WU1 (10%): Answer questions A, B and C for both OAA and AVA. (Questions about “indicative” are open-ended. Any reasonable answers with analysis will be credited.)

Note my definition of "words that are indicative of label x" is basically what combination of words result in the most label x observations in the leaves of the corresponding tree(s), given that the majority of observations in that leaf is of label x. This is because you have to look at the words in context of the splits (hence the combination criteria) and you want the words that influence the most observations (hence the greedy criteria), and lastly you still want to make sure you predict x (hence the majority criteria).

(A) Train depth 3 decision trees on the WineDataSmall task. What words are most indicative of being Sauvignon-Blanc? Which words are most indicative of not being Sauvignon-Blanc? What about Pinot-Noir (label==2)? (OAA)

The words most indicative of being Sauvignon-Blanc are the combination of words (considering both their appearance (Y) or lack thereof (N)) that result in the most class 1 observations in the Sauvignon-Blanc vs All tree's leaves, granted that the number of class 1 observations are greater than the number of class 0 observations (i.e. prediction is class 1). The top 2 are:

1. no citrus, has lime, no apple (15/16 class 1)
2. has citrus, has grapefruit, no extremely (14/14 class 1)

The words most indicative of NOT being Sauvignon-Blanc are the words that result in the most class 0 observations in the Sauvignon-Blanc vs All tree's leaves, granted that the number of class 0 observations are greater than the number of class 1 observations (i.e. prediction is class 0). The top 2 are:

1. no citrus, no lime, no gooseberry (356/366 class 0)
2. has citrus, no grapefruit, has flavors (11/16 class 0)

The words most indicative of being Pinot-Noir are the combination of words (considering both their appearance (Y) or lack thereof (N)) that result in the most class 1 observations in the Pinot-Noir vs All tree's leaves, granted that the number of class 1 observations are greater than the number of class 0 observations (i.e. prediction is class 1). The top 2 are:

1. has cherry, no cassis, no verdot (68/104 class 1)
2. no cherry, has raspberries, no integrated (12/12 class 1)

The words most indicative of NOT being Pinot-Noir are the combination of words that result in the most class 0 observations in the Pinot-Noir vs All tree's leaves, granted that the number of class 0 observations are greater than the number of class 1 observations (i.e. prediction is class 0). The top 2 are:

1. no cherry, no raspberries, and no strawberry (225/283 class 0)
2. has cherry, has cassis, no allspice (21/21 class 0)

(B) Train depth 3 decision trees on the full WineData task (with 20 labels). What accuracy do you get? How long does this take (in seconds)? One of my least favorite wines is Viognier -- what words are indicative of this? (OAA)

I got a 36.7% test accuracy. It takes 0.3 seconds to train the model and 2 seconds to predict using the test set.

The words most indicative of being Viognier are the combination of words (considering both their appearance (Y) or lack thereof (N)) that result in the most class 1 observations in the Viognier vs All tree's leaves, granted that the number of class 1 observations are greater than the number of class 0 observations (i.e. prediction is class 1). The top 2 are:

1. has peaches, has milk (3/3 class 1)
2. has peaches, no milk, has straw (1/1 class 1)

(C) Compare the accuracy using zero-one predictions versus using confidence. How much difference does it make? (OAA)

The test accuracy I got using zero-one is 24.6%. This is about 12% less accuracy compared to when using confidence which is a big difference (for the worse). Hence, we can say using confidence is better than using zero-one when using OAA.

(A) Train depth 3 decision trees on the WineDataSmall task. What words are most indicative of being Sauvignon-Blanc? Which words are most indicative of not being Sauvignon-Blanc? What about Pinot-Noir (label==2)? (AVA)

The words most indicative of being Sauvignon-Blanc are the combination of words (considering both their appearance (Y) or lack thereof (N)) that result in the most Sauvignon-Blanc labeled observations in all the trees involving Sauvignon-Blanc, granted that the final prediction is Sauvignon-Blanc. The top 4 are:

1. no thai, no very, no planted (56/60 Sauvignon-Blanc)
2. no apple, no pasta, no quite (56/67 Sauvignon-Blanc)
3. has citrus (31/31 Sauvignon-Blanc)
4. has crisp, no red (30/30 Sauvignon-Blanc)

The words most indicative of NOT being Sauvignon-Blanc are the combination of words that result in the most non-Sauvignon-Blanc labeled observations in all the trees involving Sauvignon-Blanc, granted that the final prediction is NOT Sauvignon-Blanc. The top 4 are:

1. no citrus, no lime, no melon (187/196 not Sauvignon-Blanc)
2. no crisp, no lime, no lemon (141/150 not Sauvignon-Blanc)
3. has apple, no bright (10/10 not Sauvignon-Blanc)
4. has thai (5/5 not Sauvignon-Blanc)

The words most indicative of being Pinot-Noir are the combination of words (considering both their appearance (Y) or lack thereof (N)) that result in the most Pinot-Noir labeled observations in all the trees involving Pinot-Noir, granted that the final prediction is Pinot-Noir. The top 4 are:

1. no crisp, no peach, no pear (142/145 Pinot-Noir)
2. no straw, no crisp, no example (142/150 Pinot-Noir)
3. no crisp, no lime, no lemon (141/150 Pinot-Noir)
4. no cassis, has acidity, no tannins (22/22 Pinot-Noir)

The words most indicative of NOT being Pinot-Noir are the combination of words that result in the most non-Pinot-Noir labeled observations in all the trees involving Pinot-Noir, granted that the final prediction is NOT Pinot-Noir. The top 4 are:

1. no cassis, no acidity, no duck (129/221 not Pinot-Noir)
2. has cassis, no tea, no 100 (47/48 not Pinot-Noir)
3. has crisp, no red (30/30 not Pinot-Noir)
4. no crisp, has lime, no game (13/13 not Pinot-Noir)

(B) Train depth 3 decision trees on the full WineData task (with 20 labels). What accuracy do you get? How long does this take (in seconds)? One of my least favorite wines is Viognier -- what words are indicative of this? (AVA)

I got a 26.2% test accuracy. It takes 0.6 seconds to train the model and 22 seconds to predict using the test set.

The words most indicative of being Viognier are the combination of words (considering both their appearance (Y) or lack thereof (N)) that result in the most Viognier labeled observations in all the trees involving Viognier, granted that the final prediction is Viognier. The top 4 are:

1. has fruits (4/4 Viognier)
2. has preaches (4/4 Viognier)
3. has floral (4/4 Viognier)
4. has peach (4/4 Viognier)

(C) Compare the accuracy using zero-one predictions versus using confidence. How much difference does it make? (AVA)

The test accuracy I got using zero-one is 26.1%. This is about 0.1% less accuracy compared to when using confidence which is a small difference (for the worse). This difference is so small, that we can say using zero-one or confidence when using AVA doesn't make a significant difference.

WU2 Show the test accuracy you get with a balanced tree on the WineData using a DecisionTreeClassifier with max depth 3.

The test accuracy I got using a balanced tree on the WineData using a DecisionTreeClassifier with max depth 3 is 30.8%.

WU3 (5%): What is the impact of the step size on convergence? Find values of the step size where the algorithm diverges and converges.¶

If we suppose values of x < 0.0005 are "good enough" to call it convergence (to the min of x = 0), then 0.25 <= step size <= 6.33 results in convergence when #iterations = 100. On the other hand, step size < 0.25 or step size > 6.33 results in divergence. Step size impacts convergence by influencing the rate of convergence. When the step size is too small or too big it takes longer for gradient descent to converge, for instance if we increase #iterations to 200, step sizes 0.2 and 7 end up converging but only after 100 iterations. Graphically you can see that values too small just converge really slowly, whereas values too big initially diverge, but due to our adaptive step size (which makes step size smaller), they do eventually converge (if you increase #iterations to be large enough i.e. expand the x-axis to the right).

However, clearly step size <= 0 will never converge no matter how many iterations you use. This is because a 0 step size remains a 0 step size even when using adaptive step size (0/k = 0 for all k), hence the x0 never moves. Also, step sizes < 0 will just move the x0 in the wrong direction along the gradient (step size remains negative even with adaptive step size because we only divide step size by a positive number, and neg/pos = neg), hence why it never converges. This can be seen in the first graph, where none of the lines converge even after 2000 iterations.

WU4 (10%): Come up with a non-convex univariate optimization problem. Plot the function you're trying to minimize and show two runs of gd, one where it gets caught in a local minimum and one where it manages to make it to a global minimum. (Use different starting points to accomplish this.)

Here you can see for the non-convex univariate function f(x) = x^4 - x^3 - 3x^2 + 1, gradient descent starting at x0 = -3 gets caught in a local minima which is x = -0.905, f(x) = -0.04, whereas x0 = 3 finds the global minima which is x = 1.65, f(x) = -4.24. Graphically you can see this in both the trajectory plot and the f(x) plot. This happens because x0 = 3 is the to the right of the global minima so gradient descent descends to that global valley, whereas x0 = -3 is to the left of the local minima so gradient descent descends to that local valley.

WU5 (5%): For each of the loss functions, train a model on the binary version of the wine data (called WineDataBinary) and evaluate it on the test data. You should use lambda=1 in all cases. Which works best? For that best model, look at the learned weights. Find the words corresponding to the weights with the greatest positive value and those with the greatest negative value. Hint: look at WineDataBinary.words to get the id-to-word mapping. List the top 5 positive and top 5 negative and explain.

Clearly the logistic loss worked the best because it had the highest test accuracy out of the three 97.4% (whereas squared loss had 24% and hinge loss had 75%). The reason I think logistic loss did the best is because it is smoother (compared to hinge loss) and it doesn't cater towards prediction outliers as much (compared to squared loss).

Here you can see that WineDataBinary is a dataset that only looks at the observations from label 0 or label 1, which corresponds to Sauvignon-Blanc and Cabernet-Sauvignon respectively. According to the code, we relabel Sauvignon-Blanc (label 0) as 1 and Cabernet-Sauvignon (label 1) as -1.

Since we know that the X data is either 0 or 1 (indicating if a word is present or not), we know that greater weights produce a greater prediction which drives you towards predicting Sauvignon-Blanc (+1). On the other hand, lower weights produce a lower prediction which drives you towards predicting Cabernet-Sauvignon (-1).

As a result, we can say that according to the logistic loss linear classifier, the words most indicative of Sauvignon-Blanc (i.e. having the highest weights) as opposed to Cabernet-Sauvignon are:

1. citrus
2. crisp
3. lime
4. acidity

5. tropical

*Note "indicative" meaning: if the word appears it helps the most with predicting Sauvignon-Blanc by increasing the prediction the most (the word must appear i.e. =1 in order to make use of the weight)

Also, we can say that according to the logistic loss linear classifier the words most indicative of Cabernet-Sauvignon (i.e. having the lowest weights) as opposed to Sauvignon-Blanc are:

1. tannins
2. black
3. dark
4. cherry
5. blackberry

*Note that these words are the most indicative because if they appear they help the most with predicting Cabernet-Sauvignon by decreasing the prediction the most