CS148 Homework 3

Prithvi Kannan
UID: 405110096

Collaborators: Vanessa Wang, Samuel Alsup

Problem 1:

a. E = ½*log(1/2) + ½*log(1/2) = 1
   E(Color) = ½ * (2/5*log(2/5) + 3/5*log(3/5)) + ½ * (2/5*log(2/5) + 3/5*log(3/5))
   Gain color: 0.029
   E(Type) = 3/5 * (2/3*log(2/3) + 1/3*log(1/3)) + 2/5 * (1/4*log(1/4) + 3/4*log(3/4))
   Gain type: 0.124
   E(Origin) = ½ * (2/5*log(2/5) + 3/5*log(3/5)) + ½ * (2/5*log(2/5) + 3/5*log(3/5))
   Gain origin: 0.029
   Select "type" to split on
b. First split on Type into Sports and SUV categories. Then consider splitting on either Color or Origin to further reduce the entropy
c. A possible stopping criterion for this process can be when there are no more input features to split on since there are only 3 options.
d. No, Decision trees do not require feature scaling to be performed as they are not sensitive to the variance in the data
e. Yes, Decision trees are robust to algorithms. Tree algorithms split the data points on the basis of values in the same category and so value of outlier won't affect that much to the split.

Problem 2:

a. False, boosting uses various weak learners in combination to create a stronger one.
b. TODO
c. Some strengths of Bagging is that the model is robust against outliers, less likely to overfit, and does not require advanced parameter tuning. However, Bagging can be computationally expensive. Boosting puts more weight on weak classifiers from the previous phase so it is more vulnerable to overfitting and outliers. Both of these ensemble methods are less interpretable than the original models they are built upon.
d. ReLU helps address vanishing gradient by giving an output of 1 for all input values greater than 0, no matter how large the input value is. This is different from a sigmoid function which has gradient of 0 at the extremes. Large values of activation function will enable gradients to pass through during backprop.

Problem 3:

1. Using a larger dataset would help to mitigate overfitting since the model would be able to better understand the true population since it would have seen more examples.
2. Allowing more training iterations would not help with overfitting since the model would further specialize on the training dataset.

3. No, increasing the number of features in the model would make the model more complex which would make it more likely to overfit.
4. Randomly zeroing out half the nodes in the network would help with overfitting. This technique is known as dropout and is equivalent to training different neural networks and averaging their effects.
5. Using a GPU would not help mitigate overfitting. Using a GPU would only help speed up training, but the actual process would be the same.
6. Model initialization may help prevent overfitting. We may train a model multiple times before selecting a desired model, and each model may have different initializations. Certain runs may result in convergence at local minima, so considering various initializations can help the model achieve the best outcomes.

Problem 4:

a. Dimensionality reduction can help with overfitting since the model has fewer degrees of freedom. Dimensionality reduction can also help to reduce data storage space, computation time, and remove redundant features. Lastly, dimensionality reduction can be useful for visualizations of high dimensionality data. After dimensionality reduction, the input features are less interpretable because the new axis may not correspond to a real world measurement.
b. PCA would work well on data with a linear structure. In this case, we would be able to reduce the dimensionality of the data along the relevant axis and eliminate the other features with low variation.

   PCA would struggle on hyperbolic data since it would be hard to distil the data into only two features based on the structure of the data.

   PCA works best on data that has already been scaled. If the data is on different magnitudes then it would be more difficult to identify which features truly have more impact than others.

   PCA should not be used on data where features are independent. The idea behind PCA is to generate a new feature that is representative of the existing input features, and if features are independent this would not be possible.

Problem 5:

a. Output layers can use sigmoid (binary), softmax (multiclass), or linear (continuous). Hidden layers use ReLU, Leaky ReLU, Swish to prevent vanishing gradients in backprop. Input layers do not use activation functions.
b. TODO