Assignment #1

Anton Sitkovets
21211804

(1) $P(R=1|D=1) = P(R=0|D=0) = \theta$
$P(R=0|D=1) = P(R=1|D=0) = 1-\theta$
$P(D=1) = \alpha$
$P(D=0) = 1-\alpha$

a) $P(D=1|R=1) = \frac{P(R=1|D=1)P(D=1)}{P(R=1|D=1)P(D=1) + P(R=1|D=0)P(D=0)}$

$= \frac{\theta\alpha}{\theta\alpha + (1-\theta)(1-\alpha)}$

b) $P(D=1|R=0) = \frac{P(R=0|D=1)P(D=1)}{P(R=0|D=1)P(D=1) + P(R=0|D=0)P(D=0)}$

$= \frac{(1-\theta)\alpha}{(1-\theta)\alpha + \theta(1-\alpha)}$

$P(D=1 | R_1=1, R_2=0)$
$= P(D=1|R_1=1) \cdot P(D=1|R_2=0)$
$= \frac{\theta\alpha}{\theta\alpha + (1-\theta)(1-\alpha)} \cdot \frac{(1-\theta)\alpha}{(1-\theta)\alpha + \theta(1-\alpha)}$

c) $\theta = 0.99$, $\alpha = 0.001$
$N = 1000$ patients

$P(D=1|R=0) = \frac{(1-0.99)(0.001)}{(1-0.99)(0.001) + (0.99)(1-0.01)}$

$= 0.00001011$

$M = 1000 \cdot 0.00001011 \approx 0.01$

∴ Average of 1% of patients actually have the disease.

(2) $P(S=4)=\frac{1}{4}$ $P(S=8)=\frac{1}{2}$ $P(S=12)=\frac{1}{4}$

$$\sum_{s\in S} P(X=x, S=s) = \sum_{s\in S} P(X=x \mid S=s) P(S=s)$$

a) $$P(X=3) = \sum_{s\in\{4,8,12\}} P(X=3, S=s)$$

$$= P(X=3, S=4) + P(X=3, S=8) + P(X=3, S=12)$$

$$= \frac{1}{4}\cdot\frac{1}{4} + \frac{1}{8}\cdot\frac{1}{2} + \frac{1}{12}\cdot\frac{1}{4} = 0.1458 \approx 0.15$$

b) $$P(S=12 \mid X=3) = \frac{P(X=3 \mid S=12) P(S=12)}{\sum_{s\in S} P(X=3 \mid S=s)} = \frac{\frac{1}{12}\left(\frac{1}{4}\right)}{0.15}$$

$$= 0.13\overline{8} \approx 0.14$$

c) Since the events X and Y are independent

$$P(X=3, Y=7 \mid S=12) = P(X=3 \mid S=12) P(Y=7 \mid S=12)$$

$$P(S=12 \mid X=3, Y=7) = \frac{P(X=3, Y=7 \mid S=12) P(S=12)}{\sum_{s\in\{8,12\}} P(X=3, Y=7 \mid S=s)}$$

$$= \frac{P(X=3 \mid S=12) P(Y=7 \mid S=12) P(S=12)}{\sum_{s\in\{8,12\}} P(X=3 \mid S=s) P(Y=7 \mid S=s)}$$

$$= \frac{\frac{1}{12}\cdot\frac{1}{12}\cdot\frac{1}{4}}{P(X=3 \mid S=8) P(Y=7 \mid S=8) + P(X=3 \mid S=12) P(Y=7 \mid S=12)}$$

$$= \frac{\frac{1}{12}\cdot\frac{1}{12}\cdot\frac{1}{4}}{\left(\frac{1}{8}\right)^2 + \left(\frac{1}{12}\right)^2} = \frac{1}{13} = 0.077$$

d) $P(Z=2 \mid X=3, Y=7)$ Z, Y, X are independent so $P(X=3, Y=7 \mid Z=2) = P(X=3, Y=7)$

$$= P(Z=2)$$

$$= \sum_{s\in\{8,12\}} P(Z=2, S=s)$$

$$= P(Z=2, S=8) + P(Z=2, S=12)$$

$$= \frac{1}{8}\cdot\frac{1}{2} + \frac{1}{12}\cdot\frac{1}{4} = \frac{1}{16} + \frac{1}{48} = \frac{1}{12}$$

(3) $P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$ H_i is the hypothesis that the prize is behind door i

$P(D=2|H_1) = \frac{1}{2}$ | $P(D=2|H_2) = 0$ | $P(D=2|H_3) = 1$

$P(D=3|H_1) = 1/2$ | $P(D=3|H_2) = 1$ | $P(D=3|H_3) = 0$

$$P(H_i|D=3) = \frac{P(D=3|H_i)\,P(H_i)}{P(D=3)}$$

$$P(H_1|D=3) = \frac{(1/2)(1/3)}{P(D=3)} \quad | \quad P(H_2|D=3) = \frac{(1)(1/3)}{P(D=3)}$$

$$P(H_3|D=3) = \frac{(0)(1/3)}{P(D=3)}$$

$P(D=3) = 1/2$ b/c it is normalizing constant for posterior distribution

$P(H_1|D=3) = 1/3$ | $P(H_2|D=3) = 2/3$ | $P(H_3|D=3) = 0$

★ ∴ (contestant ~~to~~ should switch to door 2 in order to have greatest chance of winning.

(4) $Var[X+Y] = Var[X] + Var[Y] + 2\,cov[X,Y]$

$Var[w] = E(w - E(w))^2$

let $w = X+Y$

$Var(X+Y) = E(X+Y - E(X+Y))^2$

$= E(X+Y-(E(X)+E(Y))^2$

$= E((X-E(X)) + (Y-E(Y)))^2$

$= E[((X-E(X))^2 + (Y-E(Y))^2 + 2(X-E(X))(Y-E(Y))]$

$= E(X-E(X))^2 + E(Y-E(Y))^2 + E[2(X-E(X))(Y-E(Y))]$

$= Var(X) + Var(Y) + 2E[(X-E(X))(Y-E(Y))]$

$= Var(X) + Var(Y) + 2Cov(X,Y)$

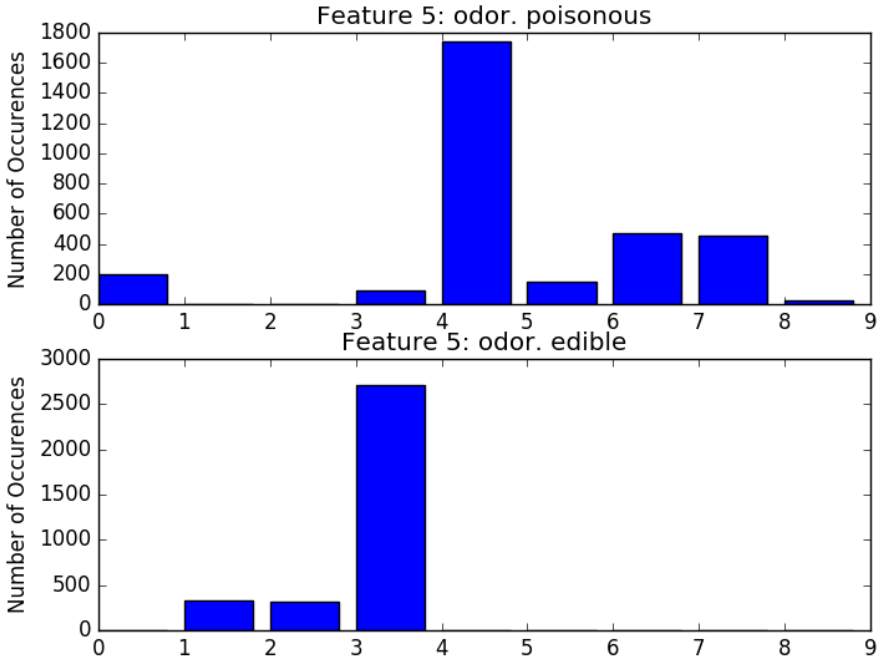
Assignment 1
EECS 4404
Anton Sitkovets 212118048

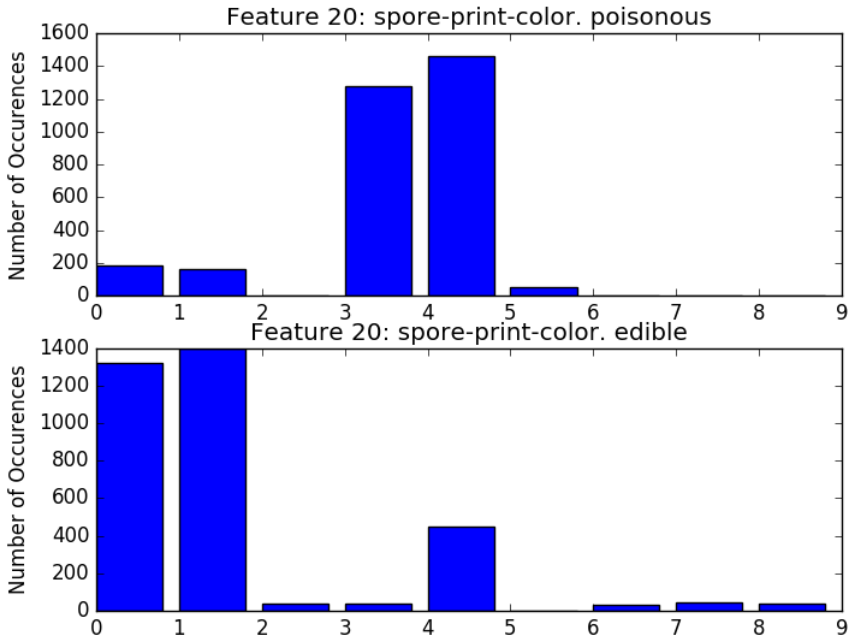
STEP 1

The result from my calculation of the percentage of edible vs poisonous data in the training set:

Out of 6499, 48.2228035082% are poisonous and 51.7771964918% are edible.

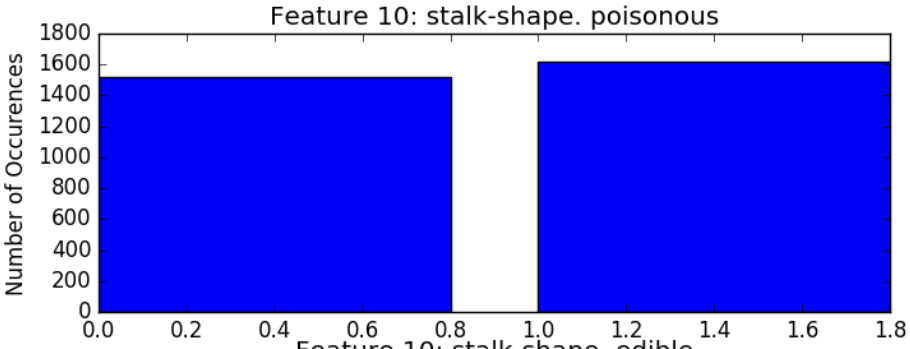
Two features that I have determined are important for determining edibility from the histogram are of features 5 and 20.

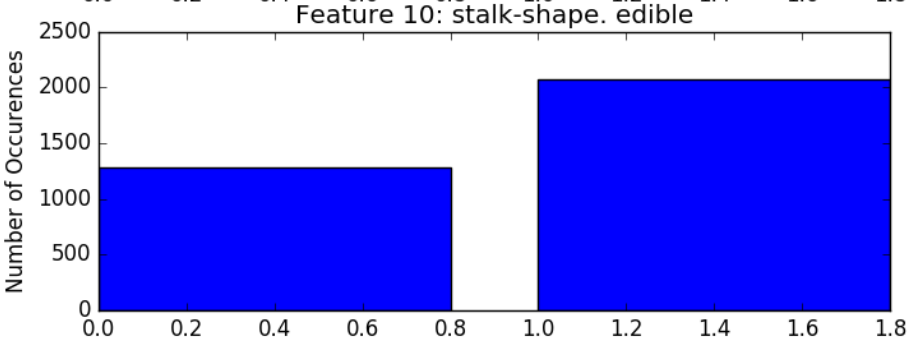


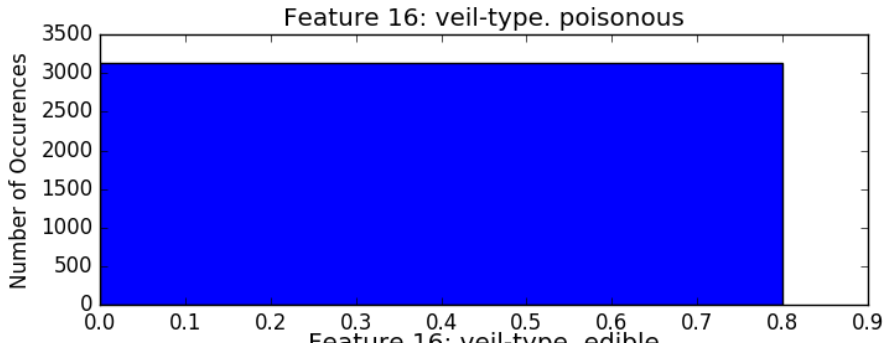


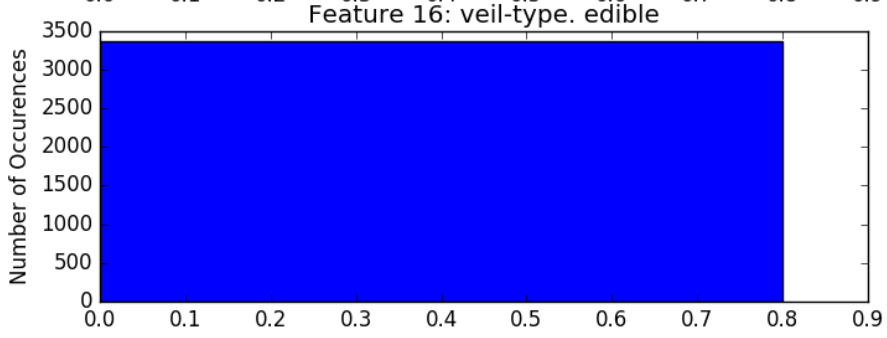
I chose these two features, because there seems to be a clear distinction between the values for the features for when a mushroom is poisonous vs when it is edible. With feature 5 we can see that when the odor is in category 1, 2, or 3 the mushroom will have a very high chance of being edible. Whereas if the mushroom does not have an odor of those three categories we know it is poisonous. As well there is a clear distinction between the edible and poisonous histograms for feature 20. Seeing as the probability distributions for feature value of feature 5 and 20, are so contrasting between the edible vs non-edible histograms, I think it would be safe to just use these two feature to determine edibility of a mushroom. I say this because all the other features do not present such clear contrasts and do not make it as obvious whether a mushroom is edible or not.

Whereas features 10 and 16, are not important for determining edibility because the probability between the feature values are not distinct enough. For the veil type, there is a 50/50 chance that a veil-type of category 0, can be edible or not edible. This type of probability doesn't not provide us with a high degree of certainty because we want the two histograms to be drastically opposing so we can have a clear definition of when a mushroom is edible or not.







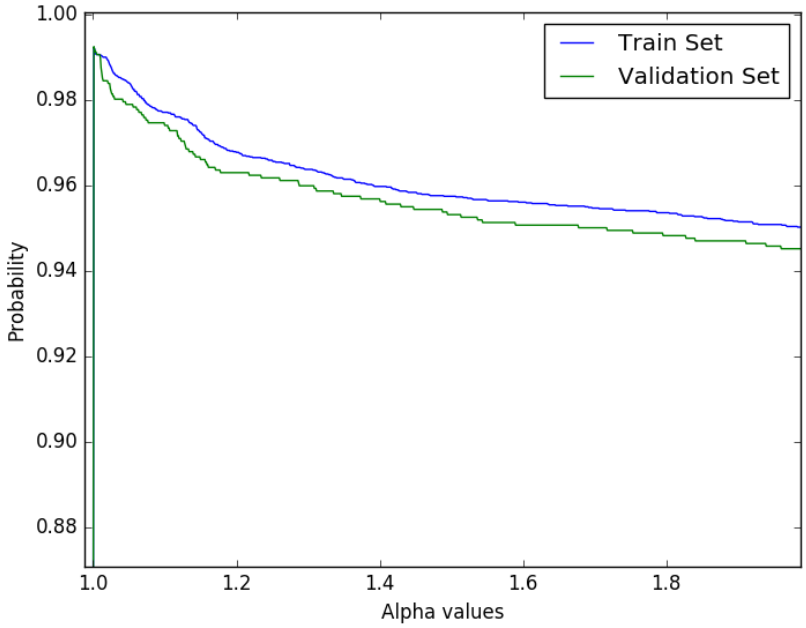


Step 2

The training set accuracy and the validation accuracy are different because we based our probability distribution values off the training set, so if there are cases that the training set has

not seen, then it is not guaranteed that it will predict the correct classification. For instance, if we have never seen a cap-color of black in the training set, then we do not have a good classifier to justify whether it is edible or not if it encounters a black cap color in the test data.

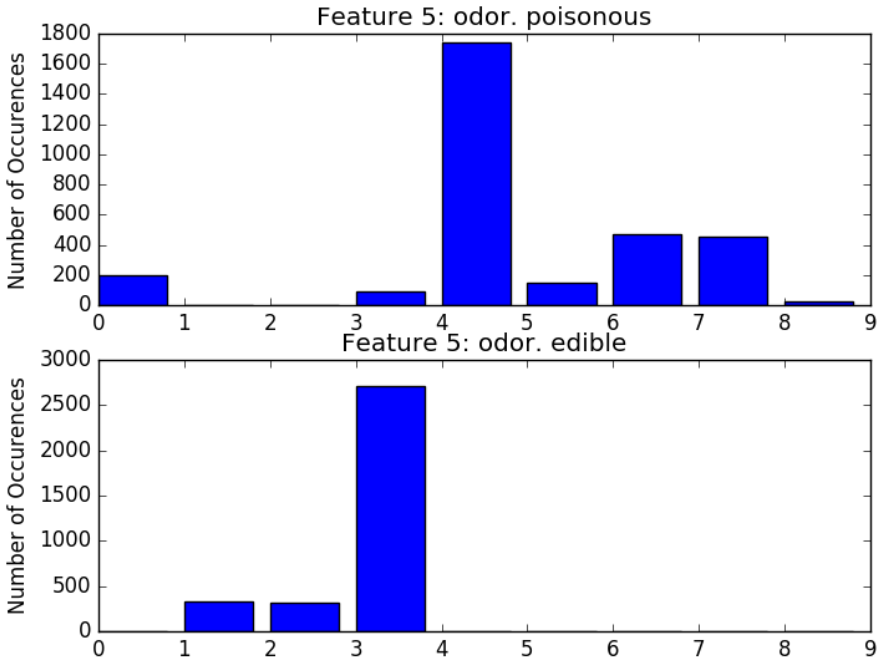
Accuracy will decrease as alpha increases because if we look at the MAP estimate in the Dirchlet model: $theta^k = (N^k + a - 1)/(N + K*a - K)$, when a is small the value for $theta^k$ will be reliant on how many values we have seen of k in N. This means the probability of feature value k depends solely on how many of those features have been seen in the training data. But as we move further away and as alpha moves to infinite, we start to care less and less about how often each outcome value k has appeared and more so that each outcome has an equal probability to appear. Say there are 5 possible outcomes, if we have a large alpha, the multinomial distribution will be normalized to show that each outcome has an equal probability. So, this leads to the notion that all outcomes are equally probable, which is usually not the case. In our training data, not all feature values are equally likely for each feature, so this would result in inaccurate results. This explains why as alpha gets bigger, the accuracy decreases as we begin to believe the probability of the feature values to be more equal leading to wrong classifications.



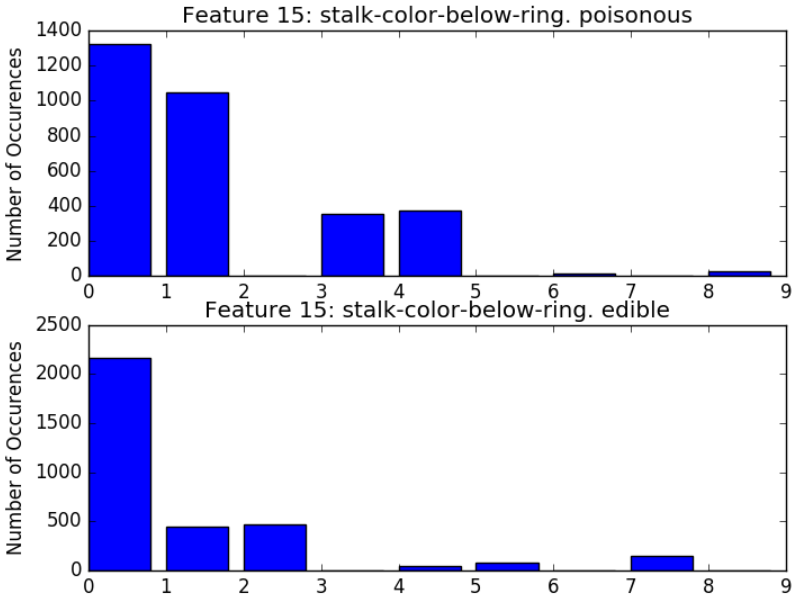
Step 3

The data that I have printed below is structured such that each array is [i, f_i, absolute difference, non-absolute difference], where i is the feature number and f_i is the value of that feature. I then sorted the values by their absolute values and then their non-absolute values. When looking at the values with high absolute value I found that feature 5(odor), attribute 4 is highly discriminative and this can also be seen in the graph, as there are zero cases in the edible

mushrooms but many in the poisonous mushrooms. Hence we can say that this means that the value when not absolute is large and negative, and therefore poisonous.



Using the not absolute difference values we can see a feature that is large and positive is feature 15, attribute 2. Looking at the graph we can see this is because there are no occurrences of the attribute among the poisonous mushrooms, but many among the edible ones. This means that it is very highly likely that a mushroom with this attribute and feature combination is edible.



Sorted absolute difference: [[16, 0, 0, 0], [1, 3, 0.029710531600497192, -0.029710531600497192], [11, 2, 0.035328625512463852, -0.035328625512463852], [6, 0, 0.040793288254338547, -0.040793288254338547], [17, 0, 0.043995295468132439, -0.043995295468132439], [1, 0, 0.061772382537118253, 0.061772382537118253], [18, 0, 0.10175617038182183, -0.10175617038182183], [3, 0, 0.13327441780378724, 0.13327441780378724], [3, 3, 0.15797336995611366, 0.15797336995611366], [10, 1, 0.17999009577518388, 0.17999009577518388], [9, 3, 0.22540059182910399, 0.22540059182910399], [2, 1, 0.22633647782790101, -0.22633647782790101], [10, 0, 0.2353031575437633, -0.2353031575437633], [2, 0, 0.26891537819757216, -0.26891537819757216], [7, 0, 0.30689402929384668, -0.30689402929384668], [22, 3, 0.31626345289441871, 0.31626345289441871], [21, 4, 0.40992443330149664, 0.40992443330149664], [14, 0, 0.41354300452887999, 0.41354300452887999], [3, 4, 0.42131943388092497, -0.42131943388092497], [15, 0, 0.4244635084975234, 0.4244635084975234], [3, 5, 0.47657513261534001, -0.47657513261534001], [12, 3, 0.53494232420385845, 0.53494232420385845], [22, 1, 0.58521471329145047, 0.58521471329145047], [3, 1, 0.58796611914735109, -0.58796611914735109], [19, 1, 0.64507582773268757, -0.64507582773268757], [2, 2, 0.64728452823658267, 0.64728452823658267], [13, 0, 0.72970956057371272, 0.72970956057371272], [8, 1, 0.76274955133475086, 0.76274955133475086], [12, 0, 0.79439341719808376, 0.79439341719808376], [9, 2, 0.79883348300666945, -0.79883348300666945], [3, 2, 0.84941686223805579, 0.84941686223805579], [21, 0, 0.86087425537861773, 0.86087425537861773], [14, 2, 0.89849069075202337, -0.89849069075202337], [4, 1, 0.903686971288164, -0.903686971288164], [15, 1, 0.92032991161409217, -0.92032991161409217], [21, 3, 0.93584235113825565, -0.93584235113825565], [13, 2, 0.93893038359214742, 0.93893038359214742], [22, 6, 0.98106943475645614, -

0.98106943475645614], [3, 8, 0.98968171642409608, 0.98968171642409608], [11, 4, 0.9916536655638714, -0.9916536655638714], [9, 5, 1.0403014034065359, -1.0403014034065359], [1, 4, 1.0559677227496773, -1.0559677227496773], [12, 1, 1.0740080792912039, 1.0740080792912039], [3, 6, 1.082700835366857, -1.082700835366857], [9, 10, 1.0846128427426169, 1.0846128427426169], [22, 0, 1.0959552556952121, -1.0959552556952121], [13, 1, 1.1124587517917983, 1.1124587517917983], [11, 0, 1.1643502152432237, 1.1643502152432237], [20, 4, 1.2533783887172643, -1.2533783887172643], [9, 4, 1.2864128242646877, 1.2864128242646877], [19, 0, 1.3187461815827908, 1.3187461815827908], [4, 0, 1.450516315323449, 1.450516315323449], [9, 0, 1.5717034184046765, 1.5717034184046765], [21, 5, 1.5938707711051561, 1.5938707711051561], [20, 0, 1.9140386895444523, 1.9140386895444523], [18, 1, 1.9328005050459409, 1.9328005050459409], [22, 2, 2.0082914691139577, 2.0082914691139577], [20, 1, 2.0710929276843637, 2.0710929276843637], [1, 1, 2.0736827510950921, 2.0736827510950921], [15, 4, 2.0779711592387988, -2.0779711592387988], [8, 0, 2.0967851130688162, -2.0967851130688162], [22, 4, 2.1007187910934482, -2.1007187910934482], [6, 1, 2.1735174636683778, 2.1735174636683778], [9, 1, 2.2054805456356372, 2.2054805456356372], [9, 6, 2.2830282874314212, 2.2830282874314212], [7, 1, 2.3278114336158167, 2.3278114336158167], [11, 1, 2.3997772681944647, 2.3997772681944647], [13, 3, 2.7959766006574638, -2.7959766006574638], [12, 2, 2.9004300615415208, -2.9004300615415208], [5, 3, 3.3326582556218476, 3.3326582556218476], [20, 3, 3.558232577104754, -3.558232577104754], [14, 3, 3.5675370333500656, -3.5675370333500656], [2, 3, 8.0778185769865409, -8.0778185769865409], [1, 5, 8.3654172965950924, -8.3654172965950924], [14, 8, 8.7707990230149946, -8.7707990230149946], [17, 3, 8.7707991325356875, -8.7707991325356875], [3, 9, 9.2346238566182723, 9.2346238566182723], [3, 7, 9.4016639564410056, 9.4016639564410056], [15, 6, 9.6180016566434059, -9.6180016566434059], [9, 9, 9.9233644446557605, -9.9233644446557605], [1, 2, 9.9277254981152439, 9.9277254981152439], [20, 6, 10.363027592905226, 10.363027592905226], [5, 8, 10.380103615447563, -10.380103615447563], [14, 7, 10.380103615447563, -10.380103615447563], [15, 8, 10.380103615447563, -10.380103615447563], [19, 4, 10.380103703064091, -10.380103703064091], [18, 2, 10.380103746872436, -10.380103746872436], [19, 3, 10.500224856166911, 10.500224856166911], [20, 8, 10.525542110758316, 10.525542110758316], [20, 2, 10.550234113607647, 10.550234113607647], [20, 7, 10.574331084480612, 10.574331084480612], [9, 11, 10.74868072789728, 10.74868072789728], [20, 5, 10.986224267892982, -10.986224267892982], [17, 2, 11.099331508885355, 11.099331508885355], [9, 7, 11.113317730477625, 11.113317730477625], [14, 5, 11.16738415505138, 11.16738415505138], [15, 5, 11.16738415505138, 11.16738415505138], [17, 1, 11.218676682031941, 11.218676682031941], [22, 5, 11.840590630480975, 11.840590630480975], [11, 3, 11.853924028649901, 11.853924028649901], [14, 6, 11.853924116266427, 11.853924116266427], [15, 7, 11.853924116266427, 11.853924116266427], [5, 5, 11.962486372007566, -11.962486372007566], [5, 0, 12.277195267543155, -

12.277195267543155], [21, 2, 12.547067919872166, 12.547067919872166], [5, 2, 12.611192308086524, 12.611192308086524], [21, 1, 12.635733276003466, 12.635733276003466], [5, 1, 12.653751793106128, 12.653751793106128], [14, 4, 12.822420216585501, -12.822420216585501], [15, 3, 12.839661973472275, -12.839661973472275], [14, 1, 12.978677213079553, 12.978677213079553], [15, 2, 12.991497874113289, 12.991497874113289], [5, 7, 13.094767233993734, -13.094767233993734], [5, 6, 13.133733118164033, -13.133733118164033], [19, 2, 13.938272450425288, -13.938272450425288], [9, 8, 14.218088526950003, -14.218088526950003], [5, 4, 14.442810078870874, -14.442810078870874]]

Sorted differences: [[5, 4, 14.442810078870874, -14.442810078870874], [9, 8, 14.218088526950003, -14.218088526950003], [19, 2, 13.938272450425288, -13.938272450425288], [5, 6, 13.133733118164033, -13.133733118164033], [5, 7, 13.094767233993734, -13.094767233993734], [15, 3, 12.839661973472275, -12.839661973472275], [14, 4, 12.822420216585501, -12.822420216585501], [5, 0, 12.277195267543155, -12.277195267543155], [5, 5, 11.962486372007566, -11.962486372007566], [20, 5, 10.986224267892982, -10.986224267892982], [18, 2, 10.380103746872436, -10.380103746872436], [19, 4, 10.380103703064091, -10.380103703064091], [5, 8, 10.380103615447563, -10.380103615447563], [14, 7, 10.380103615447563, -10.380103615447563], [15, 8, 10.380103615447563, -10.380103615447563], [9, 9, 9.9233644446557605, -9.9233644446557605], [15, 6, 9.6180016566434059, -9.6180016566434059], [17, 3, 8.7707991325356875, -8.7707991325356875], [14, 8, 8.7707990230149946, -8.7707990230149946], [1, 5, 8.3654172965950924, -8.3654172965950924], [2, 3, 8.0778185769865409, -8.0778185769865409], [14, 3, 3.5675370333500656, -3.5675370333500656], [20, 3, 3.558232577104754, -3.558232577104754], [12, 2, 2.9004300615415208, -2.9004300615415208], [13, 3, 2.7959766006574638, -2.7959766006574638], [22, 4, 2.1007187910934482, -2.1007187910934482], [8, 0, 2.0967851130688162, -2.0967851130688162], [15, 4, 2.0779711592387988, -2.0779711592387988], [20, 4, 1.2533783887172643, -1.2533783887172643], [22, 0, 1.0959552556952121, -1.0959552556952121], [3, 6, 1.082700835366857, -1.082700835366857], [1, 4, 1.0559677227496773, -1.0559677227496773], [9, 5, 1.0403014034065359, -1.0403014034065359], [11, 4, 0.99165366556387147, -0.99165366556387147], [22, 6, 0.98106943475645614, -0.98106943475645614], [21, 3, 0.93584235113825565, -0.93584235113825565], [15, 1, 0.92032991161409217, -0.92032991161409217], [4, 1, 0.903686971288164, -0.903686971288164], [14, 2, 0.89849069075202337, -0.89849069075202337], [9, 2, 0.7988334830066945, -0.7988334830066945], [19, 1, 0.64507582773268757, -0.64507582773268757], [3, 1, 0.58796611914735109, -0.58796611914735109], [3, 5, 0.47657513261534001, -0.47657513261534001], [3, 4, 0.42131943388092497, -0.42131943388092497], [7, 0, 0.30689402929384668, -0.30689402929384668], [2, 0, 0.26891537819757216, -0.26891537819757216], [10, 0, 0.2353031575437633, -0.2353031575437633], [2, 1, 0.22633647782790101, -0.22633647782790101], [18, 0, 0.10175617038182183, -0.10175617038182183], [17, 0,

0.043995295468132439, -0.043995295468132439], [6, 0, 0.040793288254338547, -0.040793288254338547], [11, 2, 0.035328625512463852, -0.035328625512463852], [1, 3, 0.029710531600497192, -0.029710531600497192], [16, 0, 0, 0], [1, 0, 0.061772382537118253, 0.061772382537118253], [3, 0, 0.13327441780378724, 0.13327441780378724], [3, 3, 0.15797336995611366, 0.15797336995611366], [10, 1, 0.17999009577518388, 0.17999009577518388], [9, 3, 0.22540059182910399, 0.22540059182910399], [22, 3, 0.31626345289441871, 0.31626345289441871], [21, 4, 0.40992443330149664, 0.40992443330149664], [14, 0, 0.41354300452887999, 0.41354300452887999], [15, 0, 0.4244635084975234, 0.4244635084975234], [12, 3, 0.53494232420385845, 0.53494232420385845], [22, 1, 0.58521471329145047, 0.58521471329145047], [2, 2, 0.64728452823658267, 0.64728452823658267], [13, 0, 0.72970956057371272, 0.72970956057371272], [8, 1, 0.76274955133475086, 0.76274955133475086], [12, 0, 0.79439341719808376, 0.79439341719808376], [3, 2, 0.84941686223805579, 0.84941686223805579], [21, 0, 0.86087425537861773, 0.86087425537861773], [13, 2, 0.93893038359214742, 0.93893038359214742], [3, 8, 0.98968171642409608, 0.98968171642409608], [12, 1, 1.0740080792912039, 1.0740080792912039], [9, 10, 1.0846128427426169, 1.0846128427426169], [13, 1, 1.1124587517917983, 1.1124587517917983], [11, 0, 1.1643502152432237, 1.1643502152432237], [9, 4, 1.2864128242646877, 1.2864128242646877], [19, 0, 1.3187461815827908, 1.3187461815827908], [4, 0, 1.450516315323449, 1.450516315323449], [9, 0, 1.5717034184046765, 1.5717034184046765], [21, 5, 1.5938707711051561, 1.5938707711051561], [20, 0, 1.9140386895444523, 1.9140386895444523], [18, 1, 1.9328005050459409, 1.9328005050459409], [22, 2, 2.0082914691139577, 2.0082914691139577], [20, 1, 2.0710929276843637, 2.0710929276843637], [1, 1, 2.0736827510950921, 2.0736827510950921], [6, 1, 2.1735174636683778, 2.1735174636683778], [9, 1, 2.2054805456356372, 2.2054805456356372], [9, 6, 2.2830282874314212, 2.2830282874314212], [7, 1, 2.3278114336158167, 2.3278114336158167], [11, 1, 2.3997772681944647, 2.3997772681944647], [5, 3, 3.3326582556218476, 3.3326582556218476], [3, 9, 9.2346238566182723, 9.2346238566182723], [3, 7, 9.4016639564410056, 9.4016639564410056], [1, 2, 9.9277254981152439, 9.9277254981152439], [20, 6, 10.363027592905226, 10.363027592905226], [19, 3, 10.500224856166911, 10.500224856166911], [20, 8, 10.525542110758316, 10.525542110758316], [20, 2, 10.550234113607647, 10.550234113607647], [20, 7, 10.574331084480612, 10.574331084480612], [9, 11, 10.74868072789728, 10.74868072789728], [17, 2, 11.099331508885355, 11.099331508885355], [9, 7, 11.113317730477625, 11.113317730477625], [14, 5, 11.16738415505138, 11.16738415505138], [15, 5, 11.16738415505138, 11.16738415505138], [17, 1, 11.218676682031941, 11.218676682031941], [22, 5, 11.840590630480975, 11.840590630480975], [11, 3, 11.853924028649901, 11.853924028649901], [14, 6, 11.853924116266427, 11.853924116266427], [15, 7, 11.853924116266427, 11.853924116266427], [21, 2, 12.547067919872166, 12.547067919872166], [5, 2, 12.611192308086524, 12.611192308086524], [21, 1, 12.635733276003466, 12.635733276003466], [5, 1, 12.653751793106128, 12.653751793106128], [14, 1,

12.978677213079553, 12.978677213079553], [15, 2, 12.991497874113289, 12.991497874113289]]