Spektrální shlukování

Návod ke čtvrtému cvičení

Jiří Kléma, klema@labe.felk.cvut.cz

ZS 2011/2012

Úvod

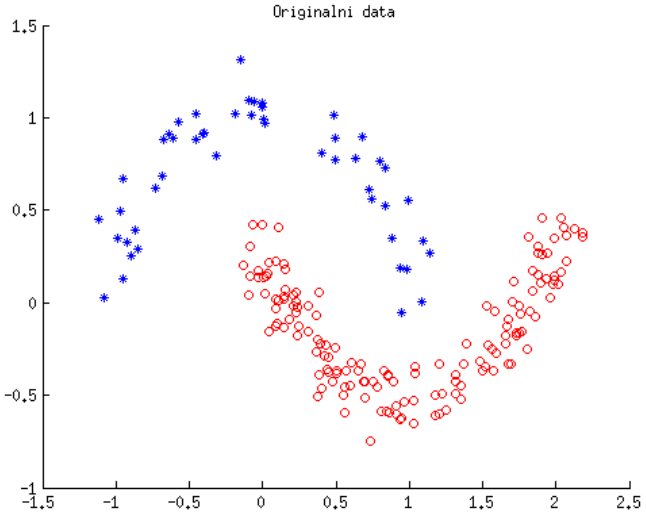
Cílem cvičení je seznámení se spektrálním shlukováním. Použijete dostupné funkční bloky a doprogramujete algoritmus spektrálního shlukování. Algoritmus aplikujete na vstupní data a výsledek srovnáte se známou anotací a s výsledkem klasického algoritmu k-středů. Srovnání provedete pro různé parametrizace vstupních dat a algoritmu spektrálního shlukování.

1 Vstupní data

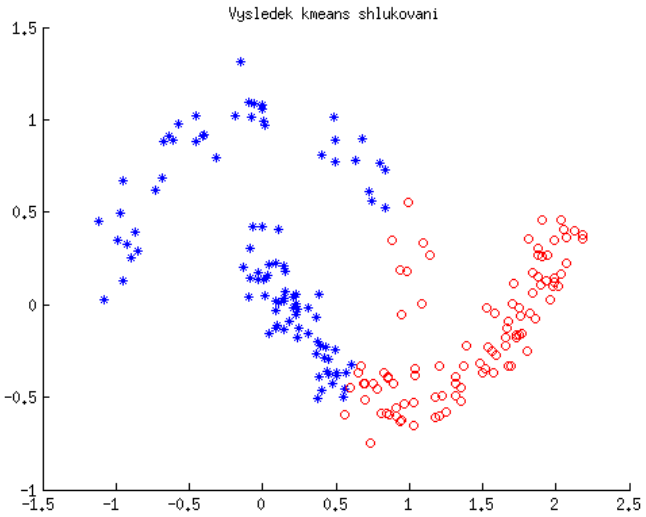
Vstupní data jsou generována funkcí *GenerateData.m*. Pevně daný je typ dat (půlměsíce), počet půlměsíců ($k = 2$) a počet reálných příznaků ($n = 2$). Pro účely vyhodnocení je dostupná anotace G (příslušnost ke shluku, resp. identifikace náhodného procesu, který příslušný půlměsíc generoval): $\mathcal{X} \subset \mathbb{R}^2 \rightarrow G = \{G_1, G_2\}$. Měnit lze rozptyl uvnitř shluků a rozdělení příkladů mezi shluky. Za obtížnější lze považovat úlohu s větším šumem a nerovnoměrným počtem příkladů ve shlucích. Vizualizace typického výstupu generátoru je na Obrázku 1.

2 Algoritmus k-středů

Jde o shlukovací úlohu s nekompaktními shluky. Algoritmus k-středů bude velmi pravděpodobně generovat shluky odlišné od zlatého standardu (aplikace Matlab funkce *kmeans* přímo na vstupní data viz Obrázek 2). Protože k-středů je i posledním krokem spektrálního shlukování, je jedním z našich cílů zjistit, jak transformace provedené ve spektrálním shlukování před aplikací k-středů ovlivní jeho výstup.



Obrázek 1: Dva půlměsíce ($n = 200$, $Pr(G) = [0.2, 0.8]$, $\sigma^2 = 0.01$).



Obrázek 2: Dva půlměsíce – shluky dle k-středů.

3 Spektrální shlukování

Spektrální shlukování lze rozdělit do několika funkčních bloků. Následující seznam prochází funkční bloky a definuje, co je dostupné na počátku cvičení a kterou funkcionalitu je naopak třeba doplnit:

- výpočet podobnostní matice $\mathcal{S}$: dostupný ve funkci *CalcSimMatrix.m*, implementace počítá euklidovskou vzdálenost mezi body a následně aplikuje gaussovský kernel, změna není nutná, je ale dobré ověřit důležitost parametru σ (vyšší hodnota znamená zvýšení podobnosti mezi vzdálenějšími body, podobnost je méně lokální),

- vytvoření grafu podobnosti: triviální variantou je ponechat $\mathcal{S}$ beze změny (úplný graf podobnosti), funkce *BuildEpsilonGraph.m* generuje graf podobnosti na základě ϵ-okolí; doplňte funkci *BuildKNNGraph.m*, která bude generovat obě varianty grafu podobnosti založené na k-nejbližších sousedech (symetrická varianta: spoj i a j pokud i je mezi k nejbližšími sousedy j nebo naopak, oboustranná varianta: spoj i a j pokud i je mezi k nejbližšími sousedy j i naopak), k dispozici je funkce *BuildDirectedKNNGraph.m*, která generuje orientovaný graf podobnosti; ke kontrole správnosti lze použít funkci *PlotGraph.m*,

- odvození Laplaceovy matice $\mathcal{L}$ s následnou projekcí do prostoru k jejích nejmenších vlastních vektorů: funkce *CalcLaplacian.m* realizuje tento krok pro nenormalizovaný Laplacián, předpokládá se, že doplníte alespoň jednu variantu výpočtu normalizovaného Laplaciánu (viz [1, 2]),

- aplikace k-středů na výstup předchozí funkce: přímočará aplikace Matlab funkce *kmeans*.

4 Zadání krok za krokem

Následující seznam shrnuje kroky, kterými byste měli během cvičení projít:

1. zvolte vhodné parametry a generujte vstupní data (ke kontrole výstupu použijte funkci *PlotData.m* zobrazující bodový graf (viz obrázky v tomto textu)),

2. proveďte shlukování pomocí k-středů, zhodnoťte jeho úspěšnost vizuálně pomocí funkce *PlotData.m* i číselně funkcí *Purity.m*,

3. ze stávajících funkcí vytvořte základní variantu algoritmu spektrálního shlukování,

4. proveďte shlukování pomocí spektrálního shlukování, zkontrolujte jeho úspěšnost vizuálně pomocí funkce *PlotData.m* i číselně funkcí *Purity.m*),

5. implementujte funkci *BuildKNNGraph.m*, k její kontrole použijte *PlotGraph.m*,

6. implementujte rozšíření funkce *CalcLaplacian.m*, popřípadě zaveďte novou funkci *CalcNormLaplacian.m*,

7. opakujte krok 4 s různými variantami algoritmu spektrálního shlukování (změny výpočtu grafu podobnosti a Laplaciánu), měňte parametry vstupních dat generovaných v kroku 1,

8. shrňte svoje zkušenosti z experimentů (která volba je důležitá a která nemá vliv, jak optimální volba souvisí s obtížností vstupních dat).

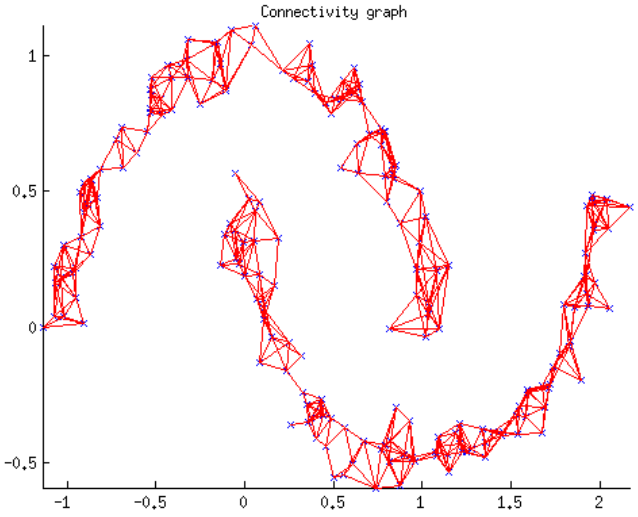
5 Očekávané výstupy

Spektrální shlukování by na rozdíl od k-středů mělo dosáhnout dokonalé shody s anotacemi (zlatým standardem). Pro vstupní data s rovnoměrným rozdělením příkladů do shluků a nízkou úrovní šumu by se to mělo podařit i bez implementace funkčních rozšíření (postačí vhodná volba parametrů σ, popřípadě ϵ). Vodítkem správnosti je vyvoření grafu podobnosti, který má právě dvě souvislé komponenty odpovídající půlměsícům (viz Obrázek 3, výstup funkce *plotGraph.m*). Dvě komponenty odpovídající skutečným anotacím nejsou ale nutnou podmínkou úspěšného řešení, graf může být i souvislý.

Pro vstupní data s nerovnoměrným rozdělením příkladů do shluků není vhodný graf podobnosti založený na ϵ-okolí. Hustota bodů v různých oblastech prostoru se liší, univerzální ϵ není možné nalézt. Naopak jsou vhodné přístupy založené na k-nejbližších sousedech. Symetrická varianta může propojit body z oblastí s různou hustotou bodů. Oboustranná varianta odděluje oblasti s konstantní hustotou bodů.

Pro rozhodnutí o Laplaciánu je důležité kritérium rozdělení stupňů uzlů v grafu podobnosti. U regulárních grafů (všechny vrcholy mají podobný stupeň) by měly být podobné výsledky pro neneormalizovanou i normalizovanou variantu. V opačném případě je vhodné normalizovat. Obecně by jednodušší nenormalizovaná varianta měla být stejně dobrá nebo horší než normalizovaná. Kromě jednoduchosti tedy není důvod pro její použití.

Nastavení parametrů spektrálního shlukování není triviální. Heuristická pravidla pro automatickou volbu lze nalézt v [2].



Obrázek 3: Dva půlměsíce – ideální graf podobnosti se dvěma souvislými komponentami.

Reference

[1] Kléma, Jiří: *Shluková analýza – specializované algoritmy*, přednáška SAD o spektrálním shlukování.

[2] Luxburg, Ulrike: *A tutorial on spectral clustering*, Statistics and Computing, 17/4, pp. 395–416, 2007.