

Vytěžování dat 6: Self Organizing Map

Michael Anděl



Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

- ▶ 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/`.

- ▶ 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".

- ▶ 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 spustíte příkazy `som_demo1`, `som_demo2`, `som_demo3` a `som_demo4`.

- ▶ Pro načtení dat z .csv souboru použijte funkce `importdata`.
- ▶ Konverze dat do formátu pro SOM toolbox:
`somData = som_data_struct(...)`
 - ▶ **Pozor**, jako parametr `'labels'` (jména vectorů) zadejte **zkrácené** názvy fór, abyste se pak vyznali v jejich vizualizaci.
- ▶ Normalizace data: `somData = som_normalize(somData)`
- ▶ Inicializace mapy:
`map = som_randinit(somData, 'msize', [height width], 'lattice', 'hexa')`
- ▶ Trénování `som_batchtrain(map, data)` nebo `som_seqtrain`).
- ▶ Anotace neuronů podle převládajících dokumentů
`som_autolabel`

- ▶ Zobrazení prázdné mapy:
`figure(1), som_show(map, 'empty', '<name>')`.
- ▶ Zobrazení U-matice, např:
`som_show(map, 'umat', 'all', 'colormap', 1-gray)`.
- ▶ Přidání anotace neuronů do mapy:
`som_show_add('label', map)`.
- ▶ Četnost h výskytů fór při jednotlivých neuronech:
`h = som_hits(map, somData)`
- ▶ Přidání histogramu do mapy:
`som_show_add('hit', h,)`
- ▶ Trakování jednotlivých fór:
 - ▶ `h1 = som_hits(map, data.data(strcmp(data.labels, '<forum1>'))`
 - ▶ `h2 = som_hits(map, data.data(strcmp(data.labels, '<forum2>'))`
 - ▶ `som_show_add('hit', [h1 h2 h3],)`

Zadání domácího úkolu (z minula)

- ▶ Ze stránek předmětu (cvičení) stáhněte data.
- ▶ Dokumenty obsahují zprávy z několika diskusních fór. Každé fórum má jeden adresář a každá zpráva v něm je jeden soubor.
- ▶ Z dokumentů extrahujte důležitá slova, tj. word-vectory
- ▶ Takto extrahovaná data uložte do CSV souboru.

Zadání domácího úkolu (2)

- ▶ Pro import do MATLABu můžete použít funkci `importdata`
 - ▶ **Pozor**, funkce `importdata` vyžaduje aby textový sloupec (v našem případě *labels*) byl v `.csv` souboru jako *první*
- ▶ Na těchto datech naučte SOM
- ▶ Vizualizujte naučenou síť:
 - ▶ Celkovou U-mapu
 - ▶ Řezy U-mapy některými pro vás zajímavými slovy
 - ▶ Anotace neuronů
 - ▶ Histogramy fór
- ▶ Shlukujte jednotlivé neurony algoritmem `kmeans_clusters`
- ▶ Zařazení neuronů do shluků vizualizujte `som_cplane(map.topol, cluster_indices)`
- ▶ Prostřednictvím U-mapy a zařazení neuronů do shluků učiňte závěry, zda se dokumenty v jednotlivých fórech podobají nebo ne.

Odevzdávání výhradně formou uploadovaného protokolu!

Bude hodnocen i způsob prezentace. Zpráva by měla obsahovat:

- ▶ Z minula (zpracování textu, extrakce příznaků):
 1. Popis proudu v Rapidmineru, kterým jste zpracovávali dokumenty, a jeho nejdůležitější parametry. Nezapomeňte uvést rozměry výsledných TFIDF dat.
 2. Popis extrakce příznaků pomocí PCA s následnou klasifikací: tj. matematický popis nebo screenshot z RM. Nezapomeňte uvést, kolik komponent jste použili.
 3. Pokus o interpretaci některé zajímavé komponenty.
 4. Odhady přesnosti klasifikace na surových TFIDF, resp. za použití PCA.
 5. Zdůvodnění pozorovaných výsledků.

Ze dnes:

1. Doporučené vizualizace
2. Závěry, zda se dokumenty v jednotlivých fórech podobají, plus nalezené souvislosti, tj. které se nejvíce podobají, resp. nepodobají kterým.

- ▶ `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>