

Cvičení z předmětu Biometrie

cvičení 2: statistika v biometrii

Eduard Bakštein, eduard.bakstein@fel.cvut.cz

26. září 2012

Úvod

Cílem cvičení je zopakovat si v krátkém teoretickém úvodu základní pojmy z testování hypotéz a uvést je do souvislosti s biometrií.

V praktické části cvičení, které je věnován tento návod, si vyzkoušíme zhodnotit identifikaci osob na základě dynamiky úhozu kláves.

Úloha: Dynamika úhozů

Jedním ze spíše experimentálních biometrických ukazatelů je identifikace na základě dynamiky úhozů kláves, která je pro každou osobu charakteristická. Existují v zásadě dvě varianty záznamu: jednak při psaní známého textu (např. hesla - umožňuje ověřit, zda osoba, zadávající heslo, není podvodník), jednak běžná práce na pc (umožňuje zvýšit bezpečnost ověřením činnosti již autentizovaného uživatele).

V této úloze použijeme reálná data dynamiky úhozů, získaná výzkumníky z Queen Mary University of London (http://www.eecs.qmul.ac.uk/~ccloy/downloads_keystroke100.html). Databáze obsahuje data od 100 měřených osob, přičemž každá zadávala desetkrát stejné heslo "try4-mbs". Toto heslo bylo všem účastníkům experimentu předem známo a byli požádáni, aby si zadávání hesla nacvičili. V úloze budeme pracovat s latencemi, tedy dobami mezi úhozy jednotlivých kláves.

Úkoly:

1. Ze stránek cvičení si stáhněte dataset `keystrokes.mat` Data si prohlédněte (vektor "ids" obsahuje identifikaci osob k příslušným řádkům v "latencies")
2. Vytvořte jednoduchou funkci `d=compareKeystrokes(a,b)`, která vypočte vzdálenost dvou vektorů. Můžete začít s jednoduchou euklidovskou vzdáleností. Funkce bude sloužit pro jednoduché porovnání latencí mezi sebou.
3. Vytvořte program, který pomocí Vámi vytvořené funkce z předchozího bodu porovná latence v databázi proti sobě (rozuměj "každý s

každým”). Vypočtené vzdálenosti uložte do dvou vektorů podle toho, zda se porovnávaly latence od téhož, či různých lidí.

4. Zobrazte v jednom grafu normované histogramy (=sčítají do jedné) pro oba vektory. (Pro vykreslení můžete použít funkci `hist` se dvěma výstupními parametry a následně funkci `bar`). Rozmyslete, co nám zobrazené průběhy říkají o úloze a naší porovnávací metodě.
5. Pomocí funkce `plotroc` zobrazte ROC křivku pro všechna data dohromady. Co z ní lze vyčíst? Jaká bude FRR (false reject rate) při FAR=0.01? Výsledek okomentujte.

Bonusy:

1. Navrhněte vlastní metodu pro srovnání vektorů latence. Zkuste porovnat ROC křivky Vámi navržené metody a původní euklidovské vzdálenosti.
2. Vypočtěte AUC (Area Under (ROC)Curve) a obě srovnávací metody s jeho pomocí porovnejte. Která je přesnější?
3. Vyzkoušejte, jak by se AUC obou metod změnila, kdybyste místo 7-prvkových vektorů latence použili prvků jen 5 (kratší heslo).