

ZÁPADOČESKÁ UNIVERZITA V PLZNI

DIAGNOSTIKA A ROZHODOVÁNÍ

KKY/DR

4. semestrální práce

Autor:
Marek Lovčí

November 21, 2020



E-M algoritmus pro odhad parametrů Gaussovské směsi

Předmět zadání

Proveďte odhad parametrů Gaussovské směsi $p(\vec{x}|\vec{\lambda}) = \sum_{m=1}^M c_m N(\vec{x}|\vec{\mu}_m, C_m)$ o neznámém počtu složek M , kde $\vec{\mu}_m$ je vektor středních hodnot m -té složky, C_m je kovarianční matice m -té složky, $\vec{\lambda}$ je vektor všech parametrů směsi a \vec{x} je posloupnost pozorování (2 rozměrných vektorů). Odhad proveďte algoritmem očekávání - maximalizace.

- Posloupnost pozorování \vec{x} načtete ze souboru `sp4_data.mat`, případně ze souboru `sp4_data.csv`, pokud vám MATLAB nevyhovuje (každý řádek obsahuje obě čárkou oddělené složky pozorování).
- Implementujte E-M algoritmus. **Zdůvodněte případné použití zjednodušujících předpokladů** (počet složek směsi, tvar kovarianční matice, atd.).
- Proveďte implementovaný algoritmus nad poskytnutými daty. Uvažujte zastavovací podmínku euklidovské vzdálenosti předcházejících a nově odhadnutých parametrů $\|\vec{\lambda}_i - \vec{\lambda}_{i-1}\| < 10^{-3}$, kde i je číslo iterace algoritmu.
- Tabelujte **všechny** hodnoty $i, \vec{\lambda}_i, \|\vec{\lambda}_i - \vec{\lambda}_{i-1}\|$.
- Do grafu vynesete závislost $\|\vec{\lambda}_i - \vec{\lambda}_{i-1}\|$ na počtu iterací.

Použité nástroje

Simulaci proveďte v prostředí MATLAB, příp. naprogramujte ve vybraném programovacím jazyce.

Co se odevzdá

V referátu ve formátu PDF slovně komentujte vaše řešení, vč. zdůvodnění použitých předpokladů. Součástí referátu bude právě jedna tabulka a jeden graf dle zadání. Spolu s referátem odevzdejte pro posouzení komentovaný programový kód, který byl k řešení použit. Dbejte na splnění všech bodů zadání.

1 Vypracování

EM (Expectation-Maximization) algoritmus je jeden ze základních a velice efektivních přístupů, jenž je základem mnoha algoritmů strojového učení. V konečném důsledku dochází k odhadu Gaussovských distribučních funkcí, které jsou generátory datové sady.

Samotný algoritmus se skládá ze 3 kroků:

1. E-step;
2. M-step;
3. Výpočet chyby a návrat na bod 1.

Před samotným výpočtem musíme stanovit prvotní odhad. Podíváme-li se na rozložení dat (obrázek 1), dokážeme odhadnout, že se v datech nachází 2 shluky.

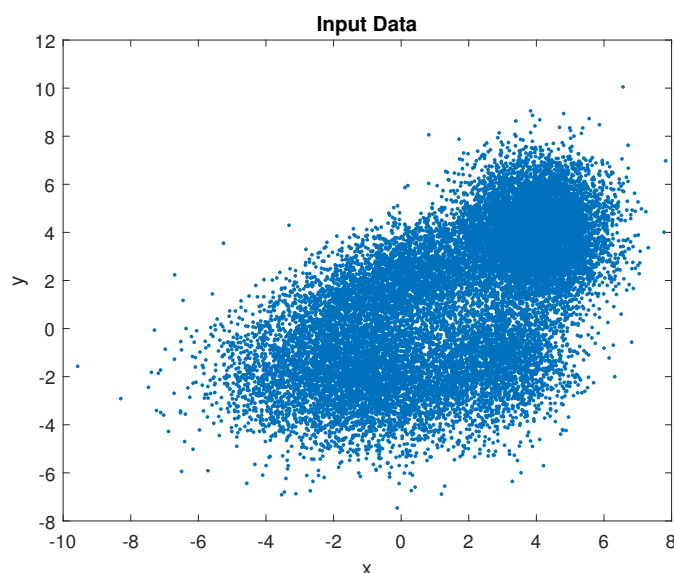


Figure 1: Vstupní data

Předpokládám symetričnost prvků Gaussovské směsi, tudíž $\Sigma = I$, protože její asymetričnost sice mohu předpokládat, ale nedokázal bych odhadnout jakým způsobem je asymetrická. Středů shluků μ_1, μ_2 byly určeny algoritmem *k-means*, tudíž prvotní odhad je i řešením ke kterému bychom měli po několika iteracích dojít. Počáteční váhy Normálních rozdělání byly zvoleny jako $\frac{1}{\text{počet tříd}} \pm 0.1$, protože z obrázku vyplývá, že váha každé ze tříd bude rozdílná.

Výpočet byl proveden dvěma způsoby. První výpočet je kontrolní, pouze s využitím funkcionalit softwaru MATLAB. Výsledek tohoto výpočtu je na obrázku 2.

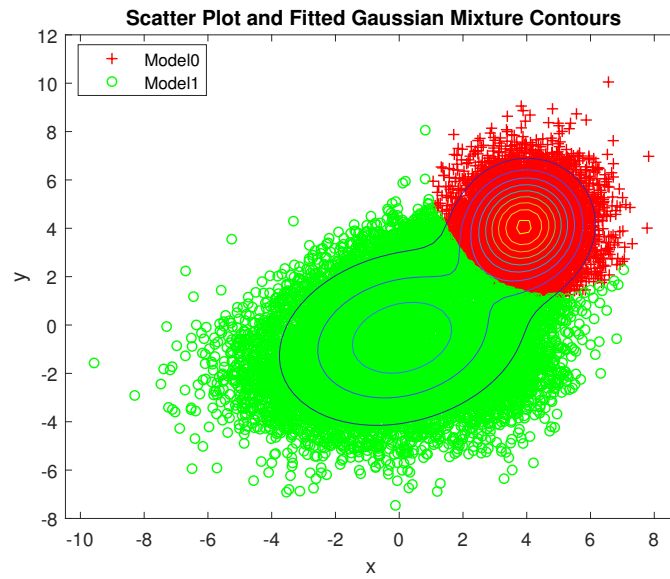


Figure 2: Výsledné rozdělení dat dle nativního EM algoritmu

Graf 3 zobrazuje výsledek po výpočtu implementovaným E-M algoritmem.

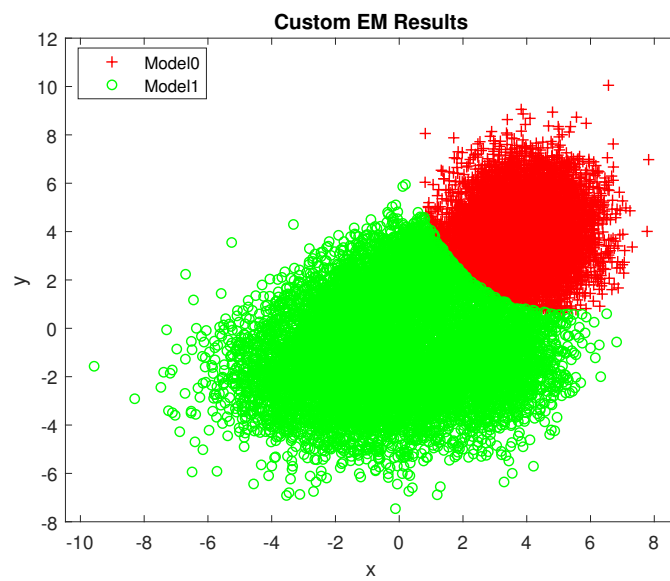


Figure 3: Výsledné rozdělení dat dle implementovaného E-M algoritmu

Tabelované hodnoty dle zadání jsou v tabulce 1.

Iterace	Chyba	μ_1	μ_2
1	0.0537	[-0.4510 -1.2142]	[3.5411 3.6698]
2	0.3032	[-0.3409 -1.1012]	[3.6216 3.7909]
3	0.2412	[-0.2684 -0.9977]	[3.7013 3.8736]
4	0.1674	[-0.2227 -0.9211]	[3.7617 3.9233]
5	0.1226	[-0.1919 -0.8618]	[3.8086 3.9536]
6	0.0952	[-0.1675 -0.8166]	[3.8453 3.9777]
7	0.0638	[-0.1499 -0.7863]	[3.8689 3.9941]
8	0.0411	[-0.1373 -0.7663]	[3.8831 4.0043]
9	0.0245	[-0.1304 -0.7548]	[3.8921 4.0107]
10	0.0123	[-0.1274 -0.7487]	[3.8970 4.0131]
11	0.0037	[-0.1262 -0.7470]	[3.8982 4.0143]
12	0.0025	[-0.1253 -0.7459]	[3.8989 4.0149]
13	0.0007	[-0.1252 -0.7455]	[3.8993 4.0150]

Table 1: Tabulované hodnoty parametrů pro každou iteraci

Závislost chyby na počtu iterací je v grafu 4. Prvotní “velmi” přesný odhad (tedy malá chyba) je způsobena tím, že první odhad je vypočtem algoritmem *k-means*. Po následné reevaluaci rozložení dat ve shlucích jsou středy přepočteny a jelikož se jedná o teprve druhý krok algoritmu, tak tím hohužel zne přesněny.

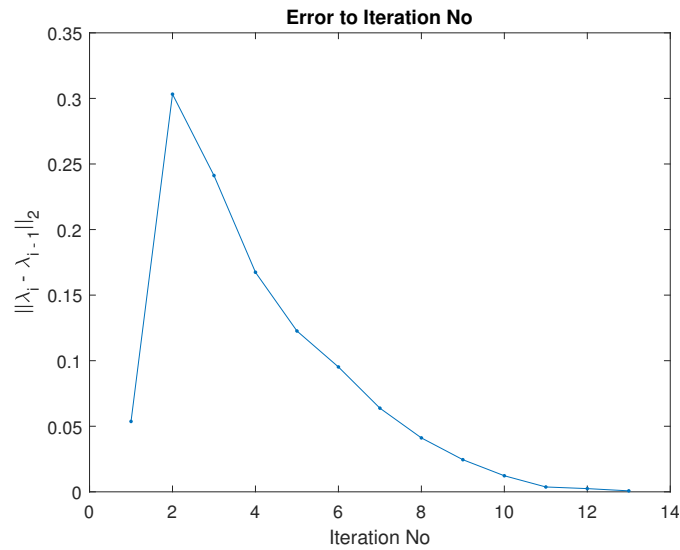


Figure 4: Hodnota chyby v každém uplynulém iteračním kroce

2 Závěr

Při řešení práce byly použity dvě varianty výpočtu. Jelikož se tyto varianty shodují, lze předpokládat, že dostáváme správné výsledky. Algoritmus dokovergoval ve 13. iteraci splněním zastavovací podmínky.