CM146, Fall 2018
Problem Set 01: Jonathan Chu

October 26, 2018

1 Problem 2

(a) **Solution:**

$$Gain(S, X_j) = Entropy(S) - \sum_{v \in V} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, X_j) = B\left(\frac{p}{p+n}\right) - \sum_{v \in V} \frac{|S_v|}{|S|} B\left(\frac{p}{p+n}\right)$$

$$Gain(S, X_j) = B\left(\frac{p}{p+n}\right)\left(1 - \sum_{v \in V} \frac{|S_v|}{|S|}\right) = 0$$

2 Problem 3

(a) **Solution:** Because a point is considered its own neighbor, the value of k that minimizes the training set error is 1. The resulting training error is 0, since our prediction will always equal the training example's value.

(b) **Solution:** Since our task is binary {-1, 1}, a very large value of k would cause the average output of neighbors to be very close to 0, meaning our prediction will always be low confidence.

A very small value of k could result in overfitting, since we only consider very similar examples in the training set when making a prediction. Outliers in the training data would have more impact on our predictions.

(c) **Solution:** By inspection, there is no value of k for which we can correctly predict the 2 upper +'s and the 2 lower -'s in a leave-one-out setting. Thus, the minimum error we can hope to achieve is 4/14.

Intuitively, we seek k values that encompass all or at least most of the other points in the same group of 7.

Values of k that achieve this are 5 and 7. 6 is not guaranteed to achieve minimum error because, for example, (1, 5) is equidistant to both (5, 1) which is not the desired prediction, and (5, 9) which is.