Exercise 7

Attached is data on gene expressions (**expressions.txt**) from five different tumor types (specified in **tcga-pancan-hiseqlabels.csv**). Simplified, the data represents how strong each gene (column) was in each sample (row). Idea is that different tumors have some overall differences in gene expressions, making them distinguishable from each other.

This dataset contains total of 795 samples, each with 2257 features. Data is preprocessed by removing features with small amount of variation as well as some invalid rows[1].

Your tasks are as follows:

1. Build a system that can say from which tumor the given gene expression data came from (i.e. a classifier to classify between tumor types). Can you reliably separate classes from each other?
2. Study which of the gene expressions are important for the classification (which genes separate the tumors the most). Provide arguments/reasoning for your conclusions. Can you find simple rules like "If gene X expression is higher than Z, then sample is from tumor Y"? You can identify different genes with the column index in the given dataset.

You may use implementations of existing classification/prediction algorithms in this task. You are free to use any technique presented during lecture or found elsewhere (cite sources and do not copy/paste existing code). **Report all steps, techniques and methods used in your submission.**

Tips and hints:

- First column in **expressions.txt** identifies samples from each other, same with **tcga-pancan-hiseqlabels.csv**. Note that these are not aligned (i.e. Row X in first file may not be same gene as in row X in second file).
- To get general sense of "classifiability" of data, you can start by clustering the dataset with i.e. k-means.
- To visualize the very high dimensional data, you can try out *dimensionality reduction*. Core idea: Turn high-dimensional data to low-dimensional (usually two-dimensional for plotting), while also trying to preserve the structure of the data in lower dimension.
- You can turn two-class classifier (i.e. perceptron) into multi-class classifier by training one classifier for each class (one vs. rest) or training one classifier per each class pair, and taking the majority vote of classifiers (one vs. one).
- Gene expressions may be highly correlated with each other.
- You can study relation between gene expressions and tumor classes by i.e. visualizing/studying distribution of features or by studying the weights/parameters of

[1] Original data: http://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq. A shoutout to Wilhelmiina Hämäläinen for offering the project idea and doing the preprocessing!

learned models. You can also try brute-forcing the problem: By definition, if feature is not important for classification, it should not affect the classification result.