Adatbányászat: Rendellenesség keresés

10. fejezet

Tan, Steinbach, Kumar

Bevezetés az adatbányászatba

előadás-fóliák

fordította

Ispány Márton

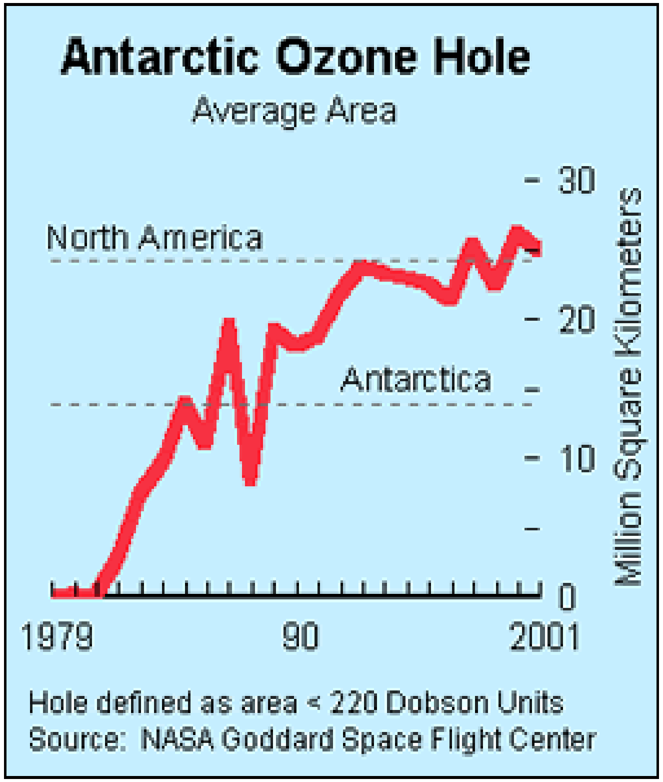
Rendellenes/kiugró adatok keresése

- Mit értünk rendellenes/kiugró adat alatt?
  - A rekordoknak egy olyan halmaza, amely számottevően eltér a többi adattól.
- Kapcsolódó feladatok:
  - Adott *D* adatbázisban találjuk meg az összes olyan $\boldsymbol{x} \in D$ rekordot, amely rendellenességi pontszáma nagyobb mint egy t küszöb.
  - Adott *D* adatbázisban találjuk meg az összes olyan $\boldsymbol{x} \in D$, rekordot, melynek $f(\boldsymbol{x})$ rendellenessége az *n* legnagyobb között van.
  - Adott *D* adatbázisban, mely jobbára normális rekordokat tartalmaz, és egy ***x*** teszt-pont esetén számoljuk ki ***x*** rendellenességi értékét *D*-re vonatkozóan.
- Alkalmazások:
  - Hitelkártya csalások, telekommunikációs csalások, hálózati betörések, csalások keresése.

Miért keressünk rendellenességeket?

Ózonréteg vékonyodása

- 1985: három kutatót (Farman, Gardinar és Shanklin) nyugtalanították a British Antarctic Survey által összegyűjtött adatok, melyek azt mutatták, hogy az Antarktiszon az ózonszint 10%-kal a normális alá csökkent.
- Miért nem jelzett a Nimbus 7 műhold, melyet ózonszint mérésre alkalmas műszerrel is felszereltek, hasonlóan alacsony koncentrációt?
- A műhold által mért ózonkoncentráció olyan alacsony volt, hogy a program kiugró adatnak kezelte és figyelmen kívül hagyta!



Forrás:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

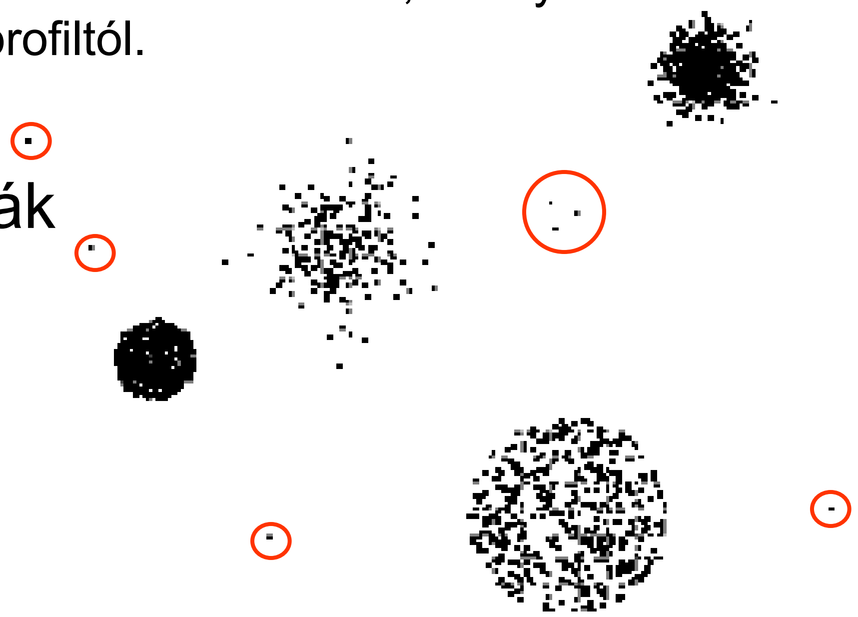
Rendellenességek keresése

- Kihívások
  - Mennyi kiugró érték van az adatok között?
  - Nemfelügyelt feladat
    - Az ellenőrzés (éppen mint a klaszterezésnél) nehéz is lehet.
  - Tű keresése a szénakazalban.

- Munka hipotézis:
  - Jóval több ,,normális” mint ,,abnormális” (kiugró/ rendellenes) megfigyelés van az adatállományban.

Rendellenesség keresési sémák

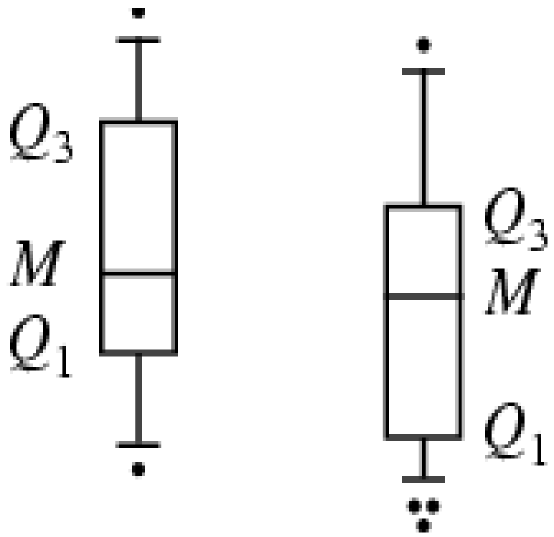
- Általános lépések
  - Alkossunk profilt a ,,normális” viselkedésről.
    - Ez lehet mintázat vagy összegző statisztika a teljes populációra.
  - Alkalmazzuk ezt a ,,normális” profilt rendellenesség keresésre.
    - Azokat a megfigyeléseket nevezzük rendellenesnek, amelyek lényegesen eltérnek a normális profiltól.

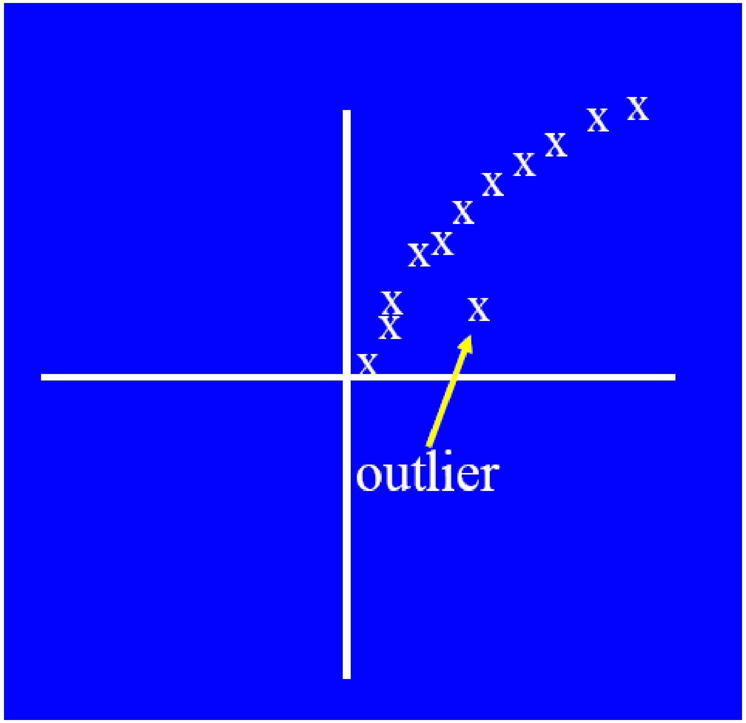


- Rendellenesség keresési sémák osztályozása
  - Grafikus és statisztikus alapú
  - Távolság alapú
  - Modell alapú

Grafikus megközelítések

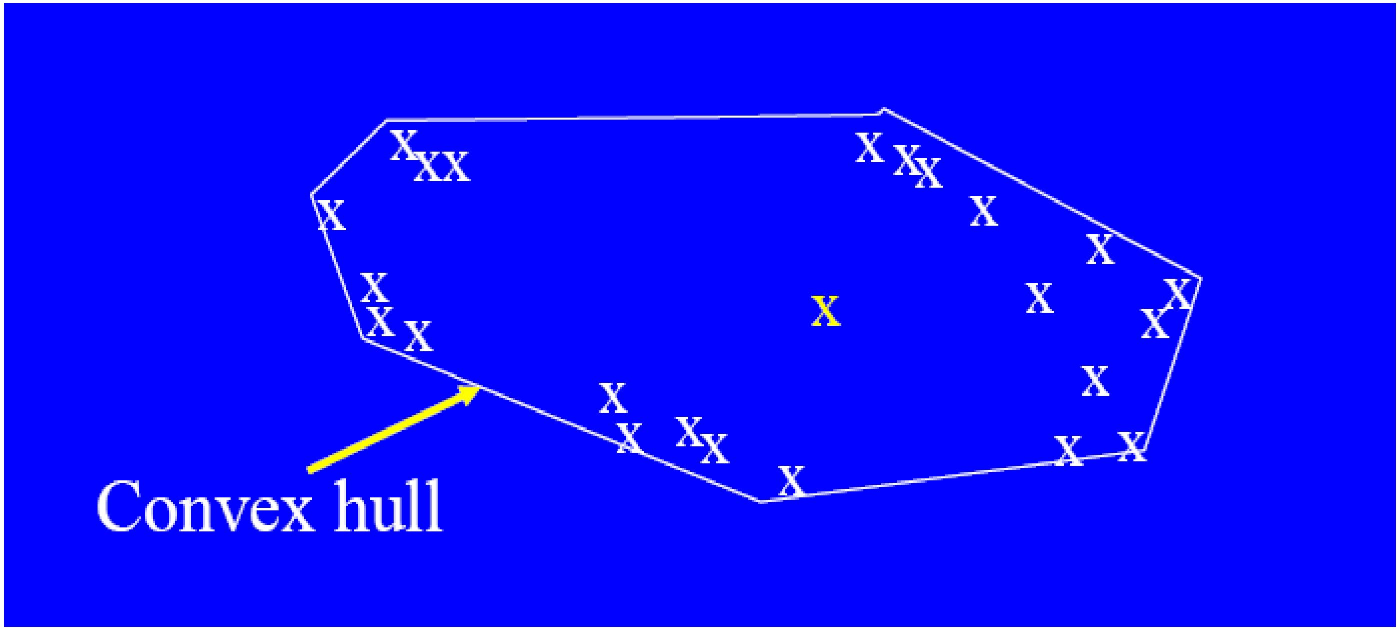
- Doboz ábra (1-D), pont diagram (2-D), térbeli diagram (3-D)
- Korlátok
  - Idő igény
  - Szubjektív





Konvex burok módszer

- Az extrém helyű pontokat kiugróaknak tekintjük.
- Használjuk a konvex burkot ezen pontok meghatározására.

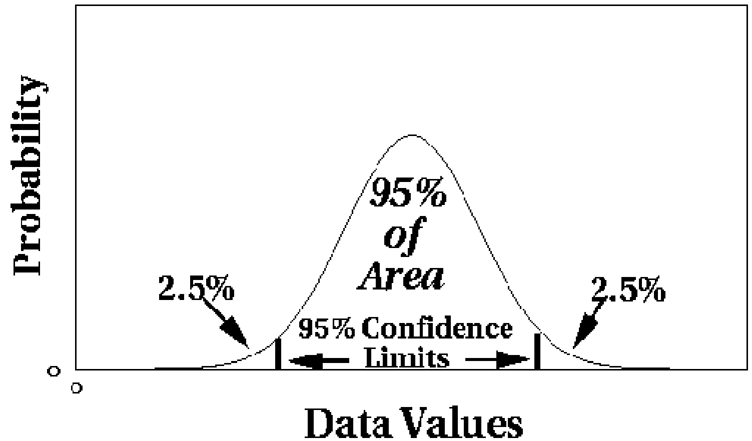


- Mi történik ha a kiugró adat középen van?

Statisztikus megközelítések

- Tegyük fel, hogy egy paraméteres modell írja le az adatok eloszlását (pl. normális eloszlás).

- Alkalmazzunk statisztikai próbákat, melyek függnek
  – az adatok eloszlásától,
  – az eloszlás paramétereitől (pl. várható érték, variancia),
  – a kiugró értékek várható számától (konfidencia határ).



Grubbs próba

- Kiugró értékeket keres egydimenziós adatokban.
- Felteszi az adatok normális eloszlását.
- Egyszerre egy kiugró értéket keres, azt eltávolít-ja, majd megvizsgálja az alábbi hipotéziseket
  - H_0: Nincs kiugró érték az adatokban
  - H_A: Van legalább egy kiugró érték
- Grubbs próba statisztika :

$$G = \frac{\max \left| X - \overline{X} \right|}{s}$$

- H_0-t elvetjük ha:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,\, N-2)}}{N-2+t^2_{(\alpha/N,\, N-2)}}}$$

Statisztikai alapú: Likelihood módszer

- Tegyük fel, hogy a *D* adatállomány két valségi eloszlás keverékéből származó mintát tartalmaz:
  - M (többségi eloszlás),
  - A (rendellenes eloszlás).
- Általános megközelítés:
  - Tegyük fel kezdetben, hogy az összes rekord *M*-beli.
  - Legyen $L_t(D)$ a *D* loglikelihoodja a *t* időpillanatban.
  - Minden *M*-hez tartozó x_t rekordot mozgassunk át *A*-ba.
    - Legyen $L_{t+1}(D)$ az új loglikelihood.
    - Számoljuk ki a differenciát $\Delta = L_t(D) – L_{t+1}(D)$.
    - Ha $\Delta > c$ (küszöbérték), akkor x_t-t rendellenesnek minősítjük és véglegesen átmozgatjuk *M*-ből *A*-ba.

Statisztikai alapú: Likelihood módszer

- Az adatok eloszlása: $D = (1 - \lambda)\, M + \lambda\, A$
- M egy az adatokból becsülhető valségi eloszlás.
  - A becslés alapulhat bármilyen modellen (naív Bayes, maximális entrópia stb).
- A-t kezdetben egyenletes eloszlásúnak feltételezzük.
- Likelihood a t időpontban:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

A statisztikus megközelítés korlátai

- A legtöbb próba csak egy attributumra működik.
- Legtöbbször nem ismert az adatok eloszlása.
- Magas dimenzióban nehéz becsülni az igazi eloszlást.

Távolság alapú módszerek

- Az adatokat jellemzők egy vektorával reprezen-táljuk.

- Három fő megközelítés
  - Legközelebbi társ módszer
  - Sűrűség alapú
  - Klaszter alapú

Legközelebbi társ alapú megközelítés

- Módszer:
  - Számoljuk ki az összes pontpár közötti távolságot.
  - A kiugró értékek definiálásának többféle módja van:
    - Azok a pontok, amelyeknek egy adott *d* sugarú környezetében kevesebb mint *p* számú pont van.
    - Az az *n* pont, amelynek a *k*-adik legközelebbi szomszédjától vett távolsága a legnagyobb.
    - Az az *n* pont, amelynek az átlagos távolsága a *k* darab legközelebbi szomszédjától a legnagyobb.

Kiugró értékek vetületekben

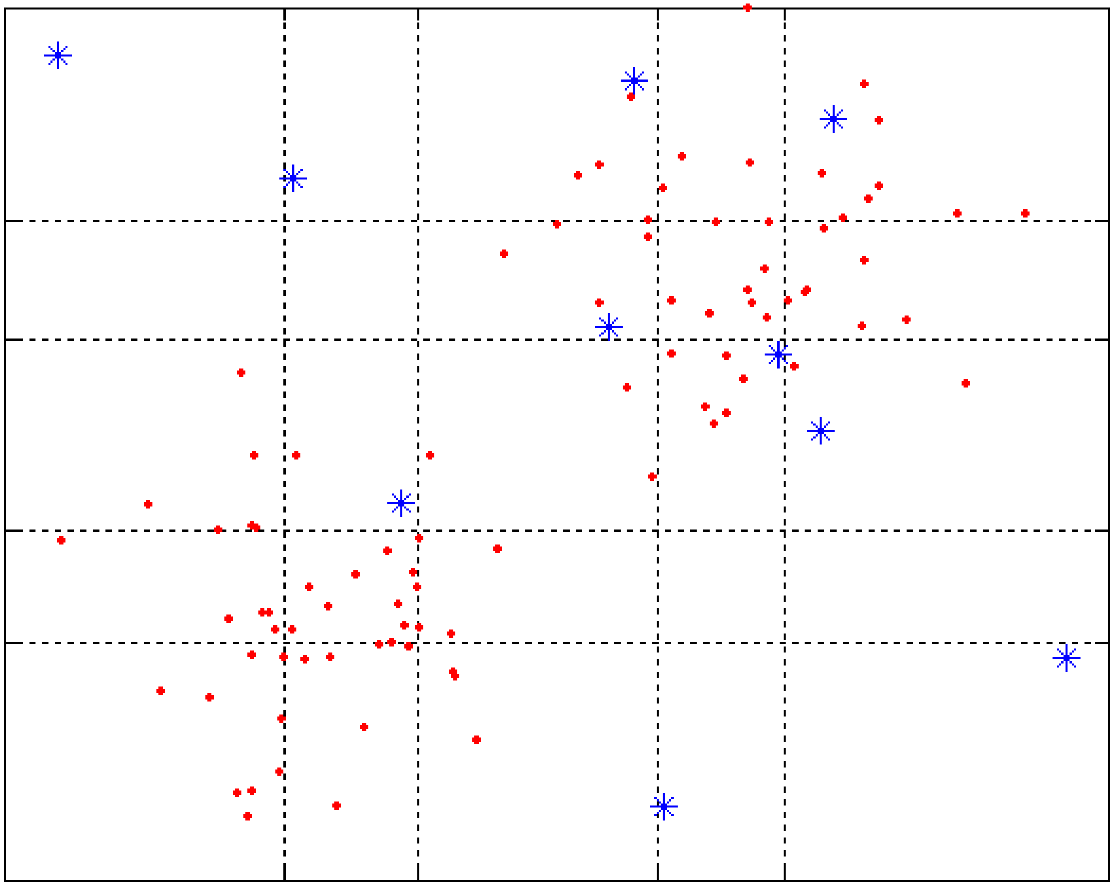
- Osszunk fel minden attributumot ϕ egyenlő mélységű intervallumra.
  - Minden intervallum $f = 1/\phi$ részt tartalmaz a rekordokból.
- Tekintsünk egy *k* dimenziós kockát, melyet *k* különböző dimenzió menti részintervallum kijelölése ad.
  - Ha az attributumok függetlenek akkor várhatóan f^k részét tartalmazza a rekordoknak.
  - *N* pont esetén a *D* kocka ritkaságát mérhetjük a következő mutatóval:

$$S(\mathcal{D}) = \frac{n(\mathcal{D}) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

  - Negatív ritkaság a vártnál kisebb számú pontot jelez a kockában.

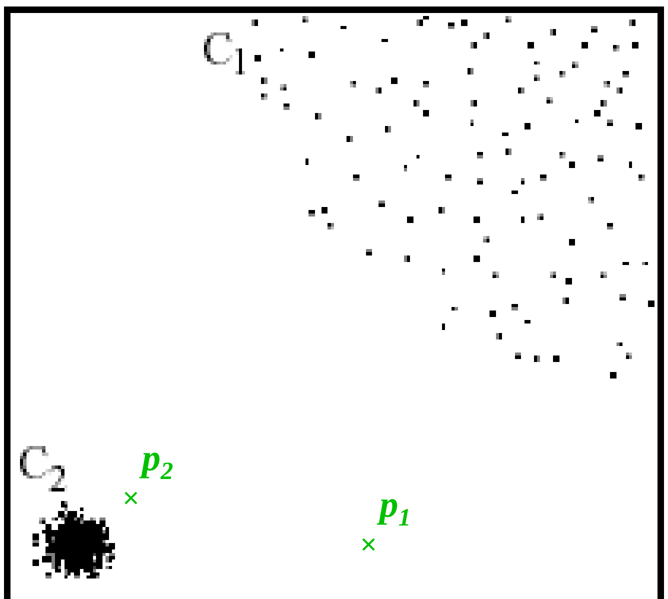
Példa

- N=100, ϕ = 5, f = 1/5 = 0.2, N $\times$ f^2 = 4



Sűrűség alapú megközelítés: LOF

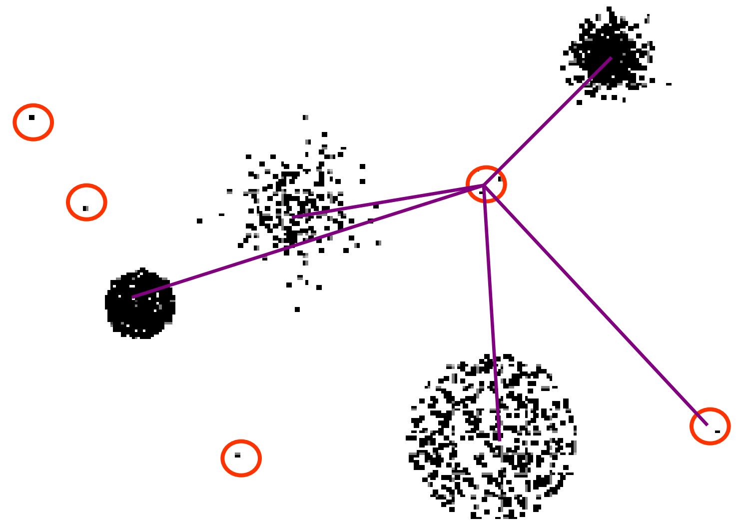
- Számoljuk ki az összes pont lokális környezetének sűrűségét.
- Számoljuk ki egy *p* minta lokális kiugró faktorát (LOF) úgy, mint a minta és az ő legközelebbi szomszédjai sűrűségének az átlagát.
- Kiugróak a legnagyobb LOF értékkel rendelkező pontok.



A legközelebbi társ módszernél p_2 nem lesz kiugró, ezzel szemben a LOF módszer p_1-t és p_2-t egyaránt kiugrónak találja

Klaszter alapú megközelítés

- Alapötlet:
  - Klaszterosítsuk az adatokat különböző sűrűségű csoportokra.
  - Válasszuk kiugró jelölteknek a kis klaszterek pontjait.
  - Számoljuk ki a kijelölt pontok és a nem kijelölt klaszterek közötti távolságot.
    - Ha a kijelölt pontok messze vannak a nem kijelölt pontoktól, akkor ők kiugróak.



Téves következtetési arány

- Bayes tétel:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

- Általánosabban:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}$$

Téves következtetési arány

- Legjobban egy példán keresztül lehet megérteni.
- Tegyük fel, hogy egy orvos által végzett teszt pontossága 99%, azaz egy beteg populáción 99%-osan jelez betegséget, és egy egészségesen 99%-ban ad negatívat.
- Vizit után az orvosnak van egy jó és egy rossz híre.
- Rossz hír: a teszt pozitív lett.
- Jó hír: a (betegség) előfordulása a teljes populációban 1/10000.
- Mennyi a valószínűsége, hogy valóban betegek vagyunk.

Téves következtetési arány

$$P(S|P) = \frac{P(S) \cdot P(P|S)}{P(S) \cdot P(P|S) + P(\neg S) \cdot P(P|\neg S)}$$

$$P(S|P) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = $$
$$= 0.00980 \ldots \approx 1\%$$

- *S* – betegség, *P* – pozitív teszt
- Bár a teszt 99%-osan pontos, annak esélye, hogy mégis betegek vagyunk 1%, mivel a populáció-ban az egészséges emberek jóval többen vannak mint a betegek.

Téves következtetés behatolás észlelésnél

- I: betolakodó viselkedés,
  ¬I: nem-betolakodó viselkedés,
  A: riasztás,
  ¬A: nincs riasztás

- Észlelési arány (igaz pozitív arány): P(A|I)
- Hamis risztás arány: P(A|¬I)

- A cél az, hogy egyaránt maximalizáljuk
  - a Bayes-i észlelési arányt, P(I|A),
  - P(¬I|¬A)

Észlelési illetve hamis riasztási arány

$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)}$$

- Tegyük fel: $P(I) = 1 \Big/ \frac{1 \cdot 10^6}{2 \cdot 10} = 2 \cdot 10^{-5};$

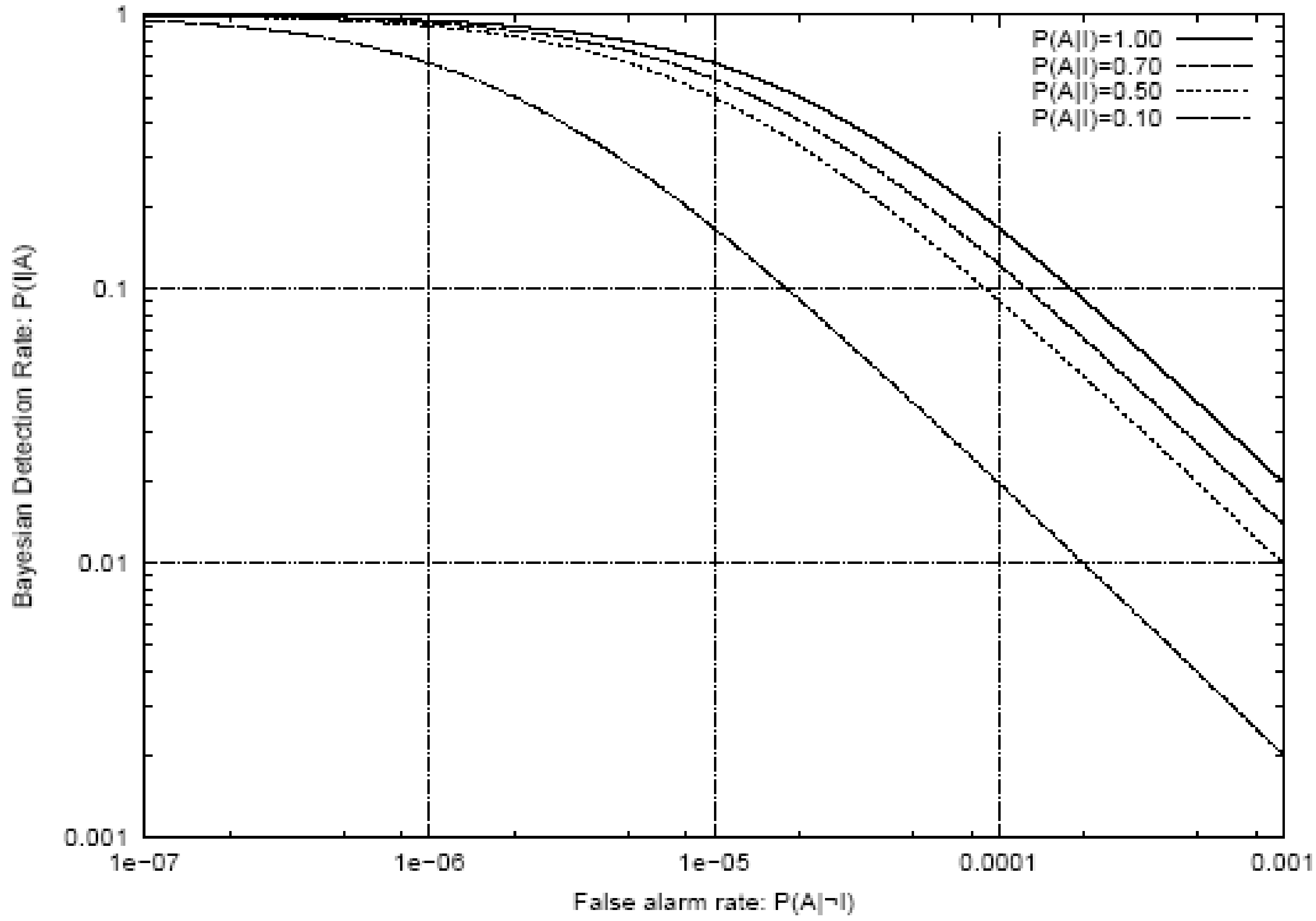
  $P(\neg I) = 1 - P(I) = 0.99998$

- Ekkor:

$$P(I|A) = \frac{2 \cdot 10^{-5} \cdot P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.99998 \cdot P(A|\neg I)}$$

- A hamis riasztási arány fog dominálni amennyiben P(I) nagyon kicsi.

Észlelési illetve hamis riasztási arány



- Axelsson: Nagyon kis hamis riasztási arány kell ahhoz, hogy ésszerű Bayes-i észlelési arányt érjünk el.