Assignment 3 Report

Dimension Reduction

I chose to use PCA and NMF to reduce the dimensions of the data to 50 components each. NMF appeared to find striking differences between tissues in the lower components but less so in the higher ones. I assume it is just finding the main signals but this could be overshadowing smaller tissue populations by treating them as noise. PCA also divides tissues well. It looks as though each component shows increasing granularity in differences between tissues whereas NMF is pretty homogenous after the first few.

Clustering

I used Gaussian Mixture Modeling which I found performed best in the first assignment. I clustered for several mixture component sizes using both PCA and NMF. Looking at their BICs, NMF was better at finding latent groupings, but as the mixture number increases to >15 it actually becomes worse than PCA.

I evaluated clustering using confusion matrices and these showed how GMM was better at assigning observations to unique tissues using PCA compared to NMF. Using NMF, the model was not able to distinguish between correlated tissue types like heart and muscle, and assigned several observations of the same tissue to different clusters.

The PCA clustering however had much more distinct groups with clearer clusters.

Ensemble Learning

Overfitting Learner

I used an ensemble of decision trees as a overfitting model which I decided to implement manually for some reason. The accuracy on the test set was much lower than on the training set meaning this model had a high variance and therefore overfit.

Underfitting Learner

For the underfitting learner I chose to use the `AdaBoostClassifier` using decision trees again. I set each learner to be weak by specifiying a max tree-depth of 1. This resulted in a underfitting model because training accuracy was low, but consistent across training and testing.

Age Prediction

To predict age from gene-expression embeddings I chose to use a Random Forest again, but this time as a regressor. I did a quick parameter search and found that 100 estimators with a max_depth of 10 would perform best.

I did followed the same workflow as with the `RandomForestRegressor` using a `GradientBoostingRegressor` with 300 estimators operating to a max_depth of 10.

Comparing between these two models, the `RandomForestRegressor` underfit much less and was slightly more accurate with the hold-out. The `GradientBoostingRegressor` performed near perfectly in training but way worse in testing, indicating that this model architecture really overfit.

I was curious how a simple linear regressor would compare to these two ensemble methods and found that a `Ridge` model performed nearly as well as the `RandomForestRegressor` with much lower variance (higher bias).

For continuous problems with most likely linear relationships I would choose a `Ridge` over any ensemble as it performs pretty well and is much more interpretable.