# Cluster Sampling

October 16, 2019

After clustering embeddings obtained from autoencoder by using `K-means` clustering algorithm, we have 10 clusters. We need to map cluster labels to actual classes, such as `cat` or `truck`. Ideally, the clusters are homogeneous, and it would be enough to do the mapping based on a single sample from each cluster. However, in reality the clusters are not homogeneous.

In order to map the labels, we are going to take a small sample of size $n$ from each cluster. We need to find $n$ such that the probability of majority class in the sample being the same as the majority class in the entire cluster is greater or equal than a certain $\alpha$.

We can do that by using the following formula:

$$n = \min \left\{ n \in \mathbb{N} : \sum_{k_1=\lceil \frac{n+9}{10} \rceil}^{n} \sum_{k_2=\lfloor \frac{n-k_1}{9} \rfloor}^{\min(k_1-1, n-k_1)} \sum_{k_3=\lfloor \frac{n-k_1-k_2}{8} \rfloor}^{\min(k_1-1, n-k_1-k_2)} \dots \sum_{k_{10}=n-\sum\limits_{i=1}^{9} k_i}^{\min(k_1-1, n-\sum\limits_{i=1}^{9} k_i)} \left( \prod_{i=1}^{10} \binom{n - \sum\limits_{j=1}^{i-1} k_j}{k_i} \mathbf{P_i^{k_i}} \right) \geq \alpha \right\}$$

For now, we assume that we know the distribution of classes in the cluster, or, in other words, probabilities of selecting each class: $\mathbf{P_1}, \mathbf{P_2}, ...\mathbf{P_{10}}$.