Inlämningsuppgift Visuell data analys

Mars 2022

Generell information

Filer för inlämning i PingPong:

- Alla Python/R-filer där kod för klustring och visualisering ingår
- Datan som används (csv)
- Textdokument där du förklarar algoritmen/algoritmernas och dina metrics funktion

Deadline 2022-04-26 kl 23:55. Zippad fil namnges med namn, Visuell dataanlys och betygsönske, tex *Eva_Hegnar_Visuell_dataanalys_VG*.

Kom ihåg att kopiera kod rakt av räknas som fusk, men man får hämta inspiration. **Ange källor!** Inlämningen är **individuell**.

Det kommer vara en **muntlig** rättning av inlämningsuppgiften under lektionstid 2022-04-27 för Stockholm och 2022-04-28 för Göteborg. Du kommer boka en tid för one-on-one med Eva där du går igenom inlämningen. Det är **inte** en muntlig redovisning som du ska förbereda, men snarare ett snabbare sätt att rätta inlämningen.

Betygskriterier

G

- Kunna på ett grundläggande sätt förklara vad informationsvisualisering, visual data mining och visual analytics är och hur det används
- Kunna på ett grundläggande sätt förklara Data- och informationsvisualiseringsmetoder
- Kunna på ett grundläggande sätt tillämpa visuell dataanalys
- Kunna på ett grundläggande sätt skapa analysresultat för beslutsfattande
- På ett grundläggande sätt självständigt kunna utföra informationsvisualisering och välja lämpliga sätt att presentera resultatet av en dataanalys för beslutsfattande

VG

- Uppnått kraven för betyget Godkänd
- Kunna på ett fördjupat sätt tillämpa visuell dataanalys
- Kunna på ett självständigt sätt skapa analysresultat för beslutsfattande
- Kunna på ett självständigt sätt utföra informationsvisualisering och välja lämpliga sätt att presentera resultatet av en dataanalys för beslutsfattande

Introduktion

Inlämningsuppgiften rör visualisering och analys av högdimensionell data. Välj klassificeringsdataset från t.ex. UCI, från sklearn eller kom med helt egna förslag för visuell analys. Ni får välja vilket språk ni vill skriva i (R eller Python) och vilket visueliseringsverktyg ni vill använda (matplotlib, seaborn, plotly).

Gör antingen G eller VG:

G

- Tillämpa och förklara en valfri klustringsalgoritm på ett valfritt dataset i R eller Python
- Förklara hur algoritmen arbetar, vilka klustringsmetrics som kan användas för att bestämma kluster och varför dessa metrics fungerar med just den valda algoritmen
- Analysera distributioner för alla variabler innan och efter klustring
- Visualisera även datasetet i 2D med t-SNE eller UMAP
- Utred ifall dimensionsreducering med PCA hjälper klustringen

VG

- Tillämpa 2 valfria klustringsalgoritmer i R eller Python på 2 olika dataset och
- Förklara hur algoritmerna arbetar, vilka klustringsmetrics som kan användas för vardera algoritm och varför dessa metrics fungerar med just dessa algoritmer
- Analysera variablernas distributioner innan och efter klustringen för respektive kluster
- Visualisera även dataseten i 2D med t-SNE eller UMAP
- Utred ifall PCA hjälper klustringen
- Finns det distinkta skillnader mellan dina valda kluster?
- Förklara även i textfilen där du beskriver algoritmerna hur resultaten mellan dina klustringsalgoritmer skiljer sig åt och försök att svara på varför.
- Ett extra krav för VG är även välformulerad kod, med en mainfil som kallar på alla funktioner där alla funktioner har informativa Docstrings.