CSE 280A
Assignment 3
List 12
Andrey Bakadze

(4) For this assignment I chose

-I 3 200 200 200.

I'm actually pretty sure that msms gives individuals sorted with respect to population. At the end I explained how I checked it.

For each sample I trained a linear SVM with multiclass rule for optimization. Train/test split is 0.8/0.2.
To test quality I used „accuracy" metric: mean #errors.
It is usually considered a poor metric as is (1) very radical for multiclass classification, (2) does not respect unequal sizes of clusters, (3) cannot be modified easily to support probabilistic classifiers. However, in this assignments (1) the classification is easy, so metric works, (2) populations are equal in size; (3) SVM is not a probabilistic estimator.

For random guess accuracy = $2/3$ (for 3 clusters).
I checked hypothesis that my accuracies are generally lower than $2/3$. // $H_0: E\xi = a$ $t = \sqrt{n}\frac{\bar{x}-a}{s} \xrightarrow[n\to\infty]{} N(0;1)$ //

T test statistic was ≈ -141.4, pvalue $< 10^{-100}$

Why I believe that individuals are sorted by population?
It is highly unlikely that I would get such a good pvalue if the labeling was wrong. To clear all doubts I repeated the analysis with presorting each sample $\Rightarrow$ the labeling is really random.

T test statistic ≈ -0.5339, pvalue ≈ 0.595.

Improvements. In this scenario the classification is very easy and probably shouldn't be improved, but for harder tasks:
(1) Metric: accuracy → Multiclass AUCROC
(2) Classifier: More advanced classifiers. F.e., more flexible kernels
(3) Transformation of data. F.e. Binary → Real valued; Feature-selection.