COSC 522 HOMEWORK 5

PROBLEM ONE

L1	W_1	W_2	W_3
W_1	30	5	5
W_2	3	24	3
W_3	0	9	21

L2	W_1	W_2	W_3
W_1	25	2	3
W_2	3	22	5
W_3	5	6	19

Probabilities from confusion matrices

for L_1

$P(L_1=1|W_1) = 30/30$ $P(L_1=2|W_1) = 5/30$ $P(L_1=3|W_1) = 5/30$

$P(L_1=1|W_2) = 3/30$ $P(L_1=2|W_2) = 24/30$ $P(L_1=3|W_2) = 3/30$

$P(L_1=1|W_3) = 0/30$ $P(L_1=2|W_3) = 9/30$ $P(L_1=3|W_3) = 21/30$

For L_2

$P(L_2=1|W_1) = 25/30$ $P(L_2=2|W_1) = 2/30$ $P(L_2=3|W_1) = 3/30$

$P(L_2=1|W_2) = 3/30$ $P(L_2=2|W_2) = 22/30$ $P(L_2=3|W_2) = 5/30$

$P(L_2=1|W_3) = 5/30$ $P(L_2=2|W_3) = 6/30$ $P(L_2=3|W_3) = 19/30$

According to Bayes theorem

$$P(W_K|L_1,L_2) \propto P(L_1,L_2|W_K)P(W_K)$$

We can ignore the prior probability $P(W_K)$ since they are equal for $K=1, 2,$ and 3.

Hence

$$P(W_K|L_1,L_2) \propto P(L_1,L_2|W_K)$$

Since Naive Bayes assume independence of the classifiers

$$P(L_1,L_2|W_K) = P(L_1|W_K)P(L_2|W_K)$$

Therefore

$$P(W_K|L_1,L_2) \propto P(L_1|W_K)P(L_2|W_K)$$

So for a given L_1 and L_2, we assign class that maximize $P(W_k | L_1 L_2)$

For $L_1 = 1$ and $L_2 = 1$

$$P(W_1 | L_1, L_2) \propto P(L_1 = 1 | W_1) P(L_2 = 1 | W_1) = \frac{20}{30} \times \frac{25}{30} = \frac{5}{9}$$

$$P(W_2 | L_1 L_2) \propto P(L_1 = 1 | W_2) P(L_2 = 1 | W_2) = \frac{3}{30} \times \frac{3}{30} = \frac{1}{100}$$

$$P(W_3 | L_1 L_2) \propto P(L_1 = 1 | W_3) P(L_2 = 1 | W_3) = \frac{0}{30} \times \frac{5}{30} = 0$$

Evidence / Normalization factor

$$P(L_1 L_2) = P(L_1 L_2 | W_1) + P(L_1 L_2 | W_2) + P(L_1 L_2 | W_3)$$

$$= \frac{5}{9} + \frac{1}{100} + 0 = \underline{0.5656}$$

So

$$P(W_1 | L_1 L_2) = \frac{5}{9} \div 0.5656 = \underline{0.98}$$

$$P(W_2 | L_1 L_2) = \frac{1}{100} \div 0.5656 = \underline{0.02}$$

$$P(W_3 | L_1 L_2) = 0 \div 0.5656 = \underline{0.00}$$

Since $P(W_1 | L_1 = 1, L_2 = 1)$ has the highest probability, the fused label is W_1.

For $L_1 = 1$ $L_2 = 2$

$$P(W_1 | L_1 L_2) \propto P(L_1 = 1 | W_1) P(L_2 = 2 | W_1) = \frac{20}{30} \times \frac{2}{30} = \frac{40}{900}$$

$$P(W_2 | L_1 L_2) \propto P(L_1 = 1 | W_2) P(L_2 = 2 | W_2) = \frac{3}{30} \times \frac{22}{30} = \frac{66}{900}$$

$$P(W_3 | L_1 L_2) \propto P(L_1 = 1 | W_3) P(L_2 = 2 | W_3) = \frac{0}{30} \times \frac{6}{30} = 0$$

Evidence / Normalization factor

$$P(L_1 = 1 L_2 = 2) = P(L_1 = 1 L_2 = 2 | W_1) + P(L_1 = 1, L_2 = 2 | W_2) + P(L_1 = 1 L_2 = 2 | W_3)$$

$$= \frac{40}{900} + \frac{66}{900} + 0 = \frac{106}{900}$$

So

$$P(W_1 | L_1 = 1, L_2 = 2) = \frac{40}{900} \div \frac{106}{900} = 40/106 = 0.377$$

$P(W_2|L_1L_2) = \frac{66}{900} \div \frac{106}{900} = 0.623$

$P(W_3|L_1L_2) = 0 \div \frac{466}{900} = 0.00$

~~Since $P(W_3|L_1L_2) =$~~

Since $P(W_2|L_1=1, L_2=2)$ has the highest probability, the fused label is W_2.

For $L_1=1$, $L_2=3$

$P(W_1|L_1=1, L_2=3) \propto P(L_1=1|W_1)P(L_2=3|W_1) = \frac{20}{30} \times \frac{3}{30} = \frac{60}{900} = \frac{2}{3}$

$P(W_2|L_1=1, L_2=3) \propto P(L_1=1|W_2)P(L_2=3|W_2) = \frac{3}{30} \times \frac{5}{30} = \frac{15}{900}$

$P(W_3|L_1=1, L_2=3) \propto P(L_1=1|W_3)P(L_2=3|W_3) = \frac{0}{30} \times \frac{19}{30} = 0$

Normalization factor $= \frac{2}{3} + \frac{15}{900} + 0 = \frac{615}{900}$

$P(W_1|L_1=1, L_2=3) = \frac{2}{3} \div \frac{615}{900} = 0.8$

$P(W_2|L_1=1, L_2=3) = \frac{15}{900} \div \frac{615}{900} = 0.2$

$P(W_3|L_1=1, L_2=3) = 0$

Since $P(W_1|L_1=1, L_2=3)$ has the highest probability, hence the fused label is W_1

for $L_1=2$, $L_2=1$

$P(W_1|L_1=2, L_2=1) \propto P(L_1=2|W_1)P(L_2=1|W_1) = \frac{5}{30} \times \frac{25}{30} = \frac{125}{900}$

$P(W_2|L_1=2, L_2=1) \propto P(L_1=2|W_2)P(L_2=1|W_2) = \frac{24}{30} \times \frac{3}{30} = \frac{72}{900}$

$P(W_3|L_1=2, L_2=1) \propto P(L_1=2|W_3)P(L_2=1|W_3) = \frac{9}{30} \times \frac{5}{30} = \frac{45}{900}$

Normalization factor $= \frac{242}{900} = 0.269$

$P(W_1|L_1=2, L_2=1) = \frac{125}{900} \times \frac{900}{242} = \underline{0.52}$

$P(W_2|L_1=2, L_2=1) = \frac{72}{900} \times \frac{900}{242} = \underline{0.298}$

$P(W_3|L_1=2, L_2=1) = \frac{45}{900} \times \frac{900}{242} = 0.186$

$P(W_1|L_1=2, L_2=1)$ is the highest probability value.
Hence the fused label is W_1

For $L_1 = 2$ $L_2 = 2$

$P(W_1|L_1=2\ L_2=2) \propto P(L_1=2|W_1)P(L_2=2|W_1) = 5/30 \times 2/30 = 10/900$

$P(W_2|L_1=2\ L_2=2) \propto P(L_1=2|W_2)P(L_2=2|W_2) = 24/30 \times 22/30 = 528/900$

$P(W_3|L_1=2, L_2=2) \propto P(L_1=2|W_3)P(L_2=2|W_3) = 9/30 \times 6/30 = 54/900$

Normalization factor $= \frac{592}{900}$

$P(W_1|L_1=2, L_2=2) = \frac{10}{900} \times \frac{900}{592} = 0.017$

$P(W_2|L_1=2, L_2=2) = \frac{528}{900} \times \frac{900}{592} = 0.892$

$P(W_3|L_1=2, L_2=2) = \frac{54}{900} \times \frac{900}{592} = 0.091$

$P(W_2|L_1=2, L_2=2)$ has the highest value hence the fused label W_2

For $L_1 = 2$ $L_2 = 3$

$P(W_1|L_1=2\ L_2=3) \propto P(L_1=2|W_1)P(L_2=3|W_1) = 5/30 \times 3/30 = 15/900$

$P(W_2|L_1=2\ L_2=3) \propto P(L_1=2|W_2)P(L_2=3|W_2) = 24/30 \times 8/30 = 120/900$

$P(W_3|L_1=2\ L_2=3) \propto P(L_1=2|W_3)P(L_2=3|W_3) = 9/30 \times 19/30 = 171/900$

Normalization factor $= \frac{306}{900}$

$P(W_1|L_1=2\ L_2=3) = \frac{15}{900} \times \frac{900}{306} = \underline{0.049}$

$P(W_2|L_1=2\ L_2=3) = \frac{120}{900} \times \frac{900}{306} = \underline{0.392}$

$P(W_3|L_1\ L_2) = \frac{171}{900} \times \frac{900}{306} = \underline{0.560}$

$P(W_3|L_1=2, L_2=3)$ has the highest value hence the fused label is W_3.

For $L_1 = 3, L_2 = 1$

$P(W_1 | L_1 = 3, L_2 = 1) \propto P(L_1 = 3 | W_1) P(L_2 = 1 | W_1) = 5/30 \times 25/30 = \frac{125}{900}$

$P(W_2 | L_1 = 3, L_2 = 1) \propto P(L_1 = 3 | W_2) P(L_2 = 1 | W_2) = 3/30 \times 3/30 = \frac{9}{900}$

$P(W_3 | L_1 = 3, L_2 = 1) \propto P(L_1 = 3 | W_3) P(L_2 = 1 | W_3) = 21/30 \times 25/30 = \frac{105}{900}$

Normalization factor $= \frac{239}{900}$

$P(W_1 | L_1 = 3\ L_2 = 1) = \frac{125}{900} \times \frac{900}{239} = 0.523$

$P(W_2 | L_1 = 3\ L_2 = 1) = \frac{9}{900} \times \frac{900}{239} = 0.038$

$P(W_3 | L_1 = 3\ L_2 = 1) = \frac{105}{900} \times \frac{900}{239} = 0.439$

$P(W_1 | L_1 = 3, L_2 = 1)$ has the highest value hence fused label is W_1

For $L_1 = 3, L_2 = 2$

$P(W_1 | L_1 = 3\ L_2 = 2) \propto P(L_1 = 3 | W_1) P(L_2 = 2 | W_1) = 5/30 \times 2/30 = \frac{10}{900}$

$P(W_2 | L_1 = 3\ L_2 = 2) \propto P(L_1 = 3 | W_2) P(L_2 = 2 | W_2) = 3/30 \times 22/30 = \frac{66}{900}$

$P(W_3 | L_1 = 3\ L_2 = 2) \propto P(L_1 = 3 | W_3) P(L_2 = 2 | W_3) = 21/30 \times 6/30 = \frac{126}{900}$

Normalization factor $= \frac{202}{900}$

$P(W_1 | L_1 = 3, L_2 = 2) = \frac{10}{900} \times \frac{900}{202} = 0.0495$

$P(W_2 | L_1 = 3\ L_2 = 2) = \frac{66}{900} \times \frac{900}{202} = 0.3267$

$P(W_3 | L_1 = 3\ L_2 = 2) = \frac{126}{900} \times \frac{900}{202} = 0.6237$

$P(W_3 | L_1 = 3, L_2 = 2)$ has the highest value hence the fused label is W_3

For $L_1 = 3\ L_2 = 3$

$P(W_1 | L_1 = 3, L_2 = 3) \propto P(L_1 | W_1) P(L_2 | W_2) = 5/30 \times 3/30 = \frac{15}{900}$

$P(W_2 | L_1 = 3, L_2 = 3) \propto P(L_1 = 3 | W_2) P(L_2 = 3 | W_2) = 3/30 \times 5/30 = \frac{15}{900}$

$P(W_3 | L_1 = 3, L_2 = 3) \propto P(L_1 = 3 | W_3) P(L_2 = 3 | W_3) = 21/30 \times 19/30 = \frac{399}{900}$

normalization factor = $\frac{429}{100}$

$P(W_3 | L_1=3, L_2=3) = \frac{319}{100} \times \frac{956}{429} = 0.93$ is the highest probability value hence fused label is W_3

L_1 L_2	Fused Label
1, 1	W_1
1, 2	W_2
1, 3	W_1
2, 1	W_1
2, 2	W_2
2, 3	W_3
3, 1	W_1
3, 2	W_3
3, 3	W_3

(b) BKS table requires joint distribution of classifiers outputs and true class labels from the training data - the confusion matrices do not provide these joint distribution information

Additional information needed include instances labeled by both classifiers to directly populate the BKS table.

Difference between Naive Bayes and BKS is that:

- Naive Bayes assumes classifier independence and combine probabilities analytically while
- BKS directly uses observed joint distribution of classifier output without independence assumption.

(c) yes, Given a BKS table, I can derive the Naive Bayes look up table first by assuming or calculating the ~~marginal~~ prior $P(W_k)$ and calculating the marginal probabilities $P(L_1, L_2 | W_k)$ by summing over the observed counts in the BKS table.

PROBLEM 2 (a) Linear Regression Implementation

sqft(x)	rent(y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
750	1160	-100	-258	25800	10000	16384
800	1200	-50	-218	10900	2500	625
850	1280	0	-138	0	0	19044
900	1450	50	32	1600	2500	25921
950	2000	100	582	58200	10000	38416
4250	7090			96500	25000	100350

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{4250}{5} = 850$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{7090}{5} = 1418$$

The regression line is

$$\hat{y} = \beta_0 + \hat{\beta}_1 x$$

Where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{96500}{25000} = 3.86$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 1418 - 3.86(850) = -1863$$

$$\boxed{\hat{y} = -1863 + 3.86x}$$

1928.50
15.21
14476.50
2467.108

$$\hat{y}_1 = -1863 + 3.86(750) = 1032$$
$$\hat{y}_2 = -1863 + 3.86(800) = 1225$$
$$\hat{y}_3 = -1863 + 3.86(850) = 1418$$
$$\hat{y}_4 = -1863 + 3.86(900) = 1611$$
$$\hat{y}_5 = -1863 + 3.86(950) = 1804$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n} = \frac{100390}{5} = \underline{20078}$$

L1 Boost Implementation

Initial prediction

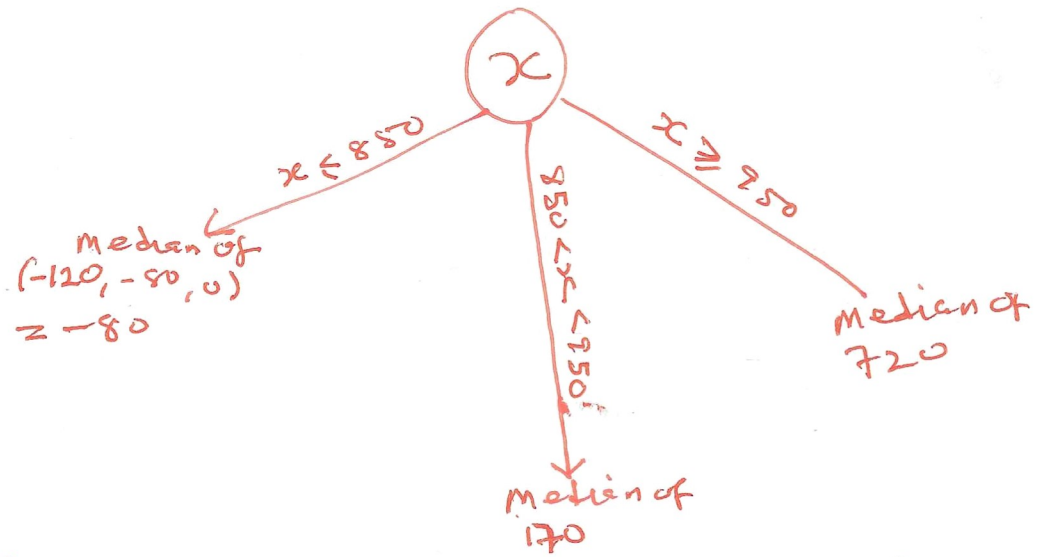
$F_0(x)$ = median of all y values = 1280

residual $r_i = y_i - F_0(x_i)$

the initial prediction and residual are tabulated below

x	y	$F_0(x)$	r_i
750	1160	1280	-120
800	1200	1280	-80
850	1280	1280	0
900	1450	1280	170
950	2000	1280	720

I Will train decision tree as the weak learner for predicting the residual, to train the weak learner, I split the feature as follow:



I will fix the learning rate of 0.8

So, for $x \leq 850$, residual is -80

for all $850 < x < 950$, residual value is 170

for all $x \geq 950$, residual value is 720

$F_1(x) = F_0(x) + \eta f(x)$ where η is learning rate and $f(x)$ is outcome of the weak learner

Table for the next iteration is below

x	y	$F_0(x)$	$f(x)$	update	$F_1(x)$	r_i
750	1160	1280	-80	-64	1216	-56
800	1200	1280	-80	-64	1216	-16
850	1280	1280	-80	-64	1216	64
900	1450	1280	170	136	1416	34
950	2000	1280	720	576	1856	144

Next iteration

I will now train a new weak learner. The partition rule for the new weak learner is

for $x \leq 800$, residual = median of $(-56, -16) = -36$

for $800 < x \leq 900$, residual = median of $(34, 64) = 49$

for $x > 900$, residual is 144

$F_2(x) = F_1(x) + \eta f(x)$

x	y	$F_1(x)$	$f(x)$	update	$F_2(x)$	r_i
750	1160	1216	-36	-28.8	1187.2	-27.2
800	1200	1216	-36	-28.8	1187.2	12.8
850	1280	1216	49	39.2	1255.2	24.8
900	1450	1416	49	39.2	1455.2	-5.2
950	2000	1856	144	115.2	1971.2	28.8

Next iteration

New weak learner

for $x \leq 750$ residual = median of $-27.2 = -27.5$

for $750 < x \leq 900$, residual = median of $-5.2, 12.8, 24.8$
$= 12.8$

for $x > 900$, residual $= 28.8$

$F_3(x) = F_2(x) + \eta f(x)$

x	y	$F_2(x)$	$f_2(x)$	update	$F_3(x)$	γ_i
750	1160	1187.2	-27.8	-22	1165.2	-5.2
800	1200	1187.2	12.8	10.24	1197.44	2.56
850	1280	1258.2	12.8	10.24	1265.44	14.56
900	1480	1455.2	12.8	10.24	1465.44	-15.44
950	2000	1971.2	28.8	23.04	1994.24	5.76

I will stop the iteration at this point and asses model performance using MSE

$$MSE = \frac{\sum_{i=1}^{5}(y_i - \hat{y}_i)^2}{n} = \frac{(-5.2)^2 + (2.56)^2 + (14.56)^2 + (-15.44)^2 + (5.76)^2}{5}$$

$$= \frac{27.04 + 6.5536 + 211.9536 + 238.3936 + 33.1776}{5}$$

$$= \frac{517.8504}{5} = 103.43168$$

L2 Boost Implementation

Initial Prediction

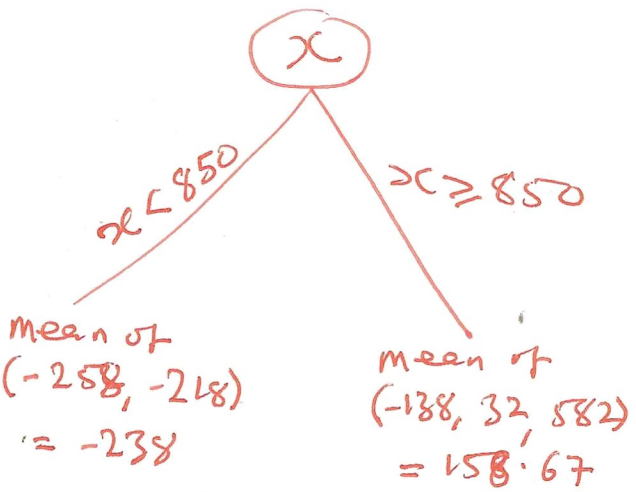
$F_0(x) = \bar{y} = \frac{1160 + 1200 + 1280 + 1450 + 2000}{5} = 1418$

$r_i = y_i - F_0(x_i)$

$F_0(x_i)$ and r_i are tabulated below

x	y	$F_0(x)$	r_i
750	1160	1418	-258
800	1200	1418	-218
850	1280	1418	-138
900	1450	1418	32
950	2000	1418	582

The weak learner for this iteration

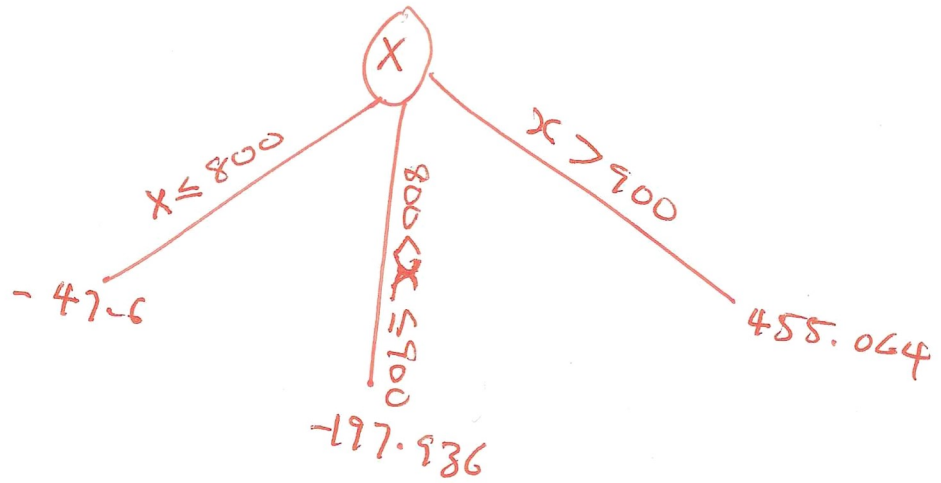


Prediction update

$F_1(x) = F_0(x) + \eta f(x) \quad \eta = 0.8$

x	y	$F_0(x)$	$f(x)$	update	$F_1(x)$	r_i
750	1160	1418	-238	-190.4	1227.6	-67.6
800	1200	1418	-238	-190.4	1227.6	-27.6
850	1280	1418	158.67	126.936	1544.936	-264.93
900	1450	1418	158.67	126.936	1544.936	-94.936
950	2000	1418	158.67	126.936	1544.936	455.064

weak learner for the next iteration

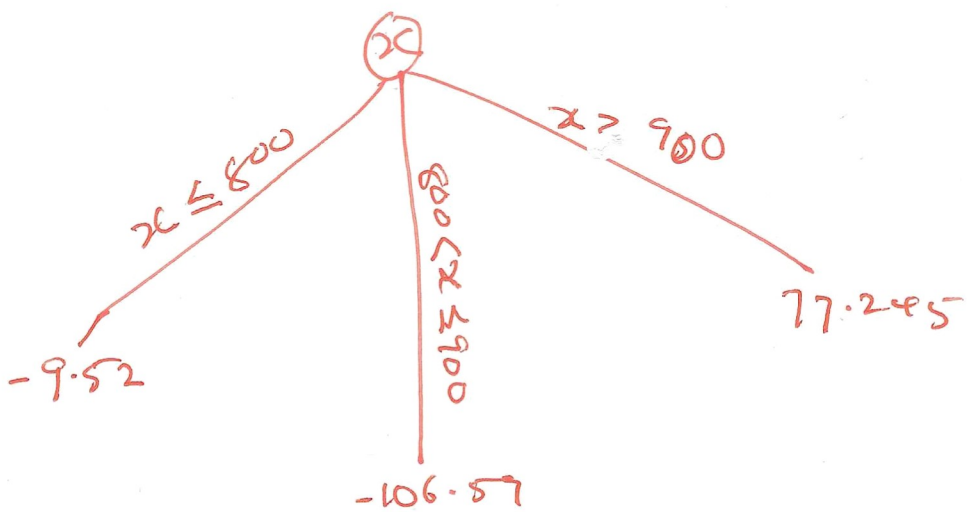


$F_2(x) = F_1(x) + \eta f(x)$

x	y	$F_1(x)$	f(x)	update	$F_2(x)$	r_i
750	1160	1227.6	-47.6	-38.08	1189.52	-29.52
800	1200	1227.6	-47.6	-38.08	1189.52	10.48
850	1280	1544.936	-197.936	-158.3488	1386.59	-106.59
900	1450	1544.936	-197.936	-158.3488	1386.59	63.41
950	2000	1544.936	455.064	364.0512	1908.99	91.08

Next Iteration

Weak learner for this iteration



$F_3(x) = F_2(x) + \eta f(x)$

x	y	$F_2(x)$	$f(x)$	update	$F_3(x)$	r_i
750	1160	1189.52	-9.52	-7.616	1181.904	-21.904
800	1200	1189.52	-9.52	-7.616	1181.904	18.096
850	1250	1386.59	-106.59	-85.272	1301.318	-21.318
900	1450	1386.59	77.245	61.796	1448.386	1.614
950	2000	1908.99	77.245	61.796	1970.788	29.212

I will stop here with the iteration and calculate the MSE

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} = \frac{(-21.904)^2 + (18.096)^2 + (-21.318)^2 + (1.614)^2 + (29.212)^2}{5}$$

$$= \frac{479.78 + 327.465 + 454.46 + 2.605 + 853.340}{5}$$

$$= 423.530$$

Comparison:

MSE for Linear Regression is 20,078, MSE for L1 Boost is 103.4318, and MSE for L2 Boost is 423.530. Hence, L1 Boost and L2 Boosy have better performance than linear regression

(b)

It is called gradient boosting because it uses the Principle of gradient descent to optimize a loss function during the boosting process. At each step of the algorithm, the model learns by minimizing either the mean absolute error (for L_1 Boost) and mean squared error (for L_2 Boost)

Boosting is an ensemble technique that combine Multiple weak learners to create a strong learner.

(c)

The Linear regression problem in Homework 4 Problem 3 involves predicting the probability of a student passing a course based on the number of hours the student reads. the data is in the table below

Hrs (x_k)	1.0	2.0	3.0	4.0	5.0
Prob (P_k)	0.07	0.26	0.61	0.87	0.97

L2 Boost Implementation

Initial prediction

$F_0(x)$ = mean of all target values

$$f_0(x) = \frac{\sum_{k=1}^{n} P_k}{n} = \frac{0.07+0.26+0.61+0.87+0.97}{5} = 0.556$$

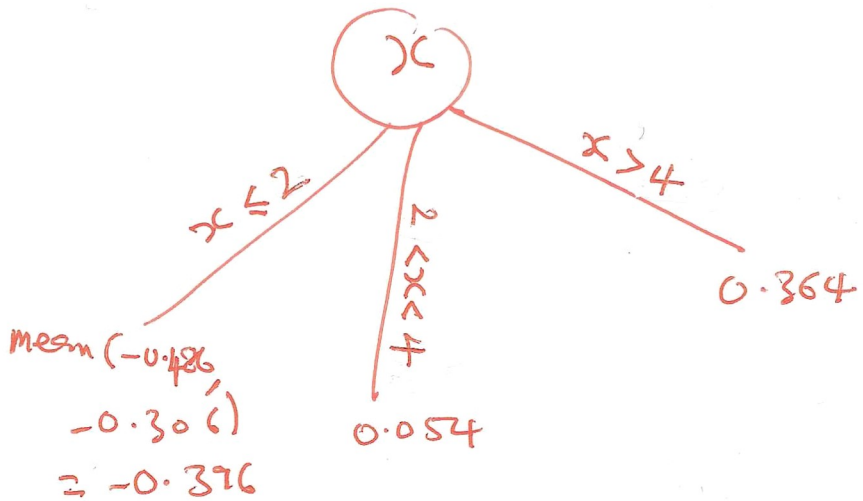
residual $r_i = P_k - F_0(x_k)$

the $F_0(x_k)$ and r_i's are tabulated below

For convenience let $y_i = P_k$ and $x_i = x_k$

x	y	$F_0(x)$	Y_i
1	0.07	0.556	-0.486
2	0.26	0.556	-0.306
3	0.61	0.556	0.054
4	0.87	0.556	0.314
5	0.97	0.556	0.414

I will train the weak learner (decision tree) as follow



for $x \leq 0$, residual is -0.396
for $0 < x < 0.1$ residual is 0.054
and for $x > 0.1$ residual is 0.364

Prediction update

$F_1(x) = F_0(x) + \eta f(x)$ learning rate $\eta = 0.8$

Next iteration table

x	y	$F_0(x)$	$f(x)$	update	$F_1(x)$	Y_i
1	0.07	0.556	-0.396	-0.3168	0.2392	-0.1692
2	0.26	0.556	-0.396	-0.3168	0.2392	0.0208
3	0.61	0.556	0.054	0.0432	0.5992	0.0108
4	0.87	0.556	0.364	0.2912	0.8472	0.0228
5	0.97	0.556	0.364	0.2912	0.8472	0.1228

Next iteration.

I will train the next weak learner (decision tree)

the partition rule will be:

for $x \leq 1$, residual = -0.1692

for $1 < x \leq 4$ residual = $\frac{0.0208 + 0.0108 + 0.0228}{3} = \underline{\underline{0.018}}$

for $x > 4$ residual $= 0.1228$

$F_2(x) = F_1(x) + \eta f(x)$ $\quad \eta = 0.8$

x	y	$F_1(x)$	$f(x)$	update	$F_2(x)$	r_i
1	0.07	0.2352	-0.1692	-0.13536	0.10384	-0.03384
2	0.26	0.2352	0.018	0.0144	0.2536	0.0064
3	0.61	0.5952	0.018	0.0144	0.6136	-0.0036
4	0.87	0.8472	0.018	0.0144	0.8616	0.0084
5	0.97	0.8472	0.1228	0.09824	0.94544	0.02456

I will stop the training here and calculate the MSE (the performance metric for the model)

$$MSE = \frac{\sum_{i=1}^{n} r_i^2}{n} = (-0.03384)^2 + 0.0064^2 + (-0.0036)^2 + (0.0084)^2 + (0.02456)^2$$

$$= 0.001145 + 0.00004096 + 0.00001296 + 0.000603$$

$$= \underline{0.0018}$$

To compare this performance with performance of the simple linear regression in Home work 4, I will bring in the regression line equation here and calculate the MSE. In home work 4, the regression equation is $\hat{y} = -0.167 + 0.241x$, $\hat{y}$ values are tabulated below

x	y	$\hat{y}$	$(y-\hat{y})^2$
1	0.07		1.6×10^{-5}
2	0.26		0.003025
3	0.61		0.002916
4	0.87		0.005329
5	0.97		0.004624

for $x=1$
$\hat{y} = -0.167 + 0.241(1) = 0.074$
for $x=2$
$\hat{y} = -0.167 + 0.241(2) = 0.315$
and so on

$$MSE = \frac{\sum_i (y-\hat{y})^2}{n} = \frac{0.020534}{5} = 0.004$$

MSE for L2 Boost is 0.0018 while MSE for linear regression is 0.004 hence L2 Boost has better performance with very few iterations

PROBLEM 3

Gini Impurity is defined as:

$$Gini = 1 - \sum_{i=1}^{K} P_i^2$$

where P_i is the proportion of samples belonging to class i (Prior probability)

$$P_1 = \frac{90}{100} = 0.1$$

$$P_2 = \frac{10}{100} = 0.1$$

$$Gini_N = 1 - [0.9^2 + 0.1^2] = 1 - 0.82 = 0.18$$

Gini Impurity for each split option

Option 1

left node

$$P_1 = \frac{70}{70} = 1.0 \qquad P_2 = \frac{0}{70} = 0.0$$

$$Gini_{left} = 1 - (1.0^2 + 0.0^2) = 0$$

Right node

$$P_1 = \frac{20}{30} = 0.67 \qquad P_2 = \frac{10}{30} = 0.33$$

$$Gini_{right} = 1 - (0.67^2 + 0.33^2) = 0.4422$$

Weighted Gini (Gini split)

$$Gini_{split} = \frac{70}{100} \times Gini_{left} + \frac{30}{100} \times Gini_{Right}$$

$$= 0.70 \times 0.0 + 0.3 \times 0.4422 = 0.1327$$

(b)

Option 2

80 samples belong to class 1

$$P_1 = \frac{80}{80} = 1.0 \qquad P_2 = \frac{0}{80} = 0.0$$

$Gini_{left} = 0.0$

Right node

$P_1 = \frac{10}{20} = 0.5 \quad P_2 = \frac{10}{20} = 0.5$

$Gini_{right} = 1 - (0.5^2 + 0.5^2) = 0.5$

Weighted Gini

$Gini_{split} = \frac{80}{100} \times Gini_{left} + \frac{20}{100} \times Gini_{Right}$

$= 0.8 \times 0 + 0.2 \times 0.5 = 0.1$

option 1: Gini Impurity = 0.1327

option 2: Gini Impurity = 0.1

Hence,

option 2 is a better split because it has lower Gini impurity.

Bonus

In KNN, the prior probability is fixed at n_k/n where
n_k = number of samples in class k
n = Total number of samples
potential modifications to KNN to vary the prior is as follow:

Step 1:
assign a weight to each class based on the desired prior probability $P(C_k)$ so that ~~voted~~ weighted vote for class k becomes

$$\sum_{i \in \text{class } k} w_i$$

Where

$$w_i = \frac{P(C_k)}{P(C_k^{\text{empirical}})}$$

Step 2
We modify the influence of neighbours based on both distance and prior

$$w_i = \frac{P(C_k)}{\text{Distance}(x, x_i)^p}$$

Where p controls the effect of the distance

Step 3
we modify the nearest neighbour search to account for the priors and re-weight by class frequencies so that instead of voting, we compute the posterior probability of each class using

Bonus

In KNN, the prior probability is fixed at n_k/n where
n_k = number of samples in class k
n = Total number of samples
potential modifications to KNN to vary the prior is as follow:

Step 1:
assign a weight to each class based on the desired prior probability $P(C_k)$ so that voted weighted vote for class k becomes

$$\sum_{i \in \text{class}k} w_i$$

Where

$$w_i = \frac{P(C_k)}{P(C_k^{\text{empirical}})}$$

Step 2
We modify the influence of neighbours based on both distance and prior

$$w_i = \frac{P(C_k)}{\text{Distance}(x, x_i)^p}$$

Where p controls the effect of the distance

Step 3
we modify the nearest neighbour search to account for the priors and re-weight by class frequencies so that instead of voting, we compute the posterior probability of each class using

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_j P(x|C_j)P(C_j)}$$

Where

$P(x|C_k)$ is estimated from kNN

with this procedure, the prior ($P(C_k)$) can always be varied.