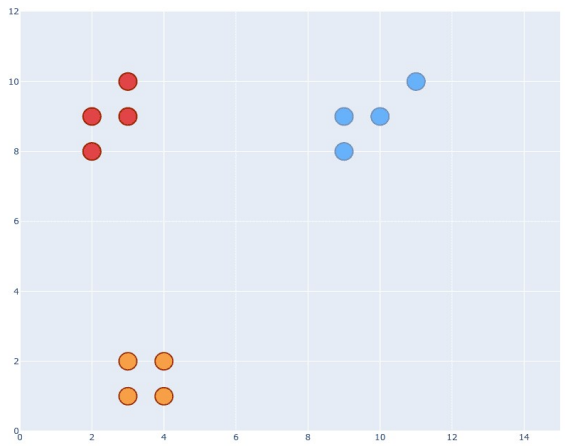
t-SNE algoritam

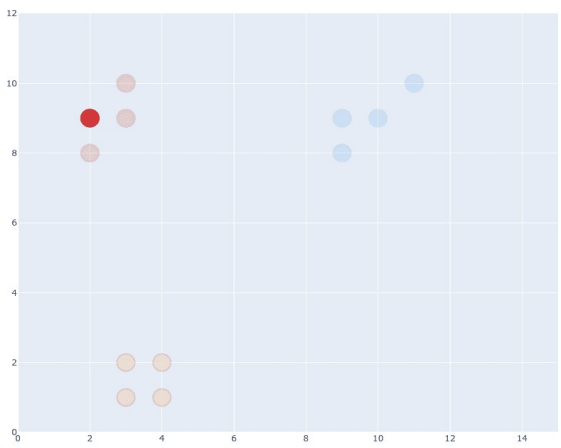
t-distributed stochastic neighbor embedding (t-SNE) je nelinearni algoritam za smanjenje dimenzionalnosti koji se koristi za istraživanje visoko-dimenzionalnih podataka. Ovaj algoritam preslikava podatke u prostor niže dimenzije uz održavanje važnog odnosa između podataka: što su objekti bliži u originalnom prostoru, to je manja udaljenost između njih u redukovanom prostoru. Jednostavnije rečeno, t-sne daje ideju o tome kako su podaci raspoređeni u višedimezionalnom prostoru. Razvili su ga Laurens van der Maaten i Geoffrey Hinton 2008 godine.

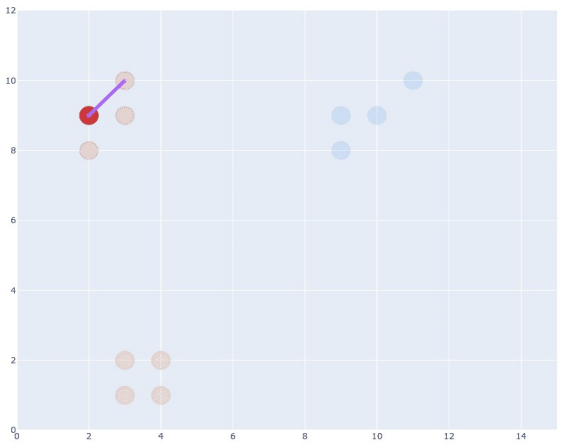
Kako radi t-SNE?

Recimo da imamo skup podataka sa 3 različite klase koje lako možemo razlikovati. Prvi deo algoritma je računanje raspodele verovatnoće koja predstavlja sličnost među susedima. Ta sličnost je zapravo uslovna verovatnoća između neke dve tačke.

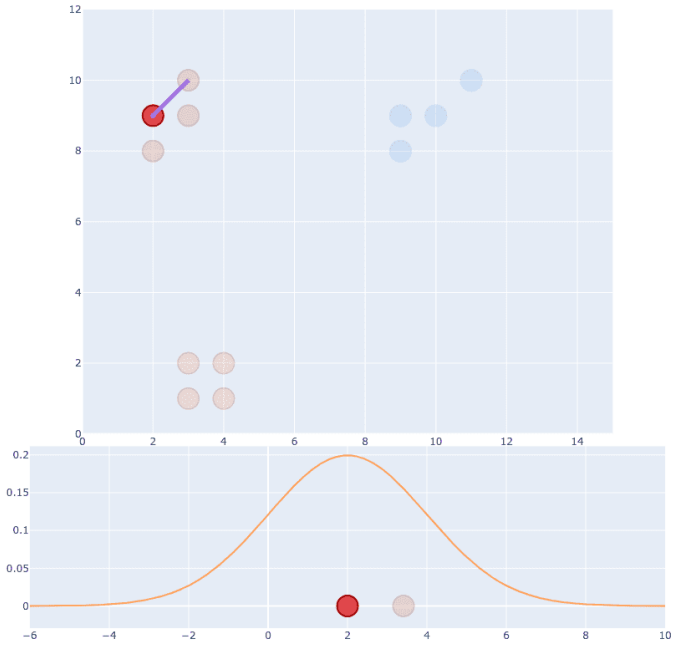


Odaberimo jednu tačku x_k iz skupa podataka. Moramo izabrati i drugu tačku x_j i izračunati euklidsko rastojanje između njih.

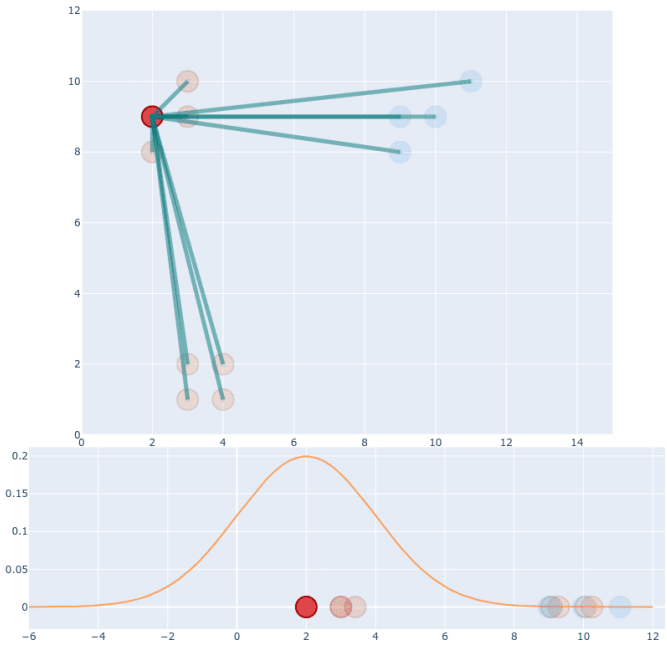




Dalje generišemo Gausovu raspodelu sa centrom u x_k i izračunato euklidsko rastojanje postavljamo na x-osu.



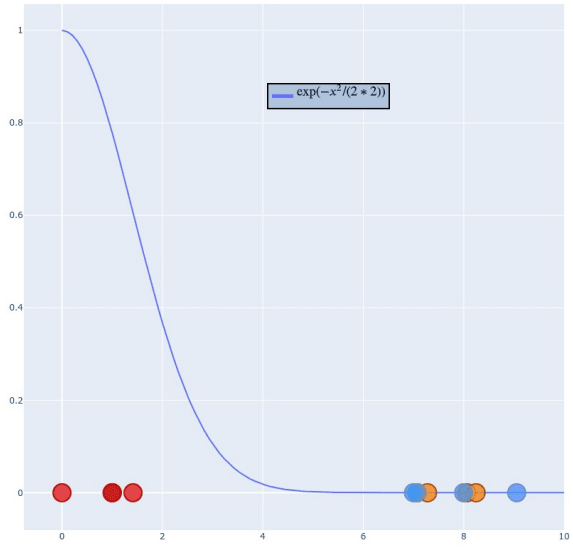
Postupak ponovimo za svaku tačku koja se nalazi u skupu.

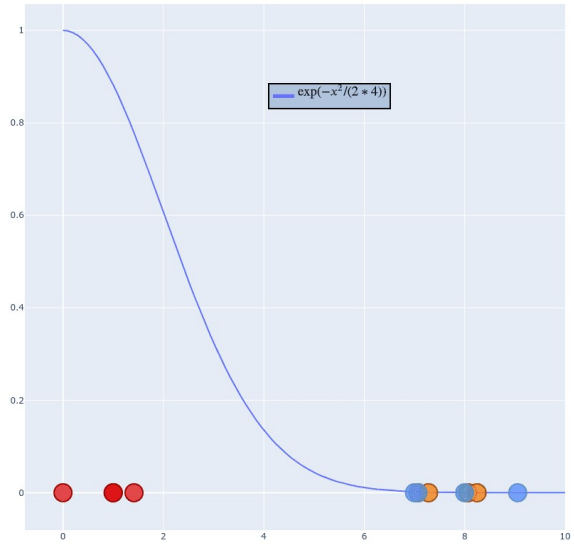


Varijansa zavisi od parametra **perplexity** koji se navodi kao argument funkcije. U osnovi, što je veći perplexity**,** veća je i varijansa. Varijansa je različita za svaku tačku i izabrana je tako da tačke u gustim oblastima dobiju manju varijansu od tačaka u retkim oblastima.

Formula:

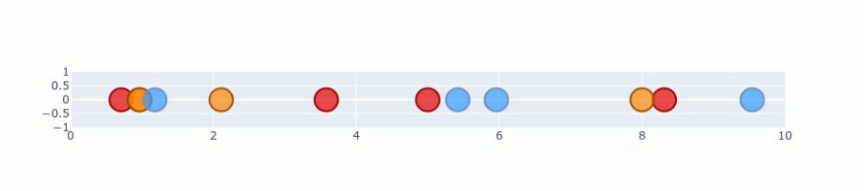
$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$





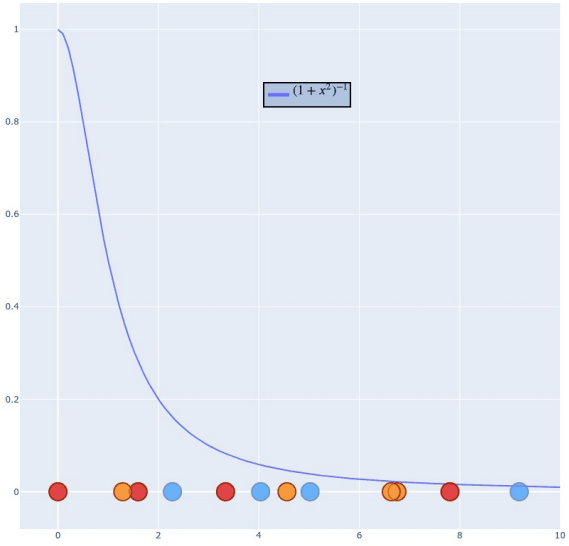
Na ovaj način smo dobili matricu skora sličnosti (matrix of score similarity).

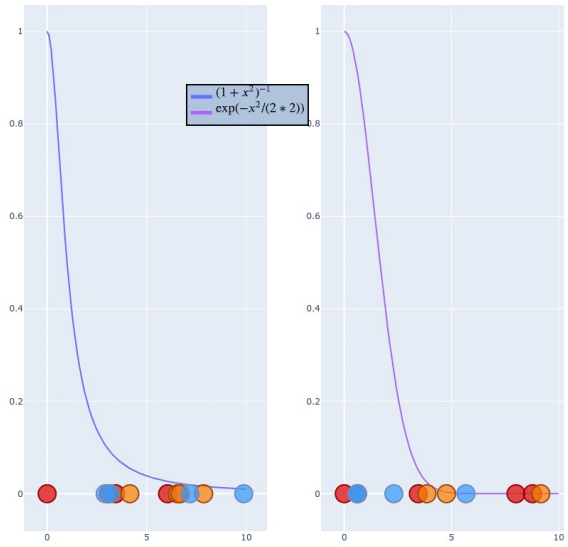
Sledeći deo je **formiranje nisko-dimenzionalnog prostora** sa istim brojem tačaka kao i u originalnom prostoru. Tačke treba nasumično rasporediti u novom prostoru.



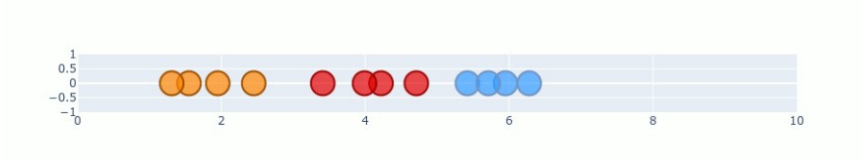
Potrebno je i ovde izračunati matricu skora sličnosti. Isto kao i malopre, biramo proizvoljnu tačku, računamo euklidsko rastojanje i uslovne verovatnoće koje predstavljaju sličnost između suseda. Samo ćemo sada koristiti Studentovu t-raspodelu.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$





Još je potrebno “grupisati” tačke u nisko-dimenzionom prostoru:

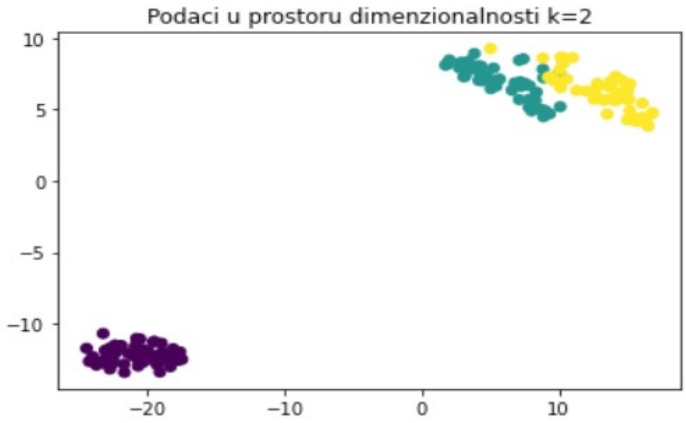


Za ovaj korak koristimo Kullback-Leibler divergenciju koju možemo da shvatimo koliko je jako privlačenje i odbijanje između tačaka.

Primer

Za primer je korišćen dataset Iris iz sklearn biblioteke.
Dataset Iris je sačinjen od 50 uzoraka iz svake od tri vrste perunike (Iris setosa, Iris virginica i Iris versicolor). Za svaki uzorak su izmereni čašićni i krunični listići.

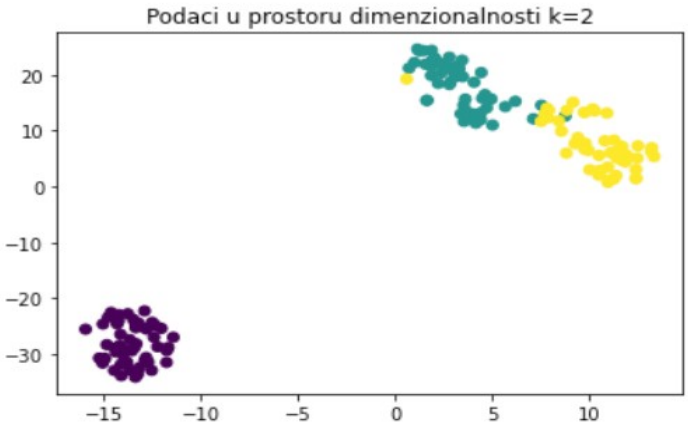
Primenjena je funkcija TSNE iz sklearn.manifold.
Od parametara je uzet perplexity=30.0



Možemo videti da je t-sne dobro formirao različite grupe od naših podataka prema različitim vrstama perunike.

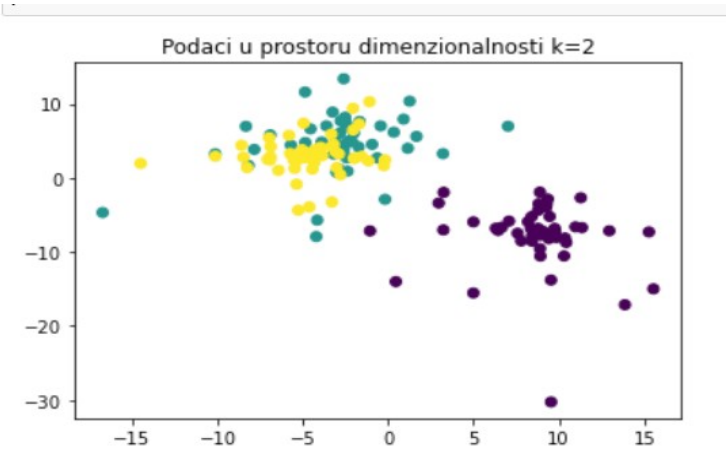
- *Promena argumenata:*

TSNE(n_components=2, perplexity=20.0)



TSNE(n_components=2, perplexity=30.0, n_iter=250)

n_iter: (default=1000) maksimalan broj iteracija za optimizaciju. minimalna vrednost: 250



Literatura

[1] https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/
[2] https://www.oreilly.com/content/an-illustrated-introduction-to-the-t-sne-algorithm/
[3] https://www.youtube.com/watch?v=NEaUSP4YerM
[4] https://www.machinelearningmastery.ru/an-introduction-to-t-sne-with-python-example-5a3a293108d1/
[5] https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a
[6] https://en.wikipedia.org/wiki/Iris_flower_data_set

Lidija Čikarić, 1080/2020