**Feature selection za unsupervised learning**

**Pojmovi:**

**Normalizacija dataseta:**

[Clearly explained: what, why and how of feature scaling-normalization & standardization | by Juhi Ramzai | Towards Data Science](https://towardsdatascience.com/clearly-explained-what-why-and-how-of-feature-scaling-normalization-standardization-e9207042d971)

Normaliziraj sve što mjeri udaljenost.

**Laplacian score:**

Procjena koliko značajka očuva lokalnost

**SVD vs PCA:**

[dimensionality reduction - Relationship between SVD and PCA. How to use SVD to perform PCA? - Cross Validated (stackexchange.com)](https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca)

**Vrste** feature selection:

Opsežan link: [A review of unsupervised feature selection methods (springer.com)](https://link.springer.com/content/pdf/10.1007/s10462-019-09682-y.pdf)

1. **Univarijantne Filter metode**

- Procjena intrinsične vrijednosti svake značajke pa odabir po pragu ili rangu

- kod supervised learninga uzima u obzir odnos između značajke i labele

* 1. **Metode bazirane na informacijskoj teoriji**

**Def:** Procjena disperzije (odvajanja) podataka kako bi identificirali grupe(clustere) unutar podataka

* + 1. Vjerojatnost – p(i) = 1/max\_bitova^2 ?
       1. vrijednosti niske vjerojatnosti sadrže više informacije, dok česte vrijednosti sadrže malo informacije
    2. Entropija – -
       1. Određuje koliko informacije sadrži značajka
       2. Distribucija vjerojatnosti da određena emisija pripada određenoj klasi
    3. Informacijska dobit –
       1. Veća informacijska dobit znači niže entropije grupa tj. manje iznenađenja
    4. Divergencija
    5. Uzajamna informacija
    6. Korelacija
    7. F-test
    8. SVD – entropija
    9. Laplacian score

– objašnjenje: [Laplacian Score for Feature Selection (neurips.cc)](https://proceedings.neurips.cc/paper/2005/file/b5b03f06271f8917685d14cea7c6c50a-Paper.pdf)

– matlab - [Rank features for unsupervised learning using Laplacian scores - MATLAB fsulaplacian - MathWorks Nordic](https://se.mathworks.com/help/stats/fsulaplacian.html)

* + 1. [Sci-Hub | Feature selection with SVD entropy: Some modification and extension. Information Sciences, 264, 118–134 | 10.1016/j.ins.2013.12.029](https://sci-hub.se/https:/dl.acm.org/doi/abs/10.1016/j.ins.2013.12.029)
    2. ~~Category utility and COBWEB – ne primjenjivo jer nemamo kategoričke podatke već kontinuirane~~
  1. **Metode bazirane na spektralnoj analizi**

**Def:** Slijede ideju modeliranja ili identificiranja

lokalne ili globalne strukturu podataka koristeći svojstveni sustav Laplacijana ili normalizirane Laplacijan matrice izvedene iz matrice sličnosti objekata

1. **Multivarijantne Filter metode**
   1. ocjenjuju relevantnost i redundantnost skupa značajki
   2. uzima u obzir odnos između značajki
2. **Redukcija dimenzionalnosti**
   1. PCA

Matlab - [Analyze Quality of Life in U.S. Cities Using PCA - MATLAB & Simulink - MathWorks Nordic](https://se.mathworks.com/help/stats/quality-of-life-in-u-s-cities.html)

* + 1. **IZBACIVANJEM OULIERA MIENJA SE PCA ANALIZA JER ODREĐENE DIMENZIJE VIŠE NEMAJU NAJVIŠE VARIJANCE KAKO JE SAMO JEDAN OUTLIER ČINIO NJU!**
    2. Potrebno predprocesiranje da je mean značajke blizu nule, varijanca 1
    3. Redukcija dimenzionalnosti na dimenziju bazirane na najvećoj varijanci unutar skupa podataka
    4. Na temelju principalnih komponenti možemo zaključiti koje značajke unose najviše varijance u skup podataka te izabrati ih po određenom pragu
  1. Whitening [Unsupervised Feature Learning and Deep Learning Tutorial (stanford.edu)](http://ufldl.stanford.edu/tutorial/unsupervised/PCAWhitening/)
     1. Transformacija značajki tako da su manje korelirane pomoću početnih koraka PCA
     2. Matrica kovarijance skupa podataka = matrica identiteta
  2. Mjere očuvanja inicijalnih odnosa točaka nakon redukcije dimenzionalnosti

[Sci-Hub | Feature selection with SVD entropy: Some modification and extension. Information Sciences, 264, 118–134 | 10.1016/j.ins.2013.12.029](https://sci-hub.se/https:/dl.acm.org/doi/abs/10.1016/j.ins.2013.12.029)

* + 1. Sammon’s Error
       1. Mjera očuvanja udaljenosti između točaka nakon redukcije dimenzionalnosti
    2. Cluster Preserving Index

1. **Wraper metode** - [dy04a.pdf (jmlr.org)](https://www.jmlr.org/papers/volume5/dy04a/dy04a.pdf)

Pretraživanje prostora od 2^n podskupova značajki i provjera točnosti model

* 1. **Metoda izbora podskupa značajki**
     1. [Feature Subset Selection Process - GeeksforGeeks](https://www.geeksforgeeks.org/feature-subset-selection-process/)7
     2. Unaprijedni odabir - kreće od praznog skupa i dodaje značajke koje najviše utječu na poboljšanje rezultata
     3. Unatražni odabir - kreće od svih značajki i uklanja značajke koje najmanje utječu na poboljšanje rezultata
     4. Stepenast odabir - unaprijedan odabir s unatražnim uklanjanjem
  2. **Clustering hyperparametri**
     1. Broj clustera – k
        1. Metoda izbora broja clustera – optimalni broj clustera je kada su maksimalno udaljeni clusteri jedan od drugog, dok najmanju varijancu

**Koristi više metoda za finalnu odluku.**

* + - * 1. Silhouette vrijednost – najbolje za konveksne i sferične podatke
        2. Akaike information criterion (AIC)
        3. Bayesian information criterion (BIC)
        4. Matalab evalclusters
    1. Kvaliteta clustera – slično kao za broj clustera
  1. **Metoda unsupervised learninga (grupiranja)** – članak [Sci-Hub | Unsupervised learning for classification of acoustic emission events from tensile and bending experiments with open-hole carbon fiber composite samples. Composites Science and Technology, 107, 89–97 | 10.1016/j.compscitech.2014.12.003](https://sci-hub.se/https:/doi.org/10.1016/j.compscitech.2014.12.003)

Matlab - [Choose Cluster Analysis Method - MATLAB & Simulink - MathWorks Nordic](https://se.mathworks.com/help/stats/choose-cluster-analysis-method.html)

* + 1. K-means – za grupiranje konveksnih podataka
       1. Matlab broj cluster - [how to determine the optimum number of cluster use K-Mean Clustering - (mathworks.com)](https://se.mathworks.com/matlabcentral/answers/34623-how-to-determine-the-optimum-number-of-cluster-use-k-mean-clustering)
    2. Gaussian mixture model – za grupiranje Gaussovih distrubucija
       1. Bolje za ne konveksne podatke s različitim skalama
    3. DBSCAN
       1. Matlab - [Density-based algorithm for clustering data - MATLAB - MathWorks Nordic](https://se.mathworks.com/help/radar/ref/clusterdbscan-system-object.html#d123e98029), [DBSCAN - MATLAB & Simulink - MathWorks Nordic](https://se.mathworks.com/help/stats/dbscan-clustering.html)
       2. Odredi epsilon - [How to determine epsilon and MinPts parameters of DBSCAN clustering - Amir Masoud Sefidian](http://www.sefidian.com/2020/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering/)
    4. OPTICS
       1. Matlab kod: [(PDF) Matlab code for OPTICS (researchgate.net)](https://www.researchgate.net/publication/280971149_Matlab_code_for_OPTICS)
       2. Matlab funkcija: [Find cluster hierarchy in data - MATLAB clusterDBSCAN.discoverClusters - MathWorks Nordic](https://se.mathworks.com/help/radar/ref/clusterdbscan.clusterdbscan.discoverclusters.html)
  1. **Metoda semi-supervised learninga**
  2. **Metoda procjene kvalietete clusteringa**
     1. [How to measure clustering performances when there are no ground truth? | by Haitian Wei | Medium](https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c)

1. **Hybrid metode**
   1. Filter metode + Hybrid metode