

# Signály a informace

12. cvičení

# Základy strojového rozpoznávání

## **Cíl úloh rozpoznávání:**

Zařadit neznámý „objekt“ reprezentovaný číselnými daty do jedné z předem daných tříd.

Příklady:

- Rozpoznávání tvarů: Je na obrázku nakreslený trojúhelník, obdélník, nebo kružnice? (3 třídy)
- Rozpoznávání obrázků: Je na obrázku pes nebo kočka? (2 třídy)
- Rozpoznávání ručně psaných číslic: Jaká číslice (0,1 ... 9) je na obrázku? (10 tříd)
- Rozpoznávání mluvených číslic: Jaká číslice (0,1 ... 9) je v nahrávce? (10 tříd)
- Rozpoznávání jazyka textu: V kterém z úředních jazyků EU je dokument? (24 tříd)
- atd.

# Základy strojového rozpoznávání

## Základní fáze vývoje rozpoznávacího (klasifikačního) systému

### 1. Sběr a příprava dat:

Shromáždění dostatečně reprezentativního vzorku dat, na nichž bude možné systém

- a) **učit a průběžně ověřovat** (trénovací sada + data vyčleněná pro průběžné testování a ladění)
- b) **testovat** (testovací sada – musí být odlišná od trénovací)

### 2. Učení (trénování) systému

Vytvoření a odladění vhodného modelu, který bude co nejlépe **reprezentovat trénovací data** a zároveň co nejlépe **klasifikovat data** vyčleněná na vývoj.

### 3. Testování systému

Ověřování systému – **rozpoznávání dat z testovací sady** a **vyhodnocování úspěšnosti**

# Rozpoznávání metodou nejmenší vzdálenosti

## Metoda nejmenší vzdálenosti

### Princip:

1. Pro objekty v dané úloze musí být vytvořena **číselná reprezentace** (čísla, vektory, matice, ...)
2. V rámci učení se **pro každou třídu** určí (vybere, vypočítá) **1 či více vzorů**.
3. Klasifikace objektu se provádí **výpočtem vzdáleností k vzorům** všech tříd a rozhodnutí se provede na základě toho, **ke kterému vzoru (a z jaké třídy) má objekt nejblíže**.

Pozn. Metodě se také říká **Metoda nejbližšího souseda** (NN – nearest neighbour)

# Rozpoznávání metodou nejmenší vzdálenosti

**Příklad: Jednoduchá třídíčka jablek a hrušek**



## Číselná reprezentace objektů:

K dispozici (např. na automatické lince) je pouze jednoduchý měřič délky a výšky objektu

## Fáze učení:

Předložíme systému  $M$  jablek a  $M$  hrušek, změříme jejich rozměry a uložíme do paměti.

Pro každou třídu máme  **$M$  vzorů** (vektorů o 2 prvcích)

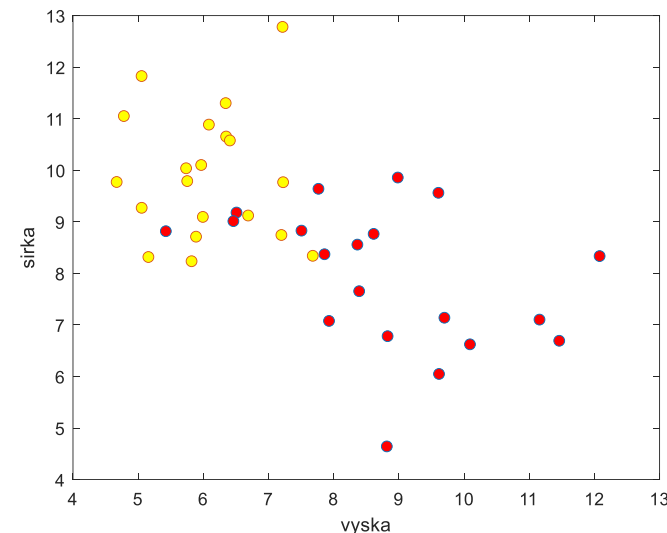
## Fáze testování:

Pro  **$N$  testovacích** objektů změříme jejich rozměry a vypočítáme vzdálenost ke všem  $2M$  vzorům.

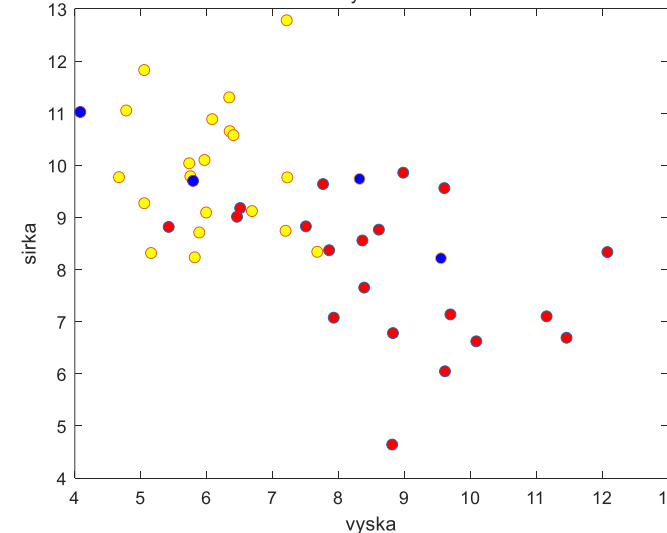
Nejbližší vzor určí, jak bude objekt klasifikován.

Výsledek klasifikace je správný/nesprávný.

**Úspěšnost** = počet\_správných\_klasifikací/ $N$



$M=20$



$N=4$   
(modré  
objekty)

# Rozpoznávání metodou nejmenší vzdálenosti

**Příklad: Jednoduchá třídička jablek a hrušek**  
- varianta s centroidy



**Číselná reprezentace objektů:**

Stejná jako předtím.

**Fáze učení:**

Předložíme systému M jablek a M hrušek, změříme jejich rozměry a pro každou třídu určíme **centroid** (vektor středních hodnot). Pro každou třídu máme **1 centroid**.

**Fáze testování:**

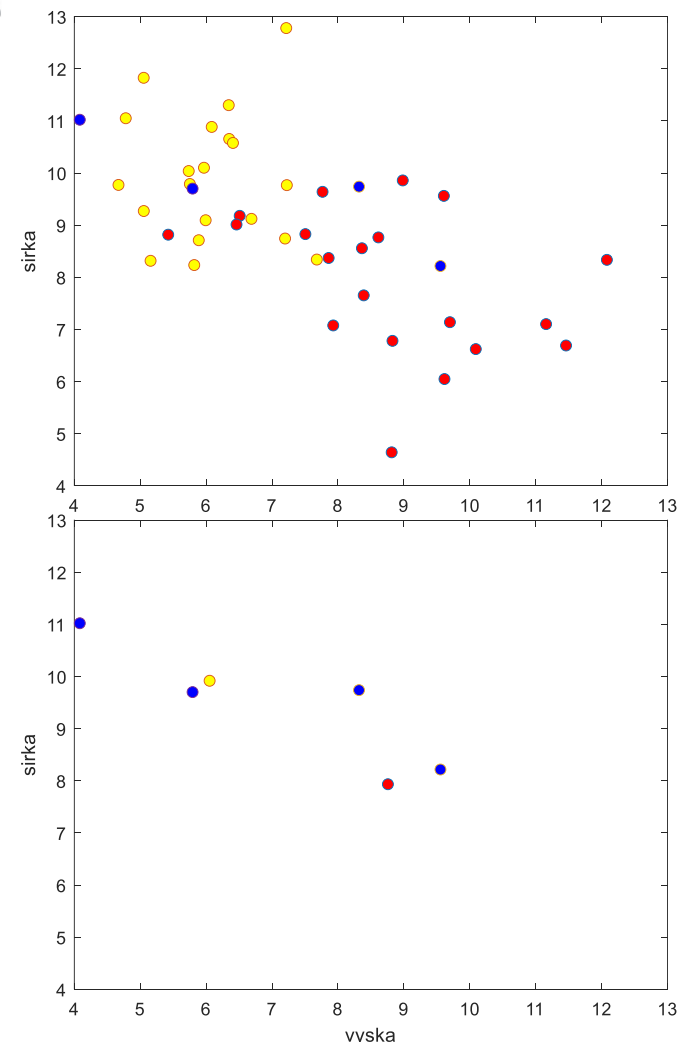
Pro **N testovacích** objektů změříme jejich rozměry a vypočítáme **vzdálenost k centroidu** každé třídy.

Nejbližší vzor určí, jak bude objekt klasifikován.

Výsledek klasifikace je správný/nesprávný.

**Úspěšnost = počet\_správných\_rozhodnutí/N**

**Pozn. Mnohem rychlejší varianta, obvykle však nižší úspěšnost.**



vzory  
M=20  
N = 4

centroidy  
M=1  
N=4

# Rozpoznávání metodou nejmenší vzdálenosti

## Jakou vzdálenost mezi vektory/maticemi používat?

( $N$  je počet prvků vektoru/matice)

### Manhatanská vzdálenost

(vzdálenost **L1**)

$$\text{dist}(\mathbf{x}, \mathbf{v}) = \sum_{n=1}^N |x[n] - v[n]|$$

### Euklidovská vzdálenost

(vzdálenost **L2**)

$$\text{dist}(\mathbf{x}, \mathbf{v}) = \sqrt{\sum_{n=1}^N (x[n] - v[n])^2}$$

### Proč **L1** a **L2** (obecně **Lp**)

Obě vycházejí ze společného vztahu:

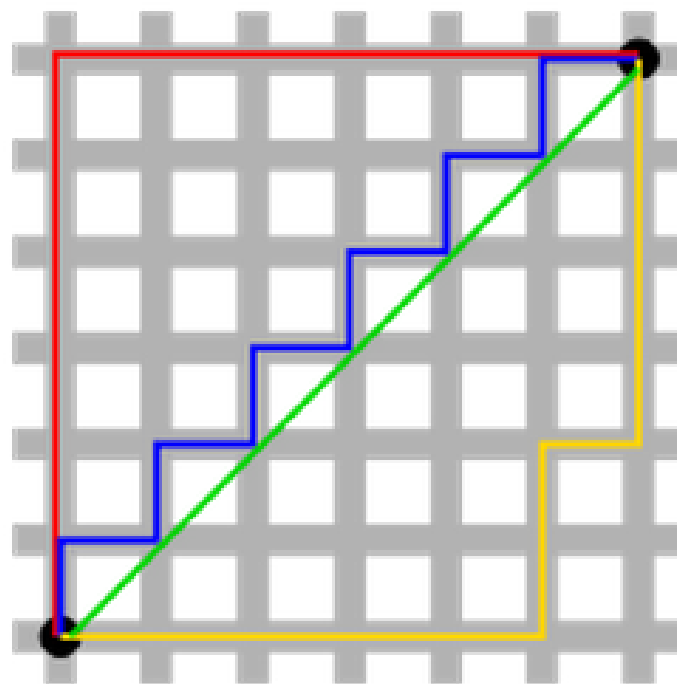
$$\text{dist}(\mathbf{x}, \mathbf{v}) = \sqrt[p]{\sum_{n=1}^N |x[n] - v[n]|^p}$$

### Manhatanská vzdálenost

(„vzdálenost naměřená podél pravoúhlých cest“)

### Euklidovská vzdálenost

(„vzdálenost naměřená vzdušnou čarou“)



# Rozpoznávání metodou nejmenší vzdálenosti (NN)

## Aplikace metody NN

Je to **nejjednodušší** metoda rozpoznávání z hlediska učení (pouze příprava vzorů či centroidů), bývá velmi pomalá (zvláště při velkém počtu vzorů pro každou třídu).

Podmínka použití: **stejné rozměry vektorů nebo matic** reprezentujících objekty

### Možné aplikace:

- Rozpoznávání stejně dlouhých sekvencí dat
- Rozpoznávání objektů popsaných N-rozměrnými příznakovými vektory (vzorky, rozměry, spektrum, ...)
- Rozpoznávání obrázků (o stejné velikosti)
- Jednoduchá diagnostika nemocí na základě měřitelných příznaků, atd.

### Výpočet vzdáleností v Matlabu:

L1:	u vektorů:	$\text{dist} = \text{sum}(\text{abs}(x-v));$	u matic:	$\text{dist} = \text{sqrt}(\text{sum}((x-v)*(x-v)));$
L2:	u vektorů:	$\text{dist} = \text{sqrt}(\text{sum}((x-v)*(x-v)))$		

Alternativní výpočet vzdálenosti nabízí funkce **norm ()**



# Rozpoznávání obrázků metodou NN - úloha

## Úloha 1:

Napište program, který bude klasifikovat ručně psané číslice z obrázků o velikosti 32x32. K dispozici máte trénovací sadu – vždy jeden vzor pro každou číslici. Testovací sada obsahuje 10 obrázků, které máte za úkol klasifikovat pomocí metody NN a vzdálenosti L1.

Stáhněte si kostru programu (soubor Uloha1.zip) a doplňte, co je třeba. (Správný výsledek je 30 %).

# Úloha k odevzdání

Na základě vyřešené Úlohy1 sestavte program Uloha2.m, který

1. Využije všechna data stažená v předchozí úloze z elearningu (celkem 100 obrázků).
  1. Názvy lze načíst jako: `file_list = dir(fullfile(datadir_XXX, '*.jpg'));`
  2. Jednotlivé znaky v názvu souboru lze vyčíst pomocí pozice: `c_part = file_name(pozice_XX);`
2. Data rozdělí do testovací sady (sady obrázků \_s01, \_s02, \_s03 a \_s04, tj. celkem 40 obrázků) a do trénovací sady (zbylé sady, tj. 60 obrázků).
3. Ve vašem programu Uloha2.m můžete využít co nejvíce kódu z Úlohy1.
4. Vzhledem k tomu, že se v Úloze2 načítá (často opakovaně) poměrně dost souborů, můžete se pokusit některé části optimalizovat.
5. V druhé části programu z trénovacích dat vytvořte pro každou třídu centroid a zopakujte rozpoznávání. Porovnejte výsledky získané pro vzory a pro centroidy.