



# KOOPERÁCIÓ ÉS GÉPI TANULÁS LABORATÓRIUM

## Nemellenőrzött és féligellenőrzött tanulás
## Felkészülési segédlet

Készítette:

*Dr. Pataki Béla*
(pataki@mit.bme.hu)

Méréstechnika és Információs Rendszerek Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem

**2009, november.**

**A gyakorlat célja:** a nemellenőrzött tanulás és a féligellenőrzött tanulás egyes tipikus algoritmusainak vizsgálata, kísérletezés az algoritmusok egyes paramétereivel a készségszintű használat céljából.

**Használt eszközök:** MATLAB környezet, és abban előre megírt speciális eljárások.

**Az otthoni felkészüléshez szükséges irodalom:**

1. Általában: a gyakorlat weblapjára feltett felkészülési anyagok.

2. Nemellenőrzött tanulás témában: *Altrichter-Horváth-Pataki-Strausz-Takács-Valyon:* ***Neurális hálózatok****, Panem Könyvkaidó Kft., 2006*. (döntően a 10. fejezet)

3. Féligellenőrzött tanulás témában: *Sugato Basu:* ***Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments*** (döntően az 1.-3. fejezetek)

## FELADATOK AZ ELŐZETES OTTHONI FELKÉSZÜLÉSHEZ:

*Olvassa el a gyakorlati feladatokat is, és ezek ismeretében írja meg az előzetesen elkészítendő programokat!*

**1.** Ismerje meg a `SOM_1.m`, önszerveződő térképet (*self organizing map*) megvalósító programot, értse meg a működését!

> *Segítség:* a program elején 6 csoportba generálunk 3-dimenzós Gauss-eloszlású adatvektorokat. A 3-dimenziós adatokat a vizualizáció kedvéért egy kijelzett színes pont színkomponenseinek fogjuk fel (RGB). A tanulás során a klasszikus Kohonen térképtől eltérően a reprezentáló mintavektorok fixen állnak az (1,1)…(Nsor,Noszlop) rácspontokban, és a színinformáció változik. Tehát 1-1 mintavektor színe azon adatvektorok színinformációjának átlaga, amelyeknél ő a győztes. (A tanítás során a győztes szomszédságában lévő mintavektorokat is módosítjuk.)

**2.** Írjon egy olyan m-fájlt, amellyel kijelezhető a megtanult MM mintavektor-mátrix a megtanult színekkel a megfelelő rácspontokban, és bekarikázza azokat a rácspontokat, amelyekhez tartozó adatvektorhalmaz legalább 80%-ban homogén. Ez alatt azt értjük, hogy azon adatvektorok közül, amelyekre ez a csomópont a győztes, legalább 80% azonos csoportba tartozik.
Opcionálisan másik hf program (a kettőből az egyiket kell megírni): Írjon egy olyan pársoros m-fájlt, amellyel kijelezhető a megtanult MM mintavektor-mátrix elemeinek a közvetlen szomszédjaiktól vett átlagos távolsága! Építse be ezt a SOM_1.m-be úgy, hogy menet közben jelezze ki ezt az értéket a hibagörbe (figure(2)) fejlécében!

**3.** Ismerje meg a K-átlagképző (*K-means*) algoritmust megvalósító `KAtlagkepzo.m` programot, értse meg a működését!

> *Segítség:* A program nemellenőrzött, féligellenőrzött vagy ellenőrzött K-átlagképző osztályozás megvalósítására képes. 6 csoportba generálunk 3-dimenziós Gauss-eloszlású adatvektorokat. A tanítás során megadható, hogy hány klaszterbe kívánjuk sorolni az adatokat, és az adatok hány százaléka ismert besorolású. (100% - ellenőrzött tanulás esete, 0% - nemellenőrzött tanulás esete, a kettő közt – féligellenőrzött tanulás esete.) Féligellenőrzött tanulás esetén minden tanítási ciklusban felhasználjuk (nem változtatjuk meg) az ismert adatok besorolását.

**4**. Írjon egy olyan m-fájlt, amellyel `KAtlagkepzo.m` többször meghívható ugyanazon paraméterekkel, és a végén visszaadja az elért átlagos lépésszámot, az elért átlagos hibát, illetve a hiba szórását!

**5.** Módosítsa úgy a `KAtlagkepzo.m` programot, hogy zajt visz az ismert adatok besorolásába (kimeneti zaj), és tegye lehetővé (futtatáskor választható módon), hogy csak az első tanítási ciklusban használjuk fel az ismert besorolás információt!

**6.** Ismerje meg a valós (elektromos fogyasztási) adatok klaszterezését végző `KAtlagkepzoFogyasztasra.m` programot, értse meg a működését!

*Segítség:* A program nemellenőrzött, féligellenőrzött vagy ellenőrzött K-átlagképző osztályozás megvalósítására képes. Egyhetes valós elektromos fogyasztási adatokból paramétereket képeztünk (Délelőtti fogyasztás, délutáni fogyasztás, hétvégi fogyasztás), ezek alapján akarjuk klaszterezni a fogyasztókat. A fogyasztók három scoportba tartoznak: iskola, üzem és iroda.Gyakorlatilag az előzőekben megismert és használt K-átlagképző eljárás speciális, konkrét alkalmazásáról van szó. A tanítás során megadható, hogy az adatok hány százaléka ismert besorolású. (100% - ellenőrzött tanulás esete, 0% - nemellenőrzött tanulás esete, a kettő közt – féligellenőrzött tanulás esete.) Féligellenőrzött tanulás esetén minden tanítási ciklusban felhasználjuk (nem változtatjuk meg) az ismert adatok besorolását.

**Ellenőrző kérdések:**

**1.** Mi történik,ha az önszerveződő térkép esetén nem tanítjuk a szomszédságot, csak a győztest?

**2.** A `SOM_1` programban az `Nsz0` paramétert 5-re állítjuk, miközben `Nsor= 10`, `Noszlop=10`. Az első tanítási ciklusban az `MM` mintavektor mátrix hány eleme fog változni, amikor a 3,4 mintavektor a győztes?

**3.** A K-átlagképző osztályozási algoritmus milyen féligellenőrzött változatait ismertük meg a Gépi tanulás tárgyban?