Data Mining Lab3 Report

Prove formally correctness Student's t-distribution as a kernel to measure the similarity between embedded point and centroid

All centroid points are independent and have random distribution. So we can use Student's t-distribution as probabilistic measure between similarity of embedded point and centroid. (Soft assignment)

$$q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1 + \|z_i - \mu_{j'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}},$$

For **α = 1**

$$q_{ij} = \frac{\left(1 + \|z_i - \mu_j\|^2\right)^{-1}}{\sum_{j'}\left(1 + \|z_i - \mu_{j'}\|^2\right)^{-1}}$$

So less distance between point and centroid $\left\|z_i - \mu_j\right\|^2$ than bigger *q.*

Prove formally correctness of p_{ij} estimation

$$p_{ij} = \frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}}$$

Another distribution which improves Q distribution. It shows how strong probability of point to centroid j in relation to other points q_{ij}

Explain how have you implemented clusterization part of DNN (SGD optimisation)

We measure KL distance between Q and P distribution as loss function of DNN and optimize it using SGD optimization

Explain how have you implemented encoder pre train part.

For create encoder we need to create autoencoder with two parts: encoder and decoder. Encoder reduces dimensions and decoder restores them. The better the data restored - the better the autoencoder learned. We trained a model with many different hyperparameters and looking for the best k-means accuracy. Out the best variant you can see in the code.

Prove formally O(nk) complexity of DEC method

At each iteration we need to calculate Q matrix and P matrix which is O(nk), because for each sample of n we have k clusters centers. Full computational complexity is bigger because of model parameters which are not small.