Marc Soda
HW 2

1) $H(all) = -5/9 \log_2 5/9 - 4/9 \log_2 4/9$
$= .991$

$H(S+) = -3/5 \log_2 3/5 - 2/5 \log_2 2/5$
$= .971$

$H(S-) = -2/4 \log_2 2/4 - 2/4 \log_2 2/4$
$= 1$

$H(all \mid S) = (5/9 \cdot .971) + (4/9 \cdot 1)$
$= .984$

$H(A+) = -5/6 \log_2(5/6) - (1/6 \log_2(1/6))$
$= .650$

$H(A-) = 0$

$H(all \mid A) = (4/9 \cdot .650) + (0)$
$= .289$

$IG(all, S) = .991 - .984$
$= .007$

$IG(all, A) = .991 - .289$
$= .702$

Not enough info provided to calculate this by hand so I had to use a calculator.

∴ we should choose feature A for the 1st split because $IG(all\ S) < IG(all, A)$

2) a)

+	1: 4.4 = +
+	3: 4.4 = + | 4.7 = + | 4.9 = +
+	5: 4.4 = + | 4.7 = + | 4.9 = + | 5.1 = − | 5.4 = −
−	9: 4.4 = + | 4.7 = + | 4.9 = + | 5.1 = − | 5.4 = − | 5.7 = + | 3.0 = − | 1.5 = − | 7.5

1 neighbor: +
3 neighbors: +
5 neighbors: +
9 neighbors −

b)

w(1.5) = .111	w(4.7) = 25	w(5.4) = 1.23
w(3.0) .444	w(4.9) = 6.25	w(5.7) = .694
w(4.4) = 100	w(5.1) = 2.78	w(7.5) = .111

K=1: +: 100	−: 0
K=3: +: 131.25	−: 0
K=5: +: 131.25	−: 4.01
k=9: +: 131.944	−: 4.676

Each K value classifies 4.5 as +

c) The distance-weighted voting approach factors the distance of each neighbor to the point being classified. This allows the closer neighbors to have a greater impact on classificatio than the farther ones. This makes the classifier less sensitive to higher values of K, reducing the impact of K for KNN classifiers compared to the majority vote approach.

3) a) Larger margins (may) reduce overfitting because it decreases the liklihood of misclassifying data that are not present in the testing data. It also reduces the effect of noise.

b) According to the textbook:
$w^Tx + b = 0$ is the equation of a generic separating hyperplane.
The distance of any point to the hyperplane is $D(x) |w^Tx + b| / ||w||$.

To find the maximum hyperplane:
maximize M, w, and b such that
$$y_i(w^Tx_i + b) \geq M \text{ for all } i$$

Solving the above optimization problem is all that needs to be done to find the maximum-margin hyperplane. The solving of this problem is agnostic to the number of datapoints from each class.

∴ A dataset consisting of just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane.

I found all of the above information directly in the textbook.

4) i 1-NN
ii 3-NN

For 1-NN, the left center point would be misclassified because it would assign each point to the same class as its nearest Neighbor

For 3-NN, the left center point would be correctly classified because each point would be assigned to the class of its 3 closest neighbors.

∴ 1-NN would have a higher Leave-one-out cross-validation score.

5) 1) ANN and CNN are both types of deep learning models that are widely used today. Both are based on the concept of inter-connected layered nodes with different weights and biases that are able to be optimally tuned using some type of backpropagation in order to allow the model to affectively "learn". The main difference between ANNs and CNNs is their applications. ANNs are able to be applied more generally to solve a much wider breadth of problems while CNNs are generally used when more spatial learning is required, such as in image or video classification. ANN's are typically used with more tabular data which allows them to be more generally applied.

2) Iterations = # samples / batch size

$300 = 600 / 2$

300

3) ~~I assume the pad length is 0~~

$((20-6+0)/2)+1) = 8$

8x8

4) $ceil((20-6+2\cdot2)+1))/2 = 10$ = height and width

depth is # filters = 5

10x10x5

5) Put simply, max pooling selects the max value within the ~~applied~~ filter or window and average pooling will select the average value, based on figure 1.

6) 4x4