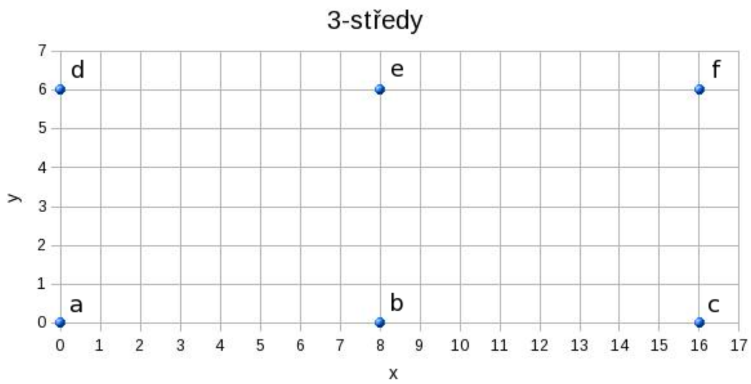
Vzorový test pro první část předmětu M33SAD

Shlukování, vyhledávání častých vzorů

1 Shlukování

Použijte algoritmus 3-středy na množinu příkladů $\mathcal{X} = \{a = (0,0), b = (8,0), c = (16,0), d = (0,6), e = (8,6), f = (16,6)\}$. Jde o prostor $\mathbb{R}^2$, algoritmus bude používat euklidovskou vzdálenost. Při shodě vzdáleností algoritmus upřednostní centroid směrem vlevo dole. Za počáteční konfiguraci označíme jakoukoli podmnožinu $\mathcal{X}$ o kardinalitě 3, půjde o počáteční volbu centroidů. Za 3-rozklad označíme jakýkoli disjunktní rozklad $\mathcal{X}$ na tři neprázdné podmnožiny (např. $\Omega = \{\{a,b,e\},\{c,d\},\{f\}\}$). Každá počáteční konfigurace jednoznačně definuje 3-rozklad. 3-rozklad je stabilní, jestliže iterací algoritmu 3-středů nedojde ke změně 3-rozkladu (a tím ani centroidů). Zodpovězte následující otázky:

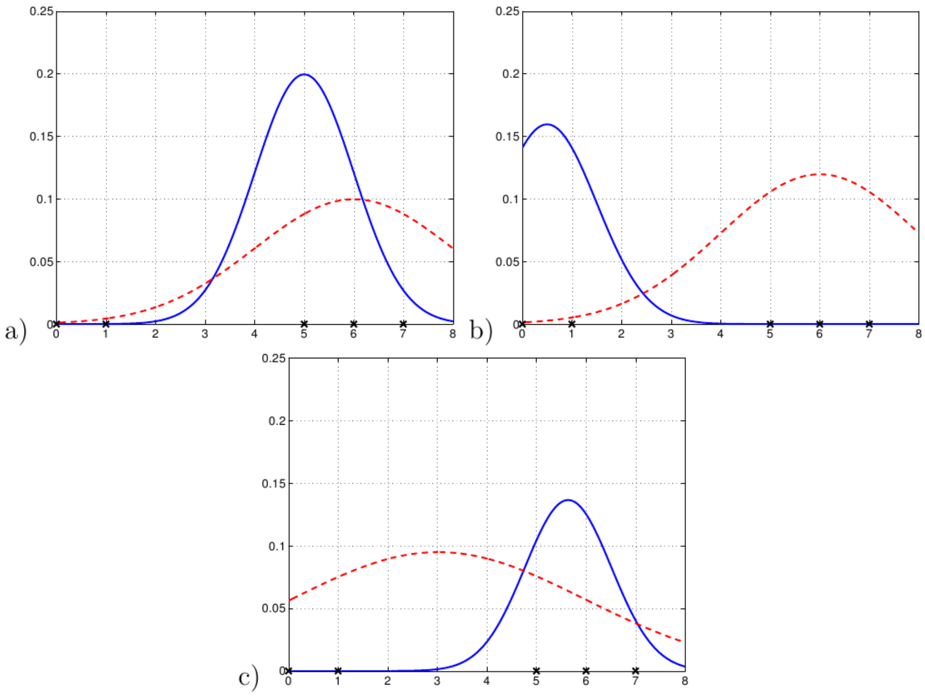


1. (1 bod) Kolik existuje různých počátečních konfigurací?
2. (0 bodů) Kolik existuje různých 3-rozkladů? (pro zajímavost)
3. (1 bod) Které 3-rozklady dosažitelné z počátečních konfigurací jsou stabilní? Kolik jich je?
4. (1 bod) Kolik počátečních konfigurací definuje stabilní 3-rozklad?
5. (1 bod) Jaký je maximální počet iterací algoritmu 3-středy z libovolné počáteční konfigurace do jejího stabilního 3-rozkladu?

2 EM algoritmus

Pomocí EM algoritmu odhadujete parametry směsi 2 normálních rozdělení. Rozdělení směsi podle příznaku x lze zapsat takto: $f(x, \theta) = \alpha N(x; \mu_1, \sigma_1^2) + (1 - \alpha)N(x; \mu_2, \sigma_2^2)$. Obrázky uvedené níže ilustrují kroky EM algoritmu (na horizontální ose je parametr x, na vertikální ose je hustota pravděpodobnosti, pozorování jsou značena křížkem). Na jednom z obrázků je uveden náhodný inicializační krok (*init*), na druhém je uveden první optimalizační krok (*step1*). Třetí z obrázků je navíc. Obrázky jsou seřazeny náhodně. Rozhodněte, která dvojice obrázků odpovídá uvedeným krokům *init* a *step1*. Vysvětlete, proč uvedené pořadí dává smysl a jak *step1* vychází z *init*.

hodnocení: 4 body (2b za určení správného pořadí, 2b za vysvětlení)



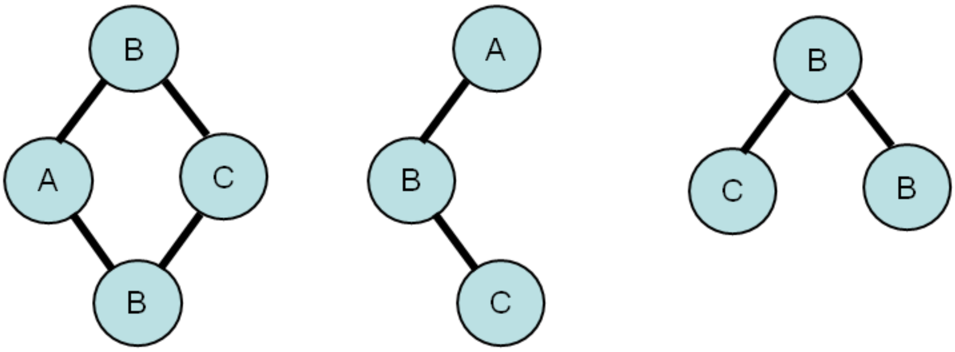
3 Časté podsekvence

Mějme abecedu dvou symbolů $\{a, b\}$. Uvažujme neorientované sekvence. Zodpovězte následující otázky:

1. (1 bod) Kolik existuje různých neorientovaných sekvencí délky 3?
2. (1 bod) Naznačte jak budete generovat různé sekvence délky 4. Ukažte alespoň jednu duplicitní sekvenci délky 4.
3. (1 bod) U sekvencí délky 3 jste ověřili, že časté jsou pouze sekvence $\{aab, bab, bbb\}$. Které sekvence délky 4 ještě stále mohou být časté? Proč?

4 Časté podgrafy

Mějme množinu tří grafů z obrázku. Vrcholy jsou anotované třemi značkami, hrany mají identické značky.



1. (2 body) Nakreslete strom všech možných podgrafů (každý musí být podgrafem alespoň jednoho grafu ze zadání).
2. (2 body) Uvažujte minimální podporu $s_{min} = 2$, vyznačte všechny uzavřené a maximální podgrafy.