Západočeská univerzita v Plzni

Diagnostika a rozhodování

KKY/DR

4. semestrální práce

Autor:
Marek Lovčí

December 2, 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 $\vec{x}$ načtěte 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 vyneste 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.

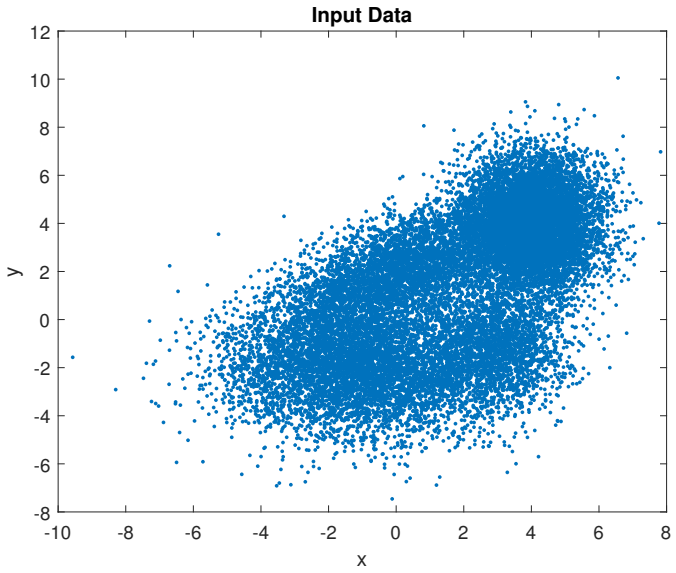

Figure 1: Vstupní data

Předpokládám symetričnost prvků Gaussovské směsi, tudíž $\Sigma = I$. Středy 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 rozdělení byly zvoleny jako $\frac{1}{\text{počet tříd}} = \frac{1}{2}$.

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

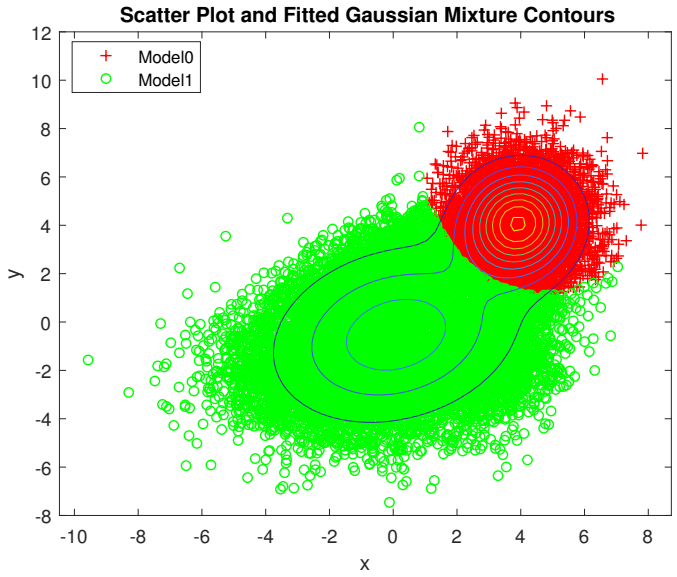

Figure 2: Výsledné rozdělení dat dle nativního EM algorimu

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

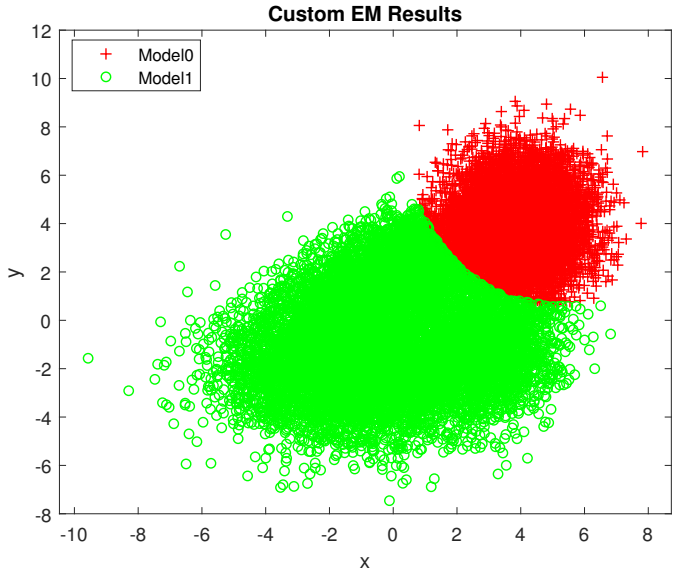

Figure 3: Výsledné rozdělení dat dle implementovaného E-M algorimu

Tabelované hodnoty dle zadání jsou v tabulce 1.

Iterace	Chyba	μ_1	μ_2	λ_1	λ_2
1	0.0285	[-0.4510 -1.2142]	[3.5411 3.6698]	0.5060	0.4940
2	0.1449	[-0.3409 -1.1012]	[3.6216 3.7909]	0.4830	0.5170
3	0.1055	[-0.2684 -0.9977]	[3.7013 3.8736]	0.4663	0.5337
4	0.0748	[-0.2227 -0.9211]	[3.7617 3.9233]	0.4546	0.5454
5	0.0553	[-0.1919 -0.8618]	[3.8086 3.9536]	0.4460	0.5540
6	0.0413	[-0.1675 -0.8166]	[3.8453 3.9777]	0.4396	0.5604
7	0.0285	[-0.1499 -0.7863]	[3.8689 3.9941]	0.4351	0.5649
8	0.0174	[-0.1373 -0.7663]	[3.8831 4.0043]	0.4324	0.5676
9	0.0112	[-0.1304 -0.7548]	[3.8921 4.0107]	0.4306	0.5694
10	0.0032	[-0.1274 -0.7487]	[3.8970 4.0131]	0.4301	0.5698
11	0.0016	[-0.1262 -0.7470]	[3.8982 4.0143]	0.4299	0.5701
12	0.0013	[-0.1253 -0.7459]	[3.8989 4.0149]	0.4297	0.5703
13	0.0000	[-0.1252 -0.7455]	[3.8993 4.0150]	0.4297	0.5703

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 znepřesněny.

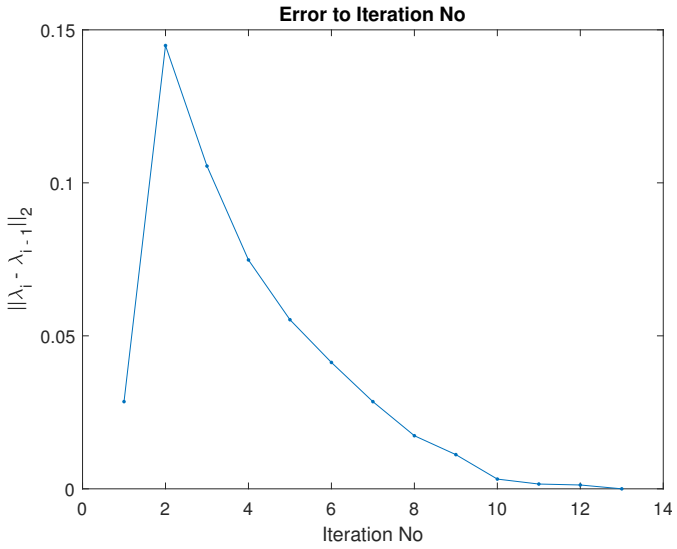

Figure 4: Hodnota chyby v každém uplynulém iteračním kroce

2 Závěr

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