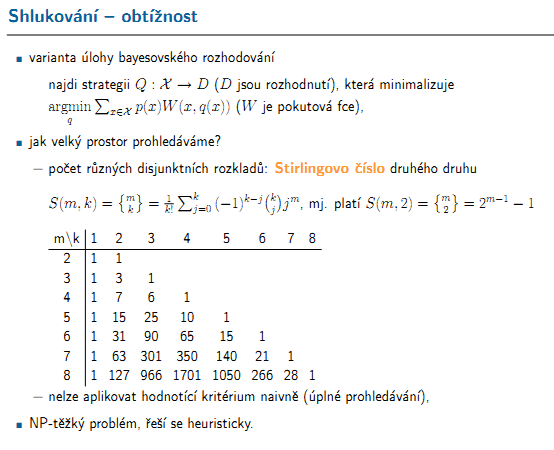
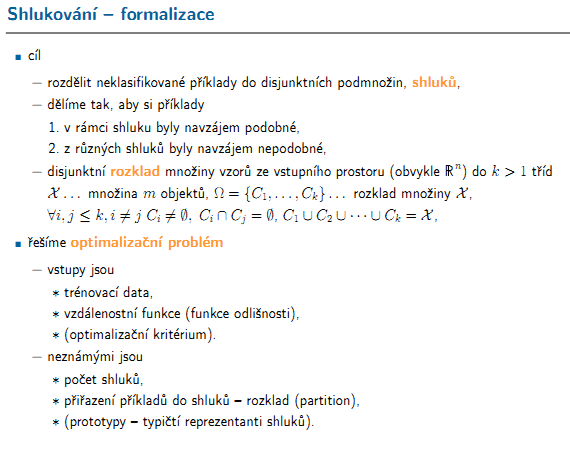
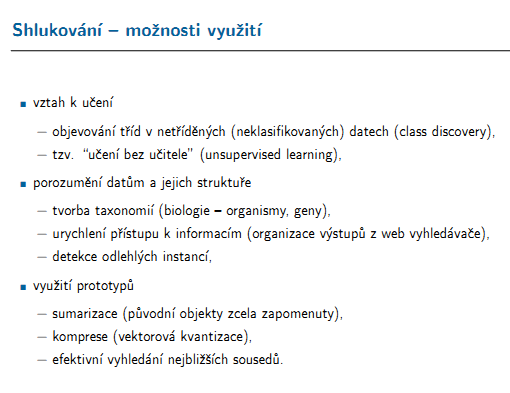
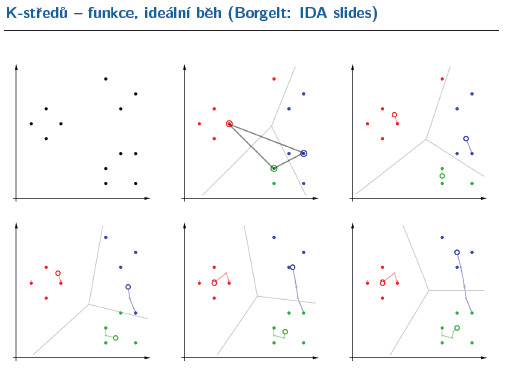
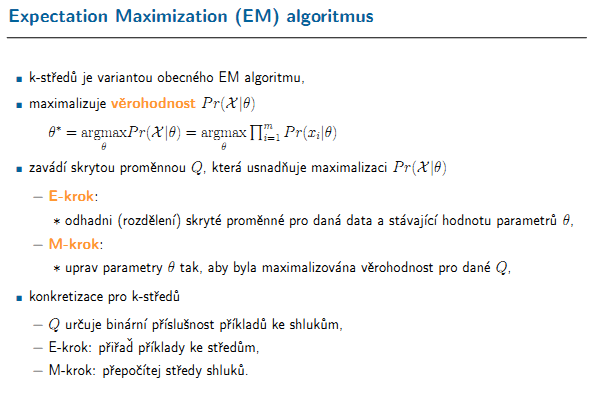
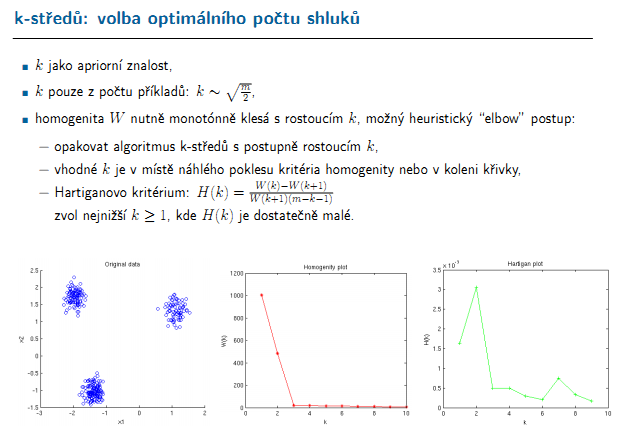
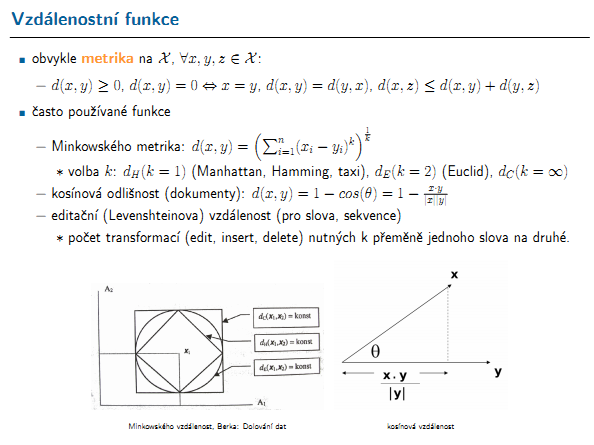
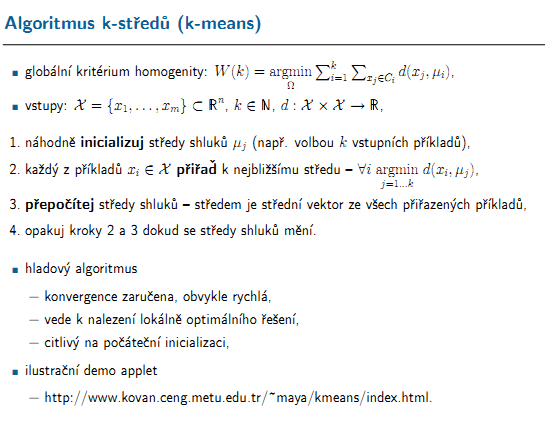
# 12. Shluková analýza: algoritmus k středů, hierarchické shlukování. Analýza hlavních a nezávislých komponent.



**(7)** **shlukování obtížnost**

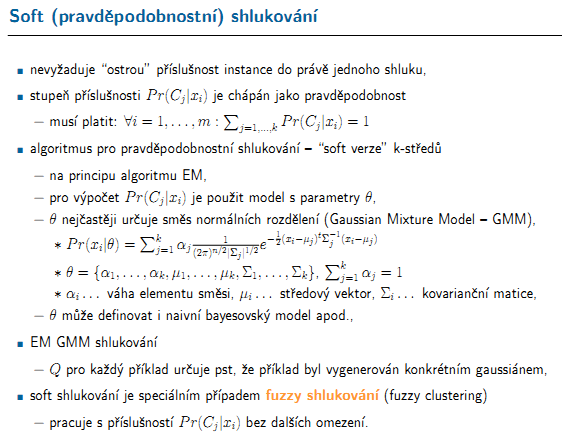
strategie je asi zpusob rozdeleni do shluku, kazde rozdeleni mohu nejak ohodnotit, tak je nejaka pokutova funkce, ktera kazdemu rozdeleni da nejakou pokutu, takze chceme najit takove rozdeleni, aby pokuta byla **nejmensi**, tedy nejlepsi rozdeleni.



**(13)**jaky je rozdil mezi EM a k-means? **(dále na 16)**

kmeans je varianta EM, jak vypada EM, ktery neni k means ? :)

napadlo me, ze treba to EM funguje i kdyz bych hledal shluky pomoci elips apod.. K-means hleda jen pomoci kruznic ze jo? jj



**(14) Soft**

specialni pripad **fuzzy shlukovani**, hmm :-D co toje? no, fuzzy logika je treba takova, ze nemas 1 / 0 = true / false, ale hodnoty od 0 do 1, treba 0.8 - spise ano, takze to vyuziva asi tohoto, nemas proste prislusnost ke shluku 100%, ale treba nejaky sample patri 80% do jednoho shluku a 20% do druheho

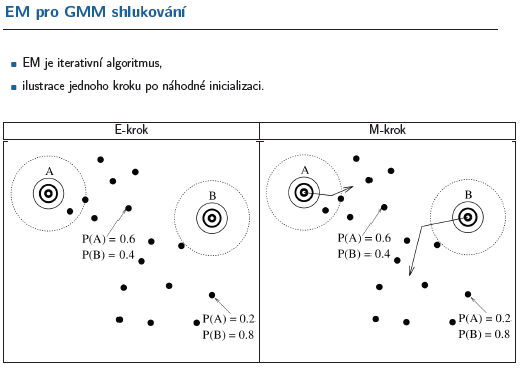
**Pr(Cj | xi) pravdepodobnost, ze bod xi patri do shluku Cj**

a dole pod tim je, ze soucet Pr pres vsechny shluky je 1, tedy **bod 100% nekam patri**, takze kdybych mel 3 shluky, tak napr. bod bude patrit 70% do jednoho, 20% do druheho a 10% do tretiho, soucet pres vsechny je 100%

**Vzorečky**

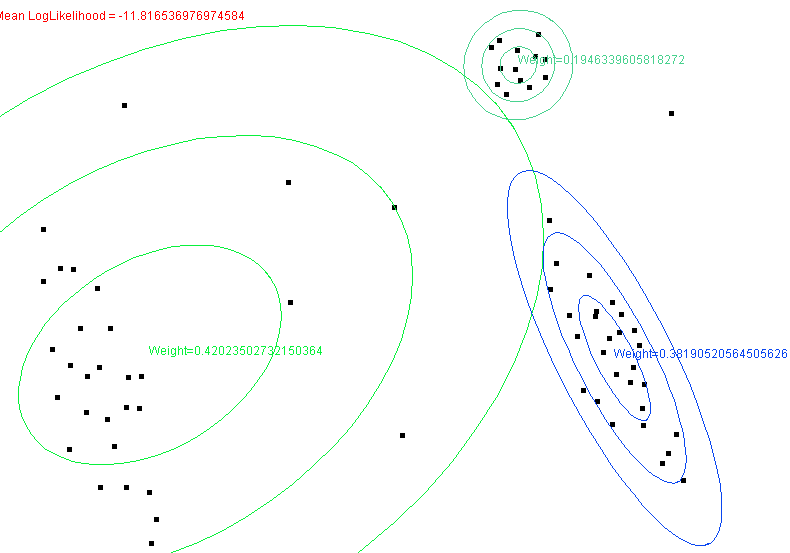
pak se tam rozebiraji **specialni pripady soft algoritmu**, treba **soft k-means,** tak ze se to resi pomoci normalniho rozdeleni, to mas nejak takhle: **k\*exp(-x^2/sigma^2),** oni to tam maji nejak normalizovane, to co jsem napsal je takove obecne normalni rozdeleni, proste kdyz je x = 0 (**vzdalenost**), tak e^0 = 1, prislusnost je 100%, kdyz je bod dale, x > 0, tak e^-(x^2/sigma^2) bude mensi nez 1, takze cim je bod dale od stredu shluku, tim mene do nej patri

oni maji ten vztah docela slozitej, protoze je v tom asi zahrnuto, ze **suma pres vsechny shluky je 1**

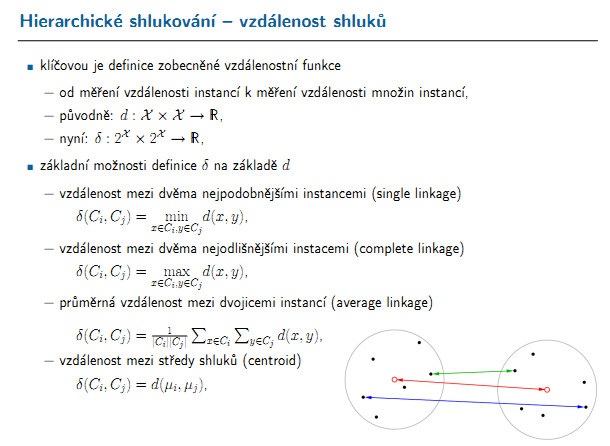
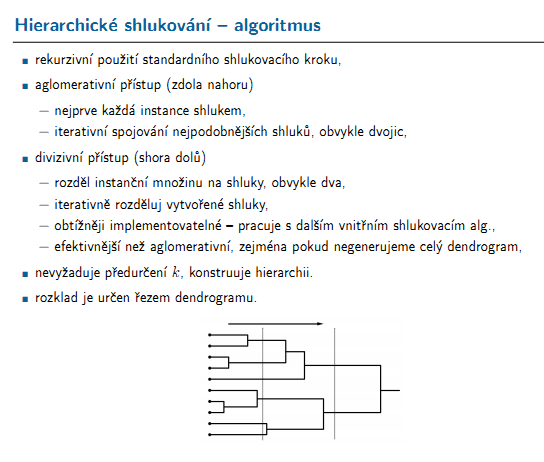
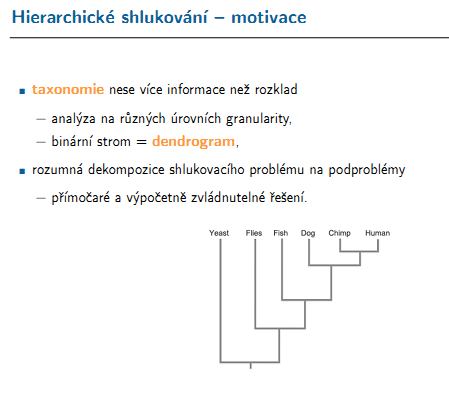
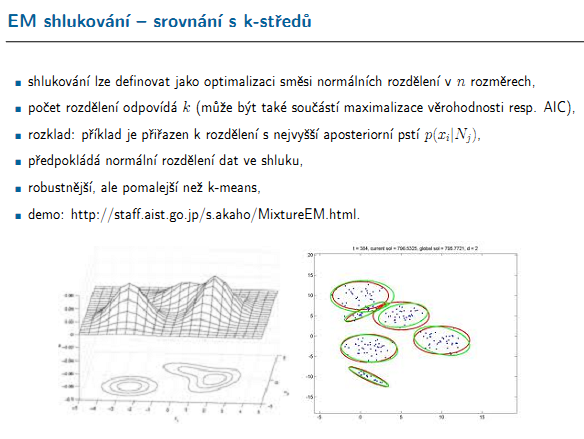
****

**(15) obr, proč se nepřepočítavájí P?**

rekl bych ze tady mame zatim jen 2 kroky E a M - E prepocitej P a M - posun stredy, takze ted by nasledoval zase krok E a teprve pak by byl prepocet



<http://www.socr.ucla.edu/Applets.dir/MixtureEM.html>



**(19, 20) Single, Complete, Centroid**

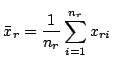
vzdy bereme **minimalni vzdalenost dvou shluku**, vzdalenost dvou shluku je ruzne definovana:  
**single** - vzdalenost dvou nejblizsich bodu ve shlucich  
**complete** - vzdalenost dvou nejvzdalenejsich bodu ve shlucich

**centroid** - vzdalenost dvou stredu (centroidu) ve shlucich, *muze to byt pomyslny bod (tedy zadny realny)*

The distance between two clusters is the Euclidean distance between their **centroids**, as calculated by arithmetic mean.

**

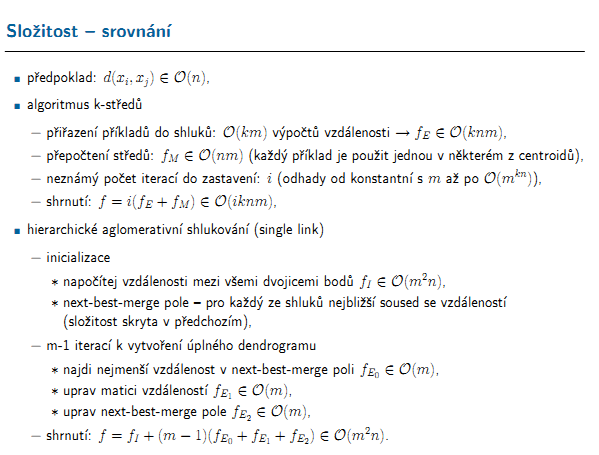
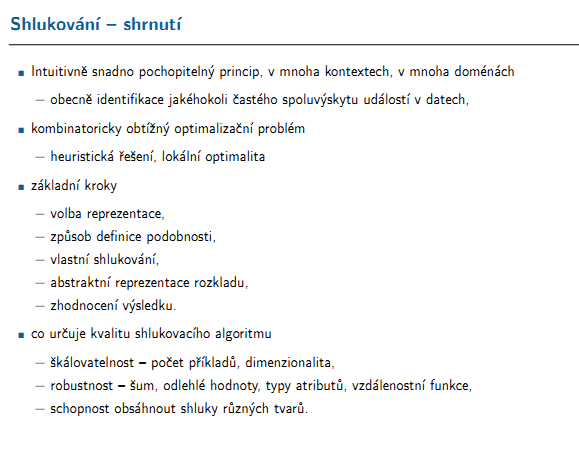
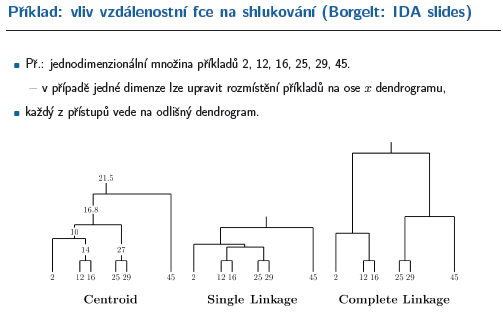
where xr is the centroid of r by arithmetic mean (ze vsech vzdalenosti bodu ve shluku):



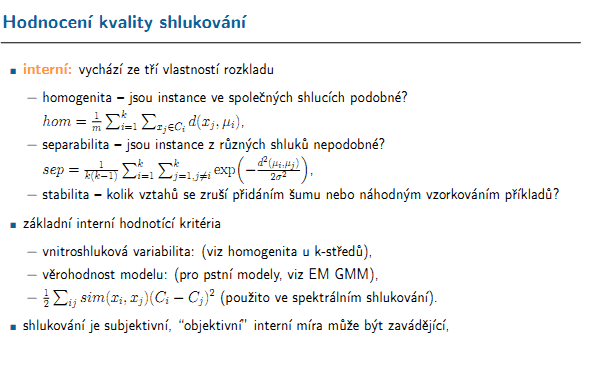
**average** – průměrná vzdálenost mezi dvojicemi instanci, vemu vsechny dvojice - napocitam jejich vzdalenosti a z toho **udelam prumer**

jinak u average ty vzdalenosti dvojic uz znam v matici vzdalenosti, takze to neni moc slozite, uz je to predpocitane, jen zprumerovat (tuto vlastnost mam i u ostatnich rezimu)

U **centoridu**: ta desítka: 2+12+16 / 3

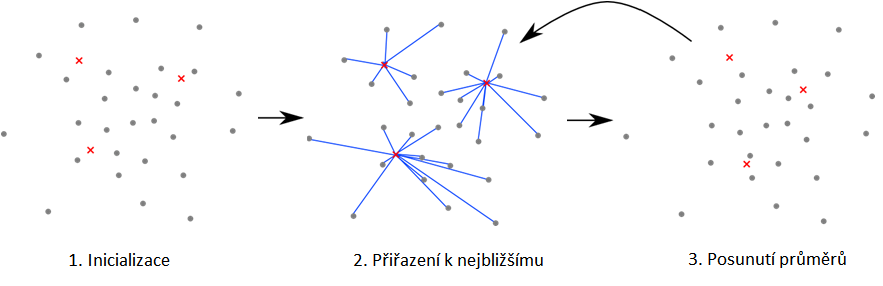


**(4) Složitost – srovnání**

nahodne mam nejake shluky, takze potrebuju zjistit vzdalenost kazdeho bodu od kazdeho shluku, **k je pocet shluku** a **m pocet bodu**, takze to je km, pak nejakej odhad poctu behu.. to nevim..prostě zapamatovat

**(5) Hodnocení kvality shlukování**

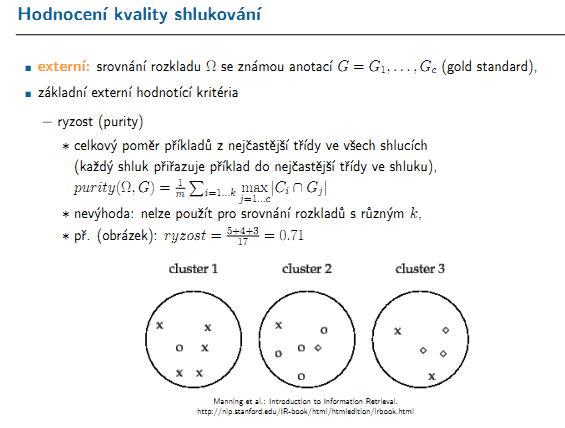
**homogenita -** asi nejaka mira podobnosti, asi to bude prumerna vzdalenost bodu od shluku ke kteremu patri, pro kazdy shluk nascitam vzdalenosti od bodu, ktere k nemu patri a pak jeste to vsechno poscitam pres vsechny centroidy a vydelim poctem bodu, takze tak bych to chapal - **prumerna vzdalenost bodu od centroidu**

****

prumerna vzdalenost tech modrych primek

**separabilita -** cim budou shluky **vzdalenejsi** tim bude **separabilita nizsi,** mi jsou stredy shluku, prochazime vsechny dvojice shluku, vypocteme jejich vzdalenost d(mi\_i,mi\_j), dame to do normalniho rozdeleni, vsechno nascitame a vyprumerujem, proč tam má nadruhou to d? to je normalni rozdeleni exp(-x^2/sigma^2)

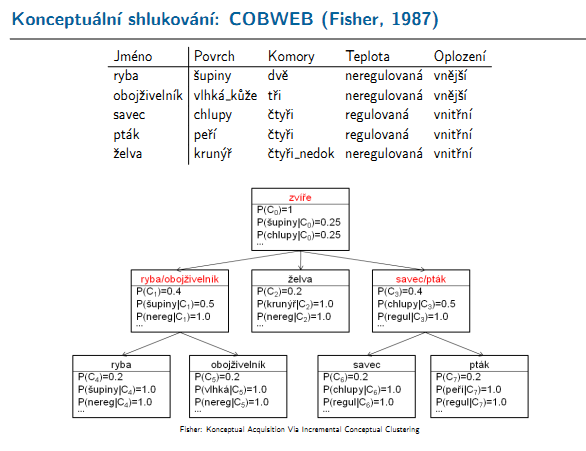
jo už to tam vidím ... ok, takze kdyz se budou vsechny shluky prekryvat, tak d = 0 nebo hodne blizke nule, exp ^ 0 = 1 podle gaussovske krivky, takze nascitame jednicky a vydelime k(k-1) - poctem vsech dvojic a dostaneme prumer 1, **takze 1, kdyz jsou shluky prilis u sebe a prekryvaji se, 0 kdyz budou hodne vzdalene od sebe**



**(6) Gold standard**

asi ze **zname jak by klasifikace mela dopadnout a porovnavame to s tim, jak to dopadlo** (viz. obr), na tom obrazku vidime samply, jsou 3 typy x, o a d (diamond), asi toto, protoze mas u toho G1, G2, ... a tady v tech vypoctech taky pouzivat Gj

**gold standard** je srovnani se znamou anotaci, srovnani je napr. pomoci ryzosti a dalsich hodnoticich hodnot



**(11) COBWEB – Konceptuální shlukování**

http://www-ai.cs.uni-dortmund.de/kdnet/auto?self=$81d91eaae317b2bebb

jednotlive samply (příklady) jsou listy ve strome, ty se spojuji do skupin - uzlu nad nimi, do skupiny (do uzlu) patri podskupiny (poduzly) s podobnymi vlastnostmi, tim jak prochazis sample po samplu, tak se stromem provadis operace

1) zaradis list do uzlu

2) vytvoris uzel (skupinu) a do nej zaradis list

3) spojis vice uzlu do jednoho

4) rozdelis jeden uzel do vice

a podobne vlastnosti se urcuji podle pravdepodobnosti, tedy kolik listu ma spolecnou vlastnost, kdyz kliknes na vnitrni uzel v appletu tak se tam vypise pocet listu, ktere maji shodnou vlastnost, takze to provadi nejake vypocty, ktere uzly jsou si podobne, ktere ne a kam to ma zaradit