Vytěžování Dat
Cvičení 6 – SOM
Self Organizing Map

Miroslav Čepek, Michael Anděl

14. 10. 2014

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

SOM Toolbox

- V dnešním cvičení vám ukážeme SOM Toolbox.
- Před použitím jej musíte stáhnout a rozbalit.
- SOM Toolbox se nachází na `http://www.cis.hut.fi/somtoolbox/`.

SOM Toolbox (II)

- Až SOM Toolbox stáhnete, rozbalte jej do "nějaké"složky (ideálně tam, kde máte ostatní vaše zdrojové soubory). Doporučuji nechat soubory SOM Toolboxu v jednom podadresáři.
- Tento podadresář musíte přidat do cesty, kde Matlab hledá skripty.
- Pravým tlačítkem klikněte na adresář se SOM Tooleboxem a vyberte "Add to Path" "Selected Folder and Subfolders".

SOM Toolbox – demo

- Společně projdeme demo skripty, které ukazují všechny možnosti SOM Toolboxu.
- Pokud si někdy nebudete vědět rady, projděte si tato dema znovu a většinou v nich najdete, co potřebujete.
- Dema spoustíte příkazy `som_demo1`, `som_demo2`, `som_demo3` a `som_demo4`.

Načtení dat

- Načtěte data pomocí `load iris_dataset`.
- Normalizujte data pomocí `data = som_normalize(irisInputs)`

SOM příklad učení

- Vytvořte náhodně inicializovanou mapu pomocí `som_randinit`.
- Pokud potřebujete vytvořit prázndou mapu, použijte `som_map_struct`.
- `map = som_randinit(data, 'msize', [10 8], 'lattice','hexa')`
- Pro trénování použijte `som_batchtrain(map, data)` (druhá možnost je `som_seqtrain`).
- Variantou je použití funkce `som_make`, která vytvoří SOM síť, inicializuje ji a naučí ji.

PCA vizualizace dat

- Zobrazení dat pomocí PCA
  - Výpočet PCA hodnot: `tmp = pcaproj(data, 2)`
  - Zobrazení `scatter(tmp(:,1), tmp(:,2))`
- Barevné rozlišení tříd:
  - `y = cell2mat(irisTargets)`
  - `scatter(tmp(irisTargets(1,:) == 1,1), tmp(irisTargets(1,:) == 1,2), 'ok')`
  - `hold on`
  - `scatter(tmp(irisTargets(2,:) == 1,1), tmp(irisTargets(2,:) == 1,2), '+r')`
  - `scatter(tmp(irisTargets(3,:) == 1,1), tmp(irisTargets(3,:) == 1,2), '*b')`

SOM – vizualizace

- Zobrazte U-Matici `som_show`.
- `som_show(map, 'umat', 'all')`.
- Jak zobrazit, který neuron je reprezentantem pro která data?
- Musíme použít `som_show_add` a k U-Matici přidat informace o počtu a typu dat.
- Nejprve je potřeba zjistit, který neuron je BMU pro které vstupní instance. K tomu slouží `som_hits`.
- `h1 = som_hits(map, data(irisTargets(1,:) == 1,:)); h2 = som_hits(map, data(irisTargets(2,:) == 1, :));`
- `som_show_add('hit', h1, 'MarkerColor', [1 0 0]); som_show_add('hit', h2, 'MarkerColor', [0 1 0]);`

SOM – vizualizace

- Zobrazte U-Matici `som_show`.
- `som_show(map, 'umat', 'all')`.
- Jak zobrazit, který neuron je reprezentantem pro která data?
- Musíme použít `som_show_add` a k U-Matici přidat informace o počtu a typu dat.
- Nejprve je potřeba zjistit, který neuron je BMU pro které vstupní instance. K tomu slouží `som_hits`.
- `h1 = som_hits(map, data(irisTargets(1,:) == 1,:)); h2 = som_hits(map, data(irisTargets(2,:) == 1, :));`
- `som_show_add('hit', h1, 'MarkerColor', [1 0 0]); som_show_add('hit', h2, 'MarkerColor', [0 1 0]);`

Zadání domácího úkolu

- Je potřeba zjistit, která fóra se odlišují a naopak která jsou těžko rozlišitelná.
- Pomocí SOM shlukování (praktické povídání bude příště) vytvořte shluky dodaných dokumentů.
- Dokumenty obsahují zprávy z několika diskusních fór. Každé fórum má jeden adresář a každá zpráva v něm je jeden soubor.
- Ze stránek předmětu (cvičení) stáhněte tato data.
- Z dokumentů extrahujte důležitá slova a příznakové vektory pomocí rozšíření rapidmineru pro textmining.
- Takto extrahovaná data uložte do CSV souboru.

Zadání domácího úkolu (2)

- Tento CSV soubor načtěte do MATLABu pomocí funkce dlmread (nebo podobné).
- Pomocí SOM Toolboxu shlukněte načtená data a pomocí různých vizualizací zobrazte výsledky shlukování.
- Za pomoci vizualizací učiňte závěry, která fóra jsou si podobná z pohledu použitých slov (strojově obtížně odlišitelná) za pomoci text-miningu a která naopak zle odlišit jednoduše.
- Pomocí funkce kmeans_clusters zjistěte, kolik segmentů v datech je. A pomocí funkce som_cplane si zobrazte, které neurony patří do stejného clusteru. Pomocí som_hits zobrazte, které typy dokumentů jsou ve kterých neuronech a pak učiňte závěr, které dokumenty jsou si podobné.

Nastavení textmeningu

- Tokeny (slova) jsou odděleny znaky, která nejsou písmena.
- Doporučuji, abyste vyfiltrovali příliš krátká slova (řekněme kratší než 5 znaků) a často se vyskytující slova (stopwords) – předložky, spojky, ...
- Pro hledání kořenů slov použijte Porterův algoritmus.
- Volitelně můžete zkusit zkusit zkontruovat n-gramy (tokeny sestávající se z více slov) – doporučuji maximálně 3 slova.
- Také doporučuji odstranit slova, která se vyskytují příliš řídce (příliš málo -krát).

Obsah zprávy

- Zpráva bude obsahovat:
- Popis proudu v Rapidmineru, kterým jste vyextrahovali příznaky z dokumentů a jeho screenshot (alespoň důležité části).
- Popis postupu, jakým jste vytvořili SOM síť a její vizualizace.
- Vytvořené vizualizace a jejich popis.
- Závěr o tom, zda se diskusní fóra podobají nebo ne (případně která).

Užitečné příkazy SOM Toolboxu

- som_demo1, som_demo2, som_demo3, som_demo4
- som_randinit
- som_make
- som_quality
- som_show
- Kompletní dokumentaci všech funkcí naleznete na `http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html`

Užitečné zdroje o Textminingu

- Pokud se chcete podívat, jak se textmining provádí v Rapidmineru, doporučuji následující sérii videí:
  - `http://www.youtube.com/watch?v=hpvda_Rfg3s`
  - `http://www.youtube.com/watch?v=EjD2M4r4mBM`
  - `http://www.youtube.com/watch?v=vhMzUi-FMy0`
  - `http://www.youtube.com/watch?v=ToxzfYECxOU`
  - `http://www.youtube.com/watch?v=BRvjWLwSScQ`
  - `http://www.youtube.com/watch?v=9I0BcMuhPe8`
- Video přednáška o Textminingu
  `http://videolectures.net/ess07_grobelnik_twdmI/`

Užitečné zdroje o Textminingu (2)

- http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf
- http://www.cs.sunysb.edu/~cse634/presentations/TextMining.pdf