Homework 4

You can submit in groups of 2.
Due 4/4, 1pm.

All assignments need to be submitted via github classroom:
https://classroom.github.com/g/likkFf_q
and via gradescope.

In this homework, we approach the problem of identifying a small subset of a dataset using unsupervised and supervised methods.

The dataset we are looking at is the Annthyroid dataset
http://odds.cs.stonybrook.edu/annthyroid-dataset/

In a real world application, we often don't have labels, and clustering and outlier detection are usually applied in settings that don't have labels. In this homework, you should work without the ground truth labels that we have as much as possible. Often inspecting and visualizing the data is the only way to understand the result of clustering and outlier detection.
However, since we do have ground-truth we can do a post-hoc analysis and determine how well we actually did.
For tasks 1-3 you don't need to split the data or use cross-validation. For task 4, you need to use the standard split methods for supervised learning.

Task 1 Visualization (20Pts)

For both tasks, you should look at the plots before making use of the ground truth labels.
For your submission, use the ground truth labels to color the points in all scatter plots.

1.1 Visualize the univariate distributions of all features, jointly and per class.
Visualize the data using PCA (first two principal components).
Plot the explained variance ratio in PCA. What would be a good threshold for the number of principal components if you wanted to reduce the dimensionality of the data to compress it?

1.2 Visualize the data using t-SNE. See if tuning the perplexity parameter helps obtaining a better visualization.

Task 2 Clustering (35Pts)

2.1: Use KMeans, Agglomerative Clustering and DBSCAN to cluster the data. For each algorithm, try to manually tune the parameters for a reasonable outcome and document how you tuned the parameters. In particular pay attention to the sizes of the clusters created. Create a dendrogram for agglomerative clustering (the truncate_mode='level' might be useful). Manually inspect the outcomes as good as you can and identify if any of the resulting clusters are semantically meaningful (as far as you can tell).

2.2: Use the known ground truth labels of the outlier vs inlier class to evaluate your clustering approaches using NMI and ARI scores. How well did they do? Can you adjust parameters so they can detect the outliers better?

Task 2 Outlier Detection (35Pts)

3.1 Assume that you know the proportion of outliers. Use EllipticEnvelope, OneClassSVM and IsolationForest to detect outliers. Without using the ground-truth, can you tell which one gave the best results? Why?

3.2 Use the ground-truth to evaluate the different outlier detection approaches using AUC and average precision. How can you compare their quality to the clustering approaches from Task 1?

Task 3 Imbalanced Classification (10Pts)

Threat the problem as an imbalanced classification problem using LogisticRegression and RandomForestClassifier. Compare your results with the outlier detection in terms of AUC and average precision. Tune C and a regularization mechanism for the random forest. Does changing the class-weight to “balanced” help?