Description:

As the available data is basically limited to 6 sample populations of varying sizes, I decide to opt for a linear regression model (optionally with L1 or L2 regularization). Since the amount of cells per population as well as their activation percentages are known, I augment the data by creating all possible non-empty combinations of populations, calculating their activations as a weighted average of the population sizes. By doing this, the amount of samples increases from 6 to 2^6 – 1. For each of these samples, the mean, standard deviation, skew and kurtosis of all 12 measurements is calculated. In order to check which features are useful for predicting the activation of the population as well as which form of regularization and regularization parameter yield the best performing model, I perform a variation of 6-fold cross validation. Here, given a held-out population, all combinations of non-held-out populations are used as a train set (size 2^5 – 1) and the resulting model is validated only on the held-out population (size 1), as validating on combinations with populations used in training would leak train data into the validation set. To find the best model, different regularization parameters for L1 and L2 regularization as well as different combinations of features are used. The model with the lowest validation MAE is unregularized linear regression using only the means of the 12 measurements as features. This model has an average MAE of 6.39 when predicting the activation percentage of the left-out population during cross-validation. Moreover, binning the predicted activation percentages of the linear model into 4 classes (0-13%, 13-35%, 35-56% and 56-100%) yields a 100 accuracy when tested on all 2^6 – 1 combinations of cell populations. To further improve the interpretability and generalizability of the model, one could consider which of 12 features are actually pertinent for the model performance, further pruning any features that don’t improve validation performance. Moreover, a different type of model such as a (boosted) decision tree could be employed.