Oliver De Sa
1005457943
LMP1210 HW1 Answer Submission

1 a) Overfitting occurs when a model becomes too well defined for a specific set of training data. As a result, the model becomes very specific to that set of data and becomes less generalizable to novel data. One way to avoid over-fitting in a decision tree model is to reduce the number of decision nodes. You can also reduce overfitting in a linear model by regularizing the weight parameter to ensure weights are kept to a limit while still fitting the model.

b) As the magnitude of k increases, the number of nearest neighbors considered when assigning a label to a new point also increases. As a result of this, the label assigned to that point is decided by using more nearest neighbors with a large k, meaning you get a label that is more general, but is not very specific to outliers or unique data points. On the other hand, a small k takes less nearest neighbors into account so you get a more specific label, however it is more sensitive to outliers. As a rule of thumb you can use the formula: k = 2/n^(2+d), Where d is the number of dimensions of the data and n is the number of data points used. In practice you should tune your k value by running it against a validation set of the data at different ks and looking for the best one.

c) **Regression category**: standard curve interpolation. I frequently run cell-based assays that secrete a protein that turns blue and can be read in a spectrophotometer. The Optical Density displayed in the output is linearly dependent on the concentration of ligand that the cells are treated with. In this case I can generate a standard curve where a known concentration of ligand is treated to the cells at increasing dilutions until reaching zero. I can then build a linear model based on this relationship of ligand concentration and optical density and use it to predict the concentration of ligands in samples which the where concentration is unknown, based on the optical density. There are several different risks to using this model. One major risk is that the optical density of a test sample must lie within the range of the model's standard curve in order to have it's concentration interpolated accurately, therefore test samples with optical densities outside of this range may have incorrectly predicted concentrations. **Classification category**: Microbiota enzyme production relating to disease onset/progression. A class of enzymes produced by certain bugs present in the human gut microbiome has recently been shown to protect against the onset of Ulcerative Colitis (UC) in mouse models, as well as increase the odds of response to Immune Checkpoint Inhibitor therapy in mouse models of tumours. Using Shotgun metagenomic sequencing data gathered from patients in a longitudinal study looking at the onset of Crohn's disease (an Inflammatory Bowel disease like UC), I am building a model that looks for the relative abundance of these enzymes expressed in their gut microbiota and uses a random forest model to predict whether or not the patient is likely to develop Crohn's.

3 A)

$$J^{B}_{reg}(w) = \frac{1}{2m}\sum_{i=1}^{m}(y^{i} - t^{i})^2 + \frac{1}{2}\sum_{j=1}^{D} B_j w_j^2,$$

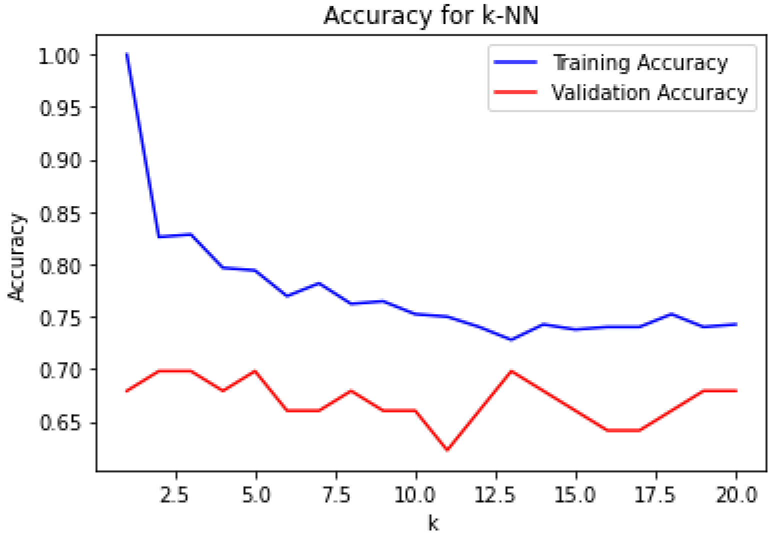
$w_j \leftarrow w_j - \alpha \cdot \partial J(w)/\partial w$ GD update rule

$\partial J(w)/\partial w$ partial derivative of cost

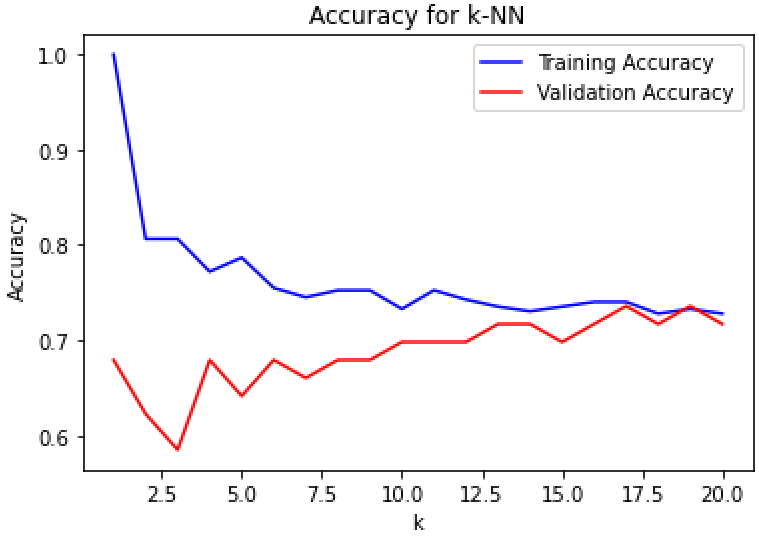
$(\partial J(w)/\partial w) + \lambda \cdot w$ Partial derivative of Reg cost w/ respect to w

$w_j \leftarrow w_j - \alpha \cdot (\partial J(w)/\partial w) + \lambda w)$

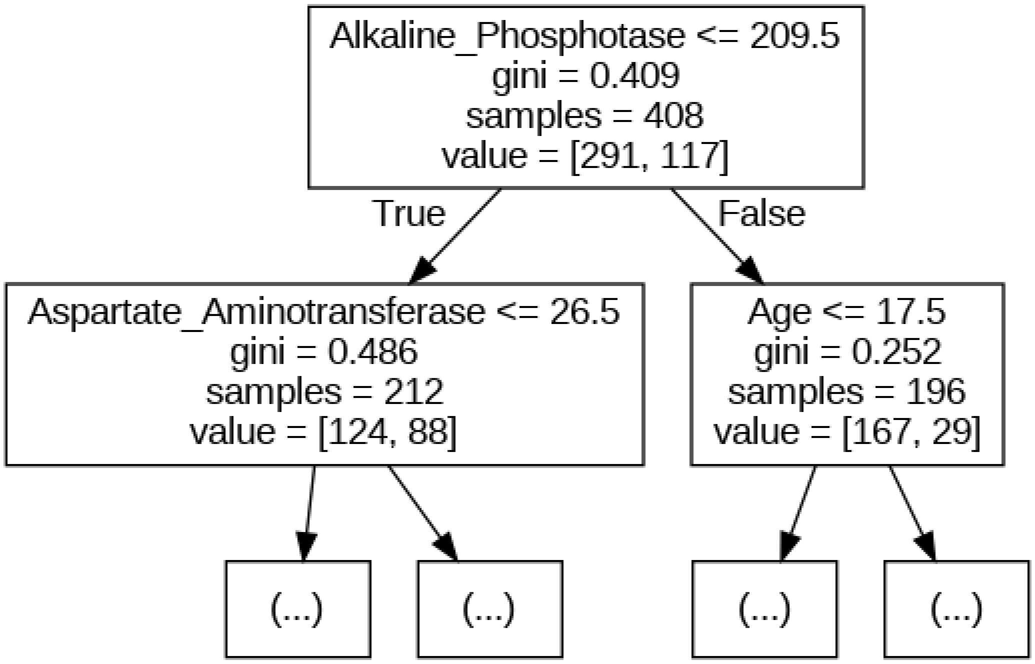
4 a) Plot of Training & Validation Accuracy in k-NN model:



4 C) Plot of Training & Validation Accuracy in k-NN model with the Cosine Metric:



5 B)


Alkaline_Phosphotase <= 209.5
gini = 0.409
samples = 408
value = [291, 117]
True
False
Aspartate_Aminotransferase <= 26.5
gini = 0.486
samples = 212
value = [124, 88]
Age <= 17.5
gini = 0.252
samples = 196
value = [167, 29]
(...)
(...)
(...)
(...)