Course project data questions

Mice protein expression

Classes:

1. Mementine / Saline
2. Control / Down syndrome
3. Behavior CS Learning / SC No Learning
4. Learning outcome No learning / Normal /Failed /Rescued
5. 8 groups of mice

Strategy :

Exploratory: apply MDS to see the grouping

To select a subset of features that are different between two classes- features are proteins that have different protein expression in the two classes. WORK WITH THESE FEATURES.

Dimensionality reduction – Principal Components as new features.

Feature selection: subset selection using forward / backward from the package leaps

WORK WITH ALL DATA.

We want **to apply supervised learning** to create a model/ classifier that has good predictive accuracy. How? Having either selected or extracted features or all 77 original expressions choose one or two or more classification rules (kNN, SVM, Naïve Bayes, Decision Tree, Bagging. Boosting, Random Forest, Logistic Regression, LDA) and do cross validation for what ? to estimate parameters and accuracy and compare it) . Practical value of this classification rule.

Regression – to predict the expression of the protein based on the values of a subset of proteins.

We want to apply unsupervised methods to reduce dimensions, to cluster to see the structure in the data – subset of features or all features - that you analyze. Methods are – Kmeans, PCA, MDS, nMDS, MeanShift, DBSCAN, Hierarchical clustering – different linkages and similarity measures can be considered different methods. Practical value of the discovered structure- how can we use it, for what?

Choose 2 supervised and 2 unsupervised methods or choose to repeat the SOM method in the paper.

Questions that arise while reading the paper:

Laima questions: What actually is this data what is unit of data, was the normalization applied, what are concentrations, what is outlier ;

Ezekiel: Do we use all 77 expressions? It is up to the student.

Erinija: CS and SC is confusing.