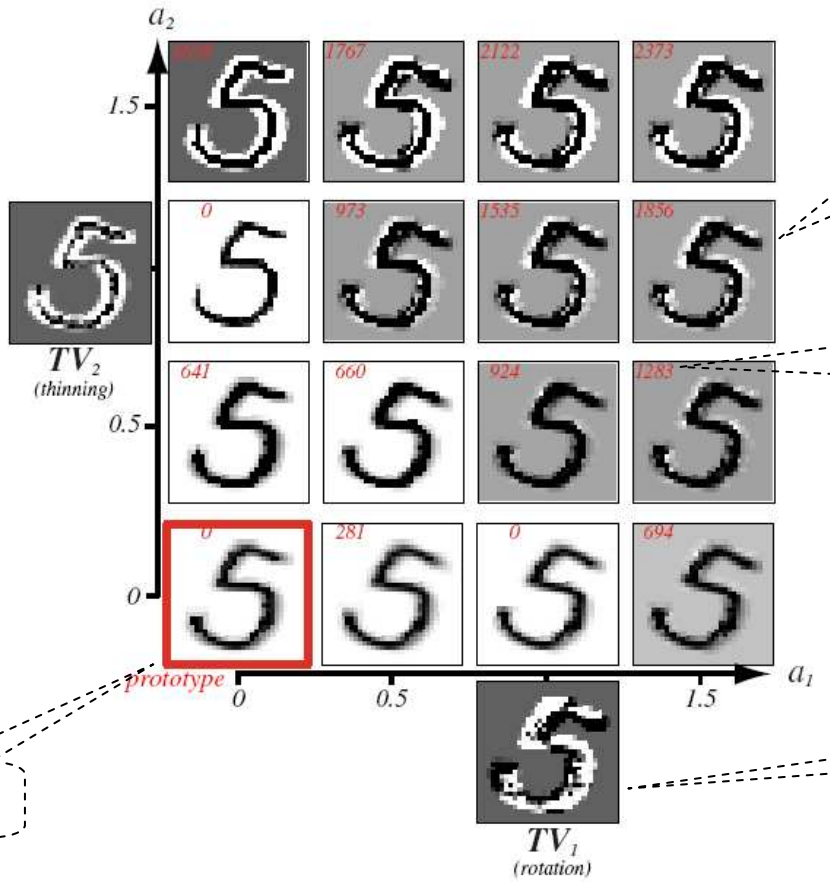
$$D_{tan}(\mathbf{x}',\mathbf{x}) = \min_{a}\left\{\left\|(\mathbf{x}'+T(\mathbf{x}')\mathbf{a})-\mathbf{x}\right\|\right\}$$




  - 2. nalezení prototypu s minimální tangentovou vzdáleností od ***x***

# Tangentová metrika – příklad

- 2 transformace
  - rotace a čárové ztenčení




lineární kombinace tangentových vektorů s koeficienty $a_1$, $a_2$

$x'_{tan} = x + a_1 TV_1 + a_2 TV_2$

Eukleidovská vzdálenost prototypu a tangentové aproximaci $x'_{tan}$

tangentový vektor $TV_2$

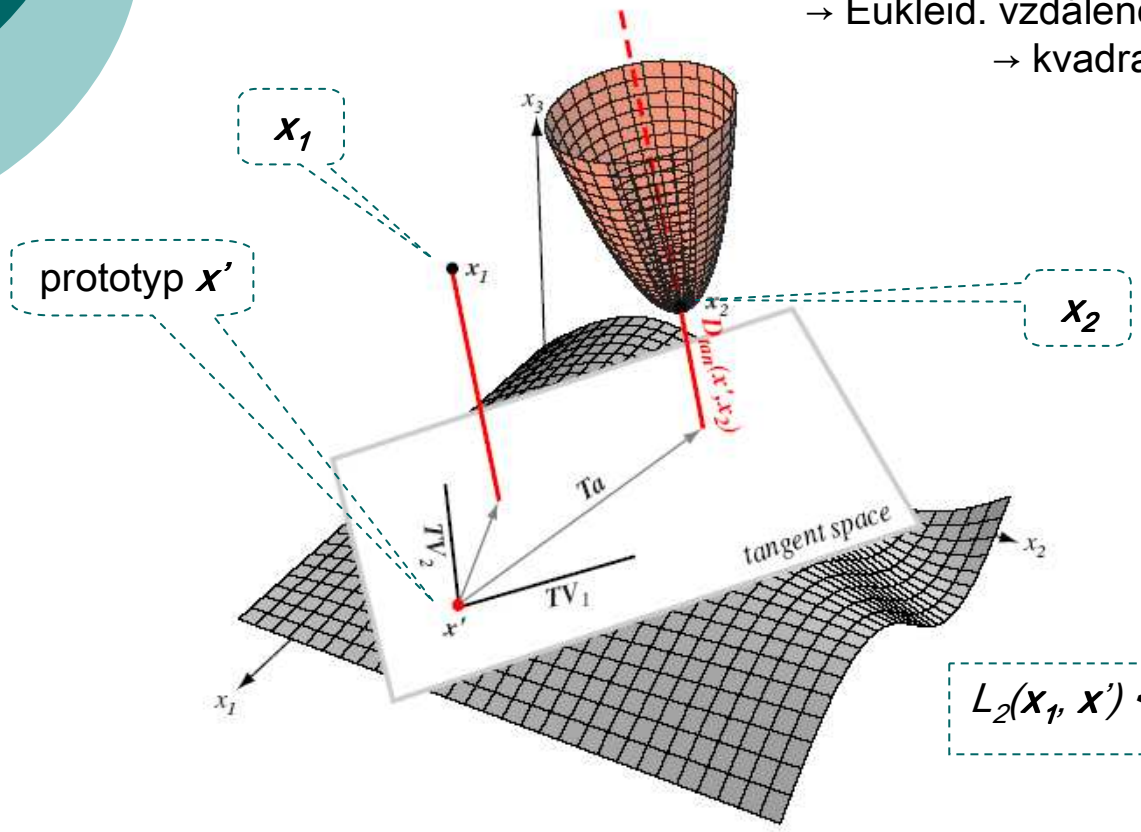
tangentový vektor $TV_1$

prototyp

# Tangentová metrika – příklad

předloženy vzory $x_1$ a $x_2$

tangentová vzdálenost $D_{tan}(x_2,x') = min_a \{|| (x'+T(x')a) - x_2 ||\}$
→ nejmenší Eukleidovská vzdálenost bodu $x_2$
k tangentovému prostoru bodu $x'$
→ Eukleid. vzdálenost $x_2$ k tangentovému prostoru $x'$
→ kvadratická funkce proměnné $a$




$L_2(x_1, x') < L_2(x_2, x')$ ale $D_{tan}(x_1, x') > D_{tan}(x_2, x')$