Jessica Massey

A1

Part (a):

True. If we use k principle components and $k = rank(k)$, the projection will capture all the variance in the data and yield zero reconstruction error.

Part (b):

False. According to singular value decomposition, it would actually be the columns of X^T, not the rows. So in this case, the columns of V^T are the eigenvalues because they are equivalent.

Part (c):

False. Sometimes it is not enough to minimize the k-means objective. Only minimizing the k-means objective can lead to overfitting and might not guarantee that the clusters will be interpretable or useful.

Part (d):

False. The singular values on the diagonal entries are unique, but the singular vectors are not unique when singular values are repeated.

Part (e):

False. The rank of a matrix is the max number of linearly independent rows or columns in the matrix, not the number of unique nonzero eigenvalues.

A2

Part (a):

Some pre-processing steps I would take for this scenario would be to handle any missing values in the given dataset. Then I would convert any categorical variables such as demographic and geographical information into numerical format (through one-hot encoding or other methods). After all the data is converted, we can standardize the data. Then we can use PCA to extract the most important features. Finally, we divide the dataset into training, validation, and testing sets.

The specific machine learning pipeline I would use is a neural network with hidden layers with ReLU or sigmoid activation and the last layer with softmax, because we're trying to represent the probability of the person having the disease or not. To train the neural network, we can use cross entropy loss and a SGD optimizer. Some techniques I would employ include cross-validation to ensure that the model is generalizing well and hyperparameter tuning with something like k-fold cross validation to find the optimal hyperparameters.

My setup acknowledges the constraints of the dataset by selecting the most relevant features in order to provide the most correlated result and uses neural networks to achieve a model complex enough to calculate these probabilities. In addition, in order to maintain accuracy, the model is thoroughly assessed on the testing set.

Part (b):

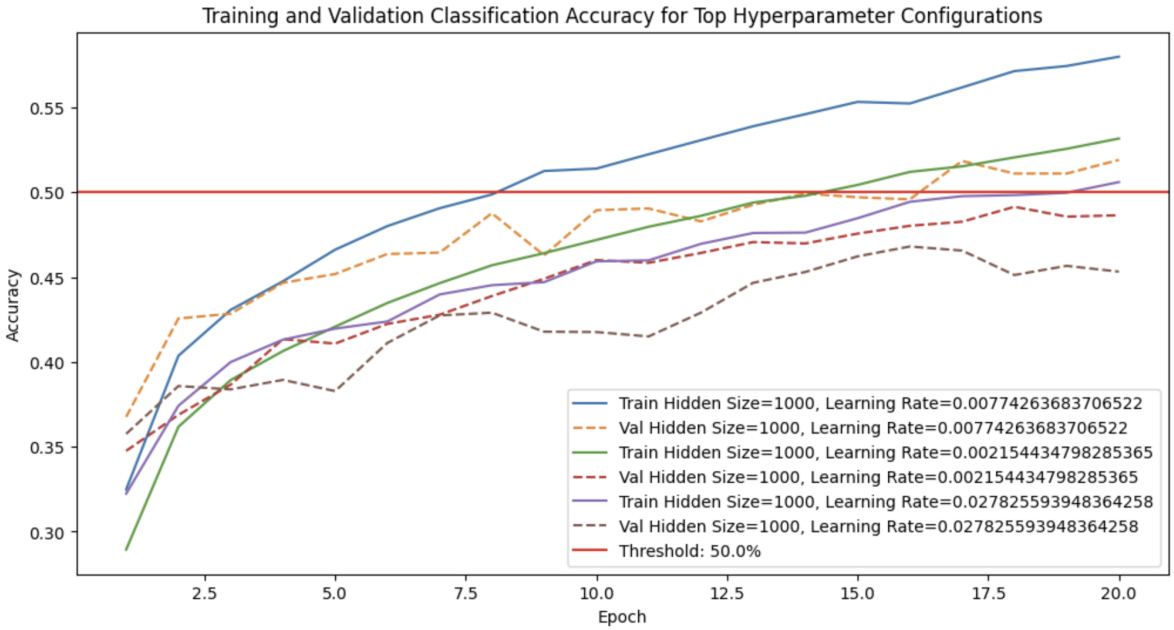
Some potential shortcomings of my training process is there may be bias in the dataset. If certain demographics are over/under-represented, the model will not accurately determine whether a person has a specific disease or not. To address these shortcomings, I would collect more representative and unbiased data.

Part (c):

Some implications of ignoring issues of crimes being reported at different rates is inequitable treatment of certain populations based on inaccurate crime data and therefore, the reduced effectiveness of crime prevention efforts.

A3

Part (a):



.

Hyperparameter values:

- Learning rate: $\boxed{[10^{-6}, 10^{-1}]}$
- Hidden size: $\boxed{[400, 500, 700, 800, 900, 1000]}$

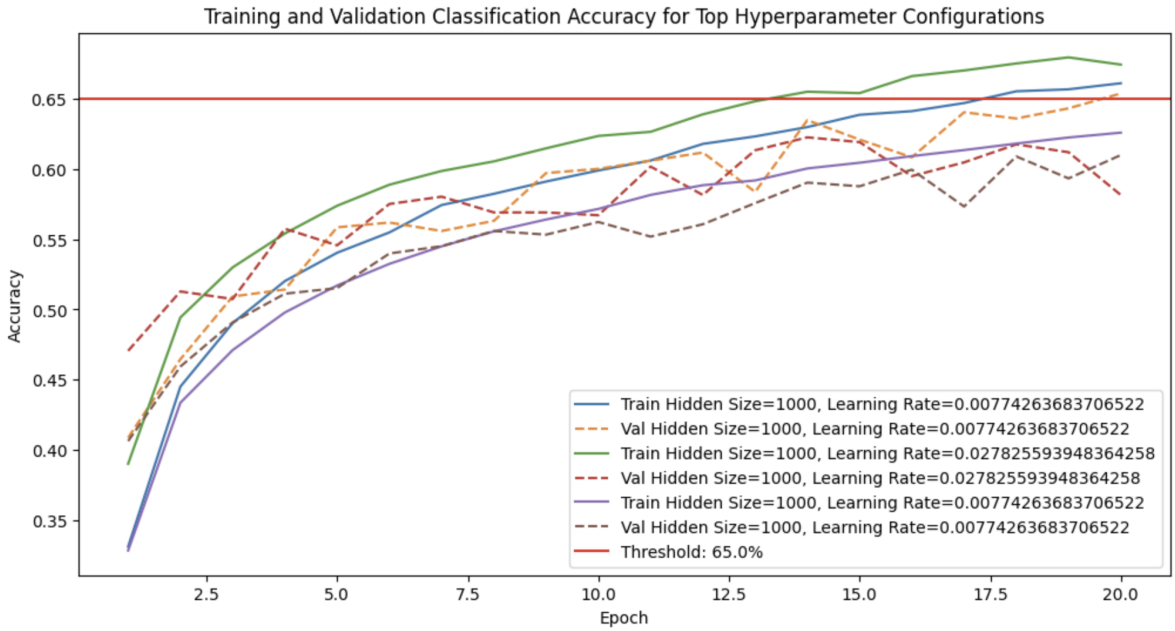
Search method: $\boxed{\text{Grid}}$

Values of best performing hyperparameter values:

- Best hidden size: $\boxed{1000}$, Best learning rate: $\boxed{0.00774263683706522}$
- Best hidden size: $\boxed{1000}$, Best learning rate: $\boxed{0.002154434798285365}$
- Best hidden size: $\boxed{1000}$, Best learning rate: $\boxed{0.027825593948364258}$

Accuracy of best model on test data: $\boxed{0.5204}$

Part (b):



.

Hyperparameter values:

- Learning rate: $\boxed{[10^{-6}, 10^{-1}]}$
- Number of filters: $\boxed{[100, 200, 1000]}$
- Filter size: $\boxed{[5]}$
- Pooling size: $\boxed{[10, 12, 14]}$

Search method: $\boxed{\text{Grid}}$

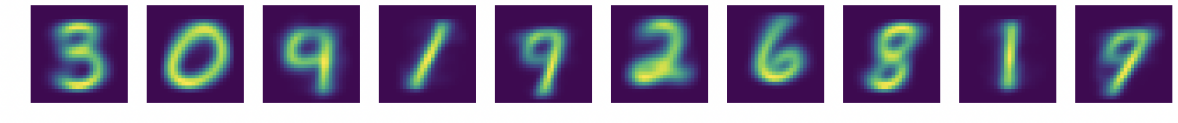
Values of best performing hyperparameter values:

- Best number of filters: $\boxed{1000}$, Best filter size: $\boxed{5}$, Best pooling size: $\boxed{14}$, Best learning rate: $\boxed{0.00774263683706522}$
- Best number of filters: $\boxed{1000}$, Best filter size: $\boxed{5}$, Best pooling size: $\boxed{10}$, Best learning rate: $\boxed{0.027825593948364258}$
- Best number of filters: $\boxed{1000}$, Best filter size: $\boxed{5}$, Best pooling size: $\boxed{10}$, Best learning rate: $\boxed{0.00774263683706522}$

Accuracy of best model on test data: $\boxed{0.6415}$

A4

Part (b):



A5

About how many hours did you spend on this homework? There is no right or wrong answer :)

Lost count!!!!!!!!