Rozpoznávanie Obrazcov

Projekty

24.2.2014

Dôležité termíny

- Nahlásenie na cvičenia
  - **3.3.2014**
- Nahlásenie skupiny
  - **3.3.2014**
- Nahlásenie databázy
  - **24.3.2014**
- Prezentácie
  - **12/19.5.2014**

Skupiny

3-4 členné skupiny

- Dohodnite sa medzi sebou
- Ak niekto nebude zaredený do skupiny, bude mu pridelená
- Vyberte si vedúceho
  - Mail s menami členov

Projekt

- Kroky
  - Výber vhodnej databázy
  - Použitie redukčných algoritmov
  - Klasifikácia dát
  - Vyhodnotenie

Databáza

- Výber:
  - http://homepages.inf.ed.ac.uk/rbf/IAPR/researchers/PPRPAGES/pprdat.htm
    - http://lib.stat.cmu.edu/DASL/
  - http://archive.ics.uci.edu/ml/

Databáza

- **Treba zvážiť, čo chcete klasifikáciou zisťovať**
- **Vybrať si jeden z príznakov – cieľovú premennú**

- **Príklad**
  - **Databáza rýb**
    - **Príznaky:**
      - **Typ ryby, váha, dĺžka, šírka, počet plutiev, počet očí, počet šupín, výskyt**
    - **Čo by ste v tejto databáze zisťovali a prečo?**
    - **Čo by bol zlý výber a prečo?**

Databáza

- **Treba zvážiť, čo chcete klasifikáciou zisťovať**
  - **Vybrať si jeden z príznakov**

    - Príklad (nepoužívať)
      - http://archive.ics.uci.edu/ml/datasets/Forest+Fires
      - http://archive.ics.uci.edu/ml/datasets/Adult

Databáza

- **Požiadavky**
  - Min. **11** príznakov
  - Min. **500** objektov
  - **Čo chcete klasifikáciou zistiť!!!!!!!**
- **Vybranú databázu je potrebné skonzultovať**
- **Môžte si vybrať aj viacero databáz**

Výber a redukcia príznakov

- **Výber príznakov:**
  - vyberieme podmnožinu
  - Dopredný, Spätný
  - Napr. Výber N príznakov s najvyšším skóre

- **Redukcia príznakov:**
  - transformujeme pôvodnú množinu do menej-dimenzionálnej

Redukčné algoritmy

- Unsupervised (minimize the information loss)
  - Latent Semantic Indexing (LSI): truncated SVD
  - Independent Component Analysis (ICA)
  - **Principal Component Analysis (PCA)**
  - Manifold learning algorithms (a manifold is a topological space which is locally Euclidean) – Nonlinear
  -
- Supervised (maximize the class discrimination)
  - Linear Discriminant Analysis (LDA)
  - Canonical Correlation Analysis (CCA)
  - Partial Least Squares (PLS)

Klasifikácia

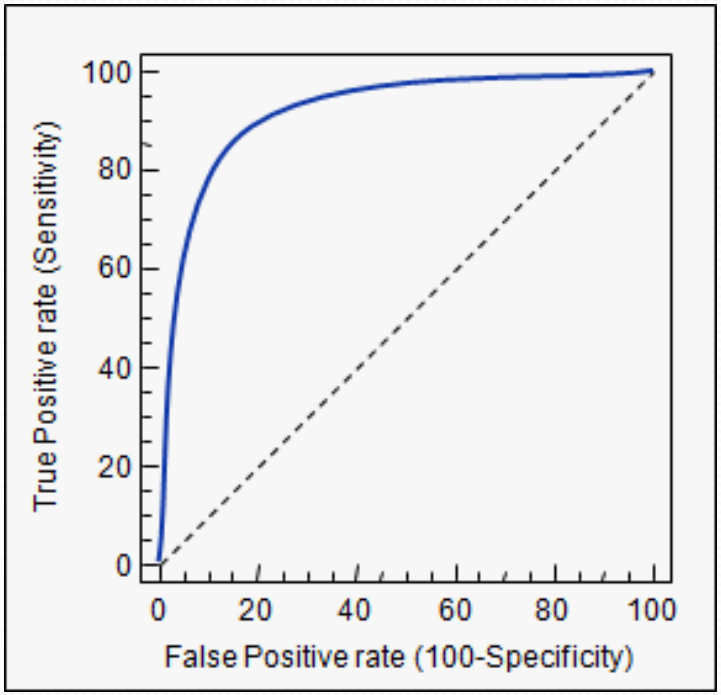
- Podstata
  - zaraďovanie objektov do kategórií na základe cieľovej premennej
  - Pre každý objekt
    - množina premenných, jedna je cieľová
- Cieľ
  - nájsť model, ktorý opisuje cieľovú premennú ako funkciu vstupných premenných (prediktorov)
- Trénovanie klasifikačného modelu vyžaduje znalosť hodnôt cieľovej premennej a prediktorov

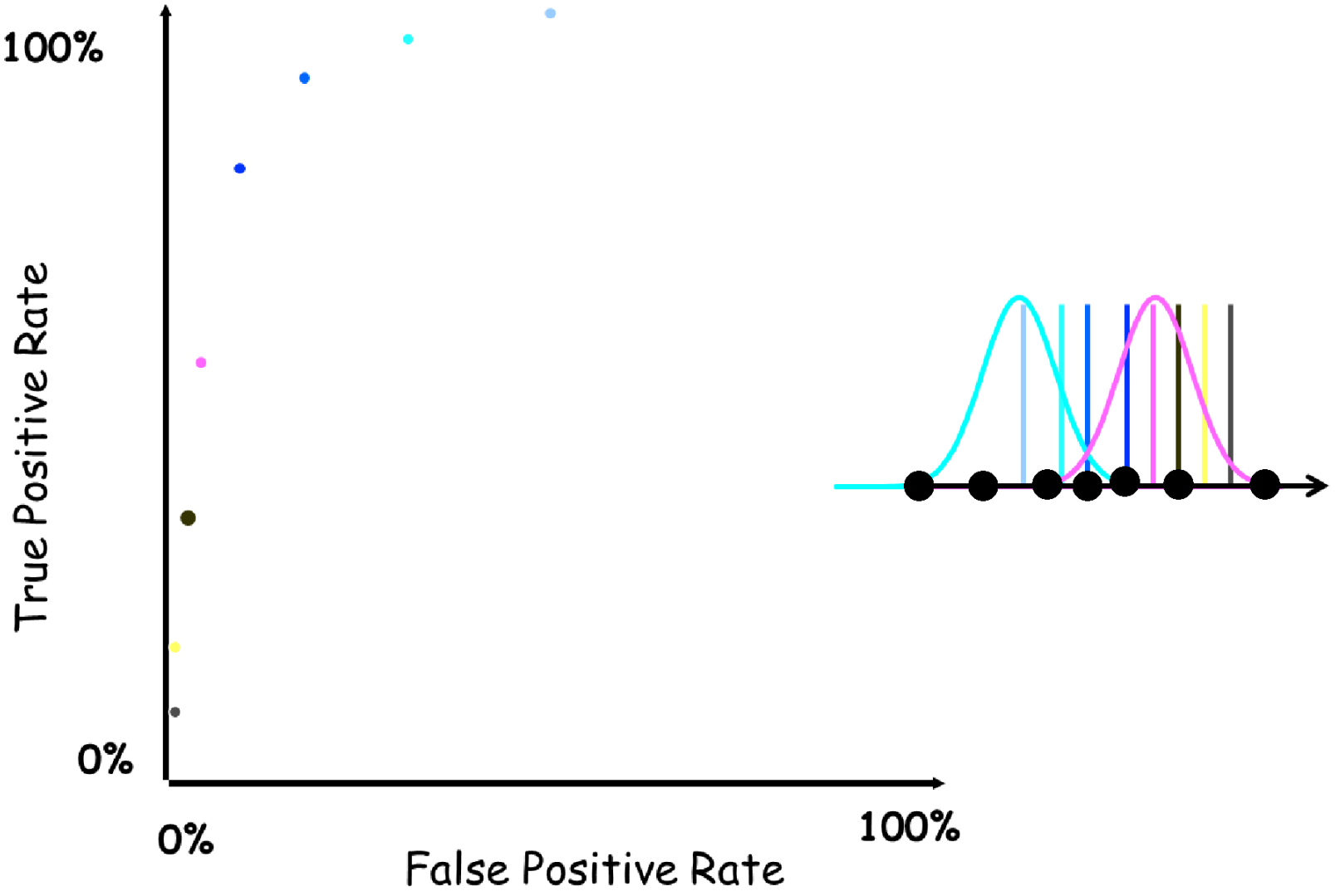
Klasifikačné algoritmy

- K-means
- K najbližších susedov
- Rozhodovacie stromy
- Bayesovský klasifikátor
- **Lineárny klasifikátor**
- SVM / Nelineárny SVM
- **ANN / SOM**
- HMM

Vyhodnotenie výsledkov

- **ROC (Receiver Operating Characteristic) krivka**
- Vzťah citlivosti a miery nesprávnej pozitivity
- vhodná pri binárnych klasifikátoroch, ktoré klasifikačné kritérium vyhodnoscujú voči nejakému prahu (obsahujú nejaký premenný parameter)
- optimálna hodnota sa dá určiť práve z analýzy ROC krivky




100%
True Positive Rate
0%
0%
100%
False Positive Rate

Vyhodnotenie výsledkov

- **Matica zámen**
  - kontignečná tabuľka, do ktorej zapisujeme počty vzoriek klasifikované do jenotlivých tried

Skutočnosť \Rozhodnutie	Negatívne	Pozitívne
Negatívne	a	b
Pozitívne	c	d

číslica	klasifikácia										
	1	2	3	4	5	6	7	8	9	0	R
1	87	0	0	0	1	0	0	0	0	0	0
2	0	88	1	0	0	0	0	0	1	1	1
3	0	0	75	1	0	0	0	10	4	0	3
4	0	0	0	79	0	0	0	0	0	0	0
5	0	0	0	0	79	6	0	0	0	4	1
6	0	0	0	0	8	80	1	0	0	2	0
7	0	1	0	0	0	0	83	0	0	0	0
8	0	0	15	0	0	1	0	65	7	0	0
9	0	0	4	0	0	0	0	10	71	0	1
0	0	0	0	1	0	1	0	0	0	90	1

Povinné časti projektu

- Redukcia príznakov
  - Povinné: **PCA**
  - 1 a viac ďalších ľubovoľných algoritmov
- Klasifikácia
  - Povinné: **Lineárny klasifikátor, Neurónové siete- ANN, SOM**
  - 1 a viac ďalších ľubovoľných algoritmov
- Vyhodnotenie
  - Matica zámen / ROC
  - Percentuálne vyhodnotenie

Záverečná správa I.

- Max 10 strán A4
- Povinné časti
  - Databáza
    - Môže byť aj viacero dostatočne rôznych databáz
    - Ukážka
    - Popis (počet prvkov, príznakov,...)
- Použité redukčné algoritmy
  - Zdôvodnenie výberu
  - Stručný popis algoritmov ich výhody/nevýhody
  - Porovnanie na databáze

Záverečná správa II.

- Povinné časti
  - Použité klasifikátory
    - Zdôvodnenie výberu
    - Stručný popis algoritmov ich výhody/nevýhody
    - Porovnanie na databáze
  - Vyhodnotenie výsledkov
    - Porovnanie výsledkov klasifikácie pri rôznych redukčných algoritmoch aj bez nich
    - Zdôvodnenie

Prezentácia

- Databáza
  - Stručné info
  - Dôvod výberu
- Redukčné algoritmy
  - Použité algoritmy
  - Dôvod výberu
  - Ukážka účinnosti na vybranej databáze
  - Klady/zápory

Prezentácia

- Klasifikátory
  - Použité algoritmy
  - Dôvod výberu
  - Ukážka účinnosti na vybranej databáze
  - Klady/zápory
- Vyhodnotenie

Hodnotenie

- 30 bodov / osoba
  - potrebných minimálne 20 spolu s dochádzkou
- hodnotí sa
  - zvládnutie danej problematiky
  - prezentácia výsledkov projektu a samotné výsledky
  - zvolené klasifikátoroy
  - konzistencia
  - redukcia prinakov
  - zvolená databáza
- hodnotíte sa aj navzájom v rámci skupiny