Vytěžování dat 6: Self Organizing Map

Miroslav Čepek

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 ionosphere`.
  - This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.
- Normalizujte data pomocí `data = som_normalize(X)`

SOM příklad učení

- Vytvořte náhodně inicializovanou mapu pomocí `som_randinit`.
- Pokud potřebujete vytvořit prázndou mapu, použijte `som_map_struct`.
- `map = som_randinit(X, '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(Y)`
  - `scatter(tmp(y == 'k',1), tmp(y == 'k',2), 'ok')`
  - `hold on`
  - `scatter(tmp(y == 'g',1), tmp(y == 'g',2), '+r')`

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(y == 'g', :)); h2 = som_hits(map, data(y == 'k', :));`
- `som_show_add('hit', h1, 'MarkerColor', [1 0 0]); som_show_add('hit', h2, 'MarkerColor', [0 1 0]);`

Zadání domácího úkolu

- Pomocí SOM 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. (bude náplní dalšího cvičení).
- 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.
- Pro počítání vzdáleností použijte Kosínovou metriku.
- Učiňte závěry, zda se dokumenty v jednotlivých fórech podobají nebo ne.

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 příspěvky v diskusních fórech podobají nebo ne.

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