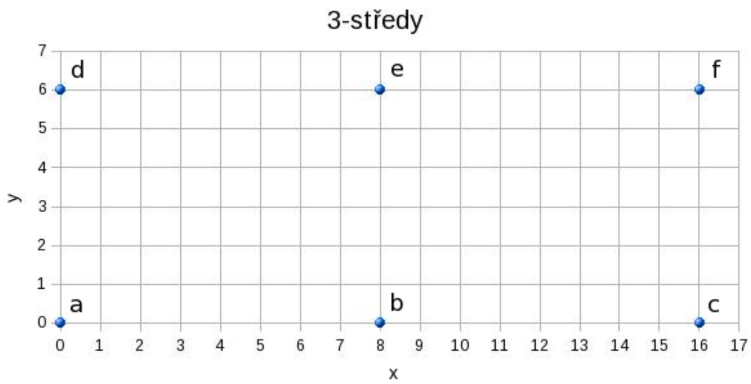
Vzorový test pro první část předmětu M33SAD

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?

Řešení:

1. Kombinace 3tí třídy z 6 prvků, bez opakování, nezáleží na pořadí prvků:

$$konf = \binom{m}{k} = \binom{6}{3} = 20$$

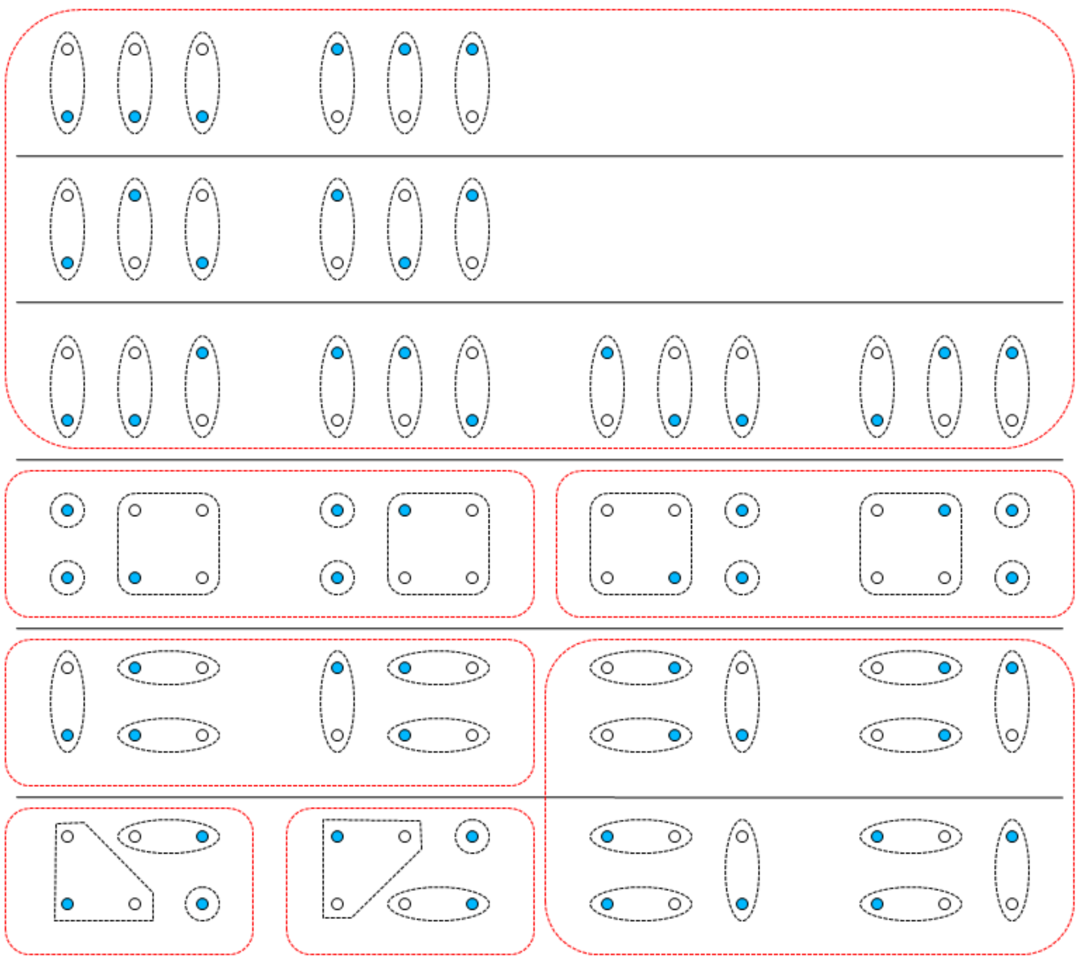
.

2. Stirlingovo číslo 2.druhu:

$$S(m,k) = \left\{ {m \atop k} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^m = 90$$

.

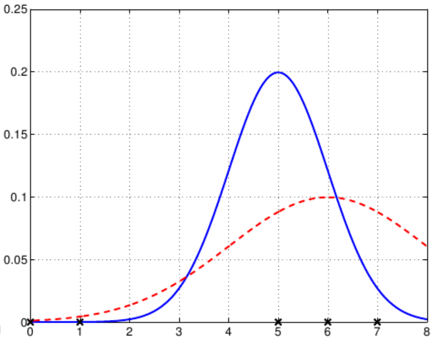
3. Uvažujeme-li symetrii, lze 20 počátečních konfigurací rozdělit na 6 tříd ekvivalence. Analýzou těchto tříd zjišťujeme, že existuje celkem 7 různých 3-rozkladů (z nichž 3 dvojice jsou symetrické), všechny 3-rozklady jsou stabilní. Viz obrázek níže – třídy ekvivalence jsou odděleny černými vodorovnými čarami, identické 3-rozklady jsou ohraničeny červenou čárkovanou čarou.
4. Všechny, plyne z předchozího.
5. Jediná iterace, která ověří stabilitu, plyne z předchozího.

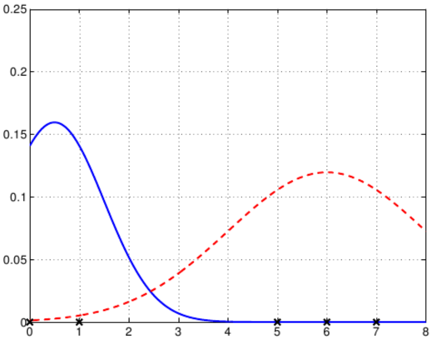


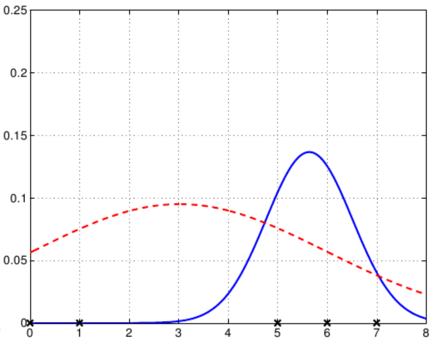
2 EM algoritmus

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)

a) 

b) 

c) 

Řešení:

Smysl dává pořadí $init = a$, $step1 = c$

Ilustrace odhadu indikátorové proměnné Z pro $init = a$ (**E krok**):

- složka 1 (f_1, $Z = 1$): modrá plná čára, složka 2 (f_2, $Z = 2$): červená čárkovaná čára,
- indikátorová proměnná určuje, který bod byl generován kterou složkou směsi (jde o skrytou proměnnou),
- hustoty psti f_1 a f_2 jsou přibližně odečteny z obrázku ad a),
- uvažujeme shodnou váhu elementů směsi $\alpha = 0.5$,
- indikátorové proměnné napočteny jako: $Pr(Z(x) = 1) = \frac{f_1}{f_1+f_2}$, $Pr(Z(x) = 2) = 1 - Pr(Z(x) = 1)$.

x	$\mathbf{f_1(x)}$	$\mathbf{f_2(x)}$	$\mathbf{Pr(Z(x)=1)}$	$\mathbf{Pr(Z(x)=2)}$
0	0	.001	0	1
1	.001	.009	.1	.9
5	.2	.08	.71	.29
6	.12	.1	.55	.45
7	.025	.08	.24	.76

Změna parametrů obou prvků směsi (**M krok**):

$$\overline{x}_1 = \frac{\sum_x x Pr(Z(x)=1)}{\sum_x Pr(Z(x)=1)} = \frac{0\times0+.1\times1+.71\times5+.55\times6+.24\times7}{0+.1+.71+.55+.24} = 5.4$$

$$\overline{x}_2 = \frac{\sum_x x Pr(Z(x)=2)}{\sum_x Pr(Z(x)=2)} = \frac{1\times0+.9\times1+.29\times5+.45\times6+.76\times7}{1+.9+.29+.45+.76} = 3$$

$$s_1 = \sqrt{\frac{\sum_x Pr(Z(x)=1)(x-\overline{x}_1)^2}{\sum_x Pr(Z(x)=1)}} = \sqrt{\frac{0\times5.4^2+.1\times4.4^2+.71\times0.4^2+.55\times0.6^2+.24\times1.6^2}{0+.1+.71+.55+.24}} = 1.3$$

$$s_2 = \sqrt{\frac{\sum_x Pr(Z(x)=2)(x-\overline{x}_1)^2}{\sum_x Pr(Z(x)=2)}} = \sqrt{\frac{1\times3^2+.9\times2^2+.29\times2^2+.45\times3^2+.76\times4^2}{1+.9+.29+.45+.76}} = 3$$

Očekáváme, že napočtené hodnoty $\overline{x}_1$, $\overline{x}_2$, s_1 a s_2 budou zhruba odpovídat obrázku c. Průměry evidentně odpovídají, směrodatné odchylky také (u normálního rozdělení by zhruba 2/3 hodnot měla ležet v rozsahu jedné směrodatné odchylky od průměru).

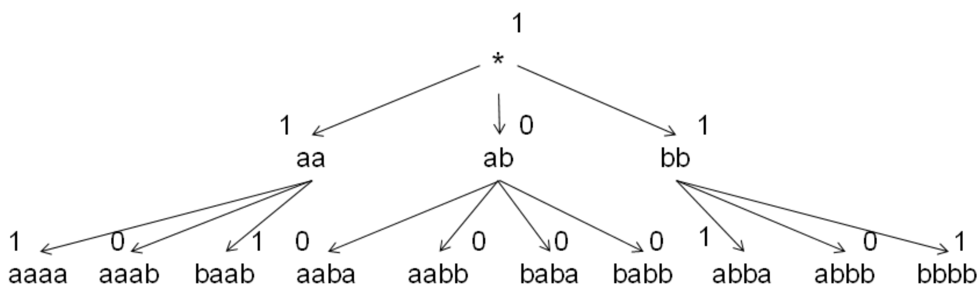
3 Časté podsekvence

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 strom, kterým budete generovat kanonické formy sekvencí délky 4. Ukažte alespoň jednu duplicitní sekvenci délky 4 (nekanonickou formu).
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?

Řešení:

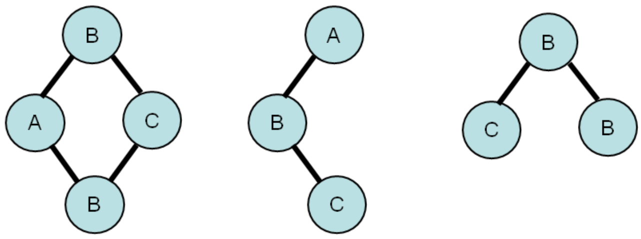
1. 6 sekvencí: $\{aaa, aab, bab, aba, abb, bbb\}$.
2. Neduplicitní neorientované sekvence délky 4 generujeme ze sekvencí délky 2, ty naopak ze sekvencí délky 0. Využíváme přitom příznaku symetrie, který určuje, zda mohu přiřazovat první a poslední symbol v lexikografickém (symetrie=1) nebo libovolném pořadí (symetrie=0). Strom je naznačen níže, sekvencí délky 4 je 10. Duplicitní je například $baaa$ vzhledem k $aaab$.



3. Na základě generalizovaného APRIORI prořezávání zjišťujeme, zda sekvence má častý prefix i zakončení délky 3. Kandidátskými sekvencemi délky 4 jsou: $\{baab, bbbb\}$. První má prefix baa (kanonická forma aab) a zakončení aab. Druhá má prefix i zakončení bbb.

4 Časté podgrafy

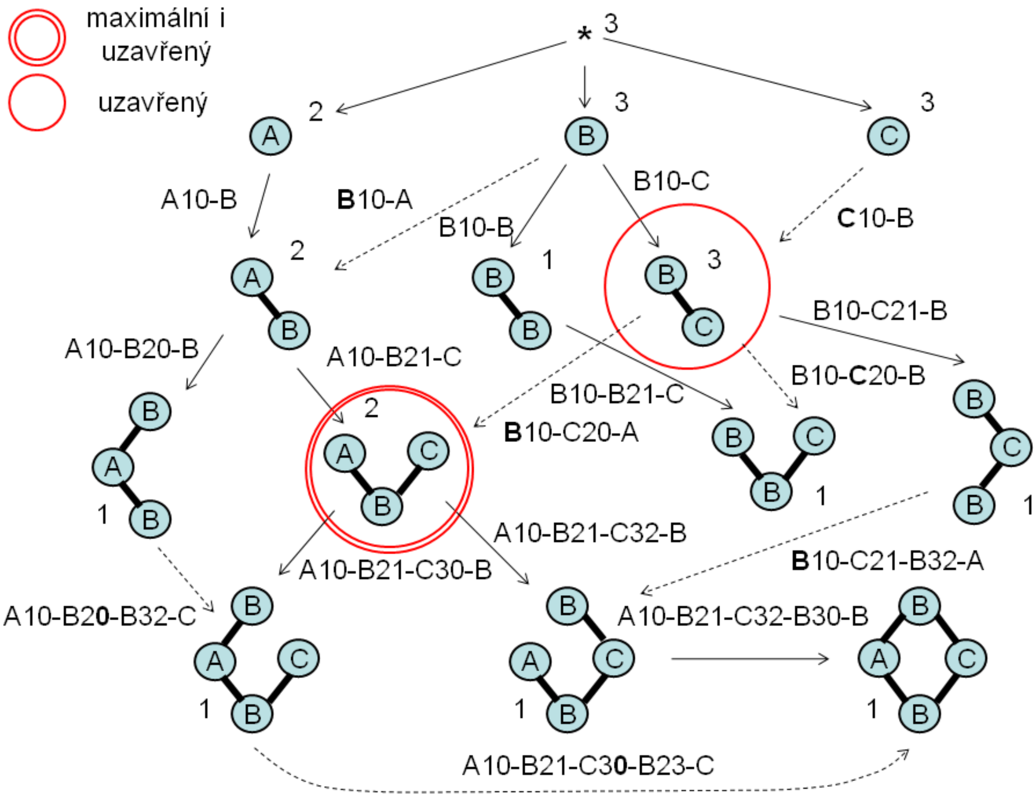
Mějme množinu tří grafů z obrázku. Vrcholy jsou anotované třemi značkami, hrany mají identické značky.



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.

Řešení:

1. Strom je konstruován metodou do hloubky. Regulární výraz pro kódová slova je: $a\,(i_d\,\underline{i_s}\,b\,a)^m$, pro všechny znaky jsou upřednostněny lexikograficky menší symboly, pouze u i_s vynutíme kvůli prohledávání do hloubky opačné pořadí. Čárkované hrany jsou odmítnuté, generují větší než minimální kódové slovo a vedou na redundantní (dříve navštívený) podgraf. Viz obrázek.
2. Dva uzavřené podgrafy a jeden maximální podgraf. Viz obrázek.


maximální i uzavřený
uzavřený
* 3
A 2
B 3
C 3
A10-B
B10-A
B10-B
B10-C
C10-B
A B 2
B B 1
B C 3
A10-B20-B
A10-B21-C
B10-B21-C
B10-C20-A
B10-C21-B
B10-C20-B
B A B 1
A B C 2
B C B 1
B C B 1
A10-B21-C32-B
A10-B21-C30-B
B10-C21-B32-A
A10-B20-B32-C
A10-B21-C32-B30-B
B A C B 1
B A C B 1
B A C B 1
A10-B21-C30-B23-C