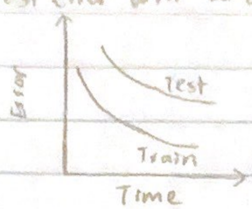
Ethan Wang
305319001

CS M148 HW 2

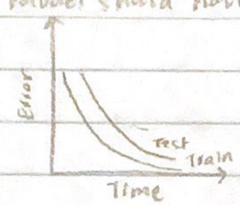
1\. a) The leftmost classifier has large bias and underfitting. In this data, it makes more sense to have a non-linear classifier. Both the training and test error will be decently high even if allowed to train over a long time.



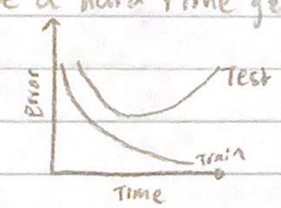
(Bias is inability for model to capture true relationship)

(Variance is difference in fits between data sets) - poor generalization

The middle classifier fits the data well without underfitting or overfitting. Thus, neither the bias nor variance should be that high. This model should have low error rates on both training and testing data.



The rightmost classifier has low bias but high variance, as it overfits the data. The model performs perfectly on training data but will have a hard time generalizing, and perform poorly on test data.



L1: Lasso

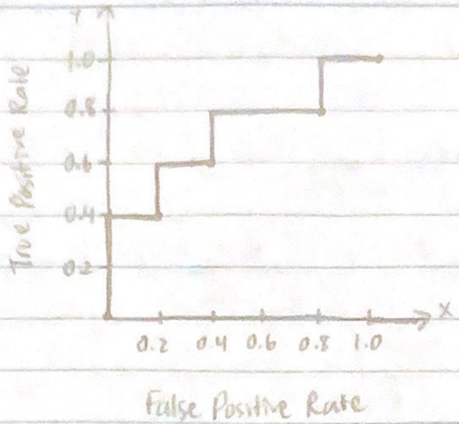
L2: Ridge

b) It seems like figure A is using L1 regression as the original black line gets flattened. This is indicative of L1 regression since L1 regression shrinks the coefficient of the least important feature to 0, effectively reducing the complexity of the model. We see that the black line is a straight split through the middle, which is significantly less complex. Also, in L1 regression the optimal slope will reach 0 (which we see on figure A), but in L2 regression the optimal slope should not reach 0.

2

1) Rank predicted probability
Set different threshold.
determine the TPR and FPR for that given threshold - then plot it
(ie 10 thresholds)

2. a)



0	1	FP	TP	TN
0	0	TN	TP	TN
1	1	TP	TP	TN
1	0	FN	TP	TN
			FN	FP

b) AUC = (0.2 × 0.4) + (0.2 × 0.6) + (0.4 × 0.8) + (0.2 × 1.0)

AUC = 0.08 + 0.12 + 0.32 + 0.2 = 0.72 $\boxed{AUC = 0.72}$

c)

	Predicted = 0	Predicted = 1
Actual = 0	TN: 1	FN: 1
Actual = 1	FP: 4	TP: 4

Threshold = 0.5

1	1	TP
1	1	TP
1	1	TP
1	1	TP
1	0	FN
0	1	FP
0	1	FP
0	1	FP
0	1	FP
0	0	TN

d) Accuracy : $\frac{5}{10}$ = 0.5 $\frac{TP+TN}{TP+FP+FN+TN}$

Precision : $\frac{4}{8}$ = 0.5 $\frac{TP}{TP+FP}$

Recall : $\frac{4}{5}$ = 0.8 $\frac{TP}{TP+FN}$

F1 Score : $\frac{2 \times (0.8 \times 0.5)}{(0.8 + 0.5)}$ = 0.615 $\frac{2(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}$

e) We could change the threshold to 0.36

	Predicted = 0	Predicted = 1
Actual = 0	TN = 1	FN = 0
Actual = 1	FP = 4	TP = 5

Threshold 0.36

1	1	TP
1	1	TP
1	1	TP
1	1	TP
1	1	TP
0	1	FP
0	1	FP
0	1	FP
0	1	FP
0	0	TN

Accuracy : $\frac{6}{10}$ = 0.6

Precision : $\frac{5}{9}$ = 0.55

Recall : $\frac{5}{5}$ = 1.0

F1 Score : $\frac{2(0.55 \times 1)}{0.55 + 1}$ = 0.719

We could change the threshold to 0.36 as it yields better results across all four statistics: accuracy, precision, recall, and F1 Score.

3

3. $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

a) $p=2$ $\beta_0 = 3$ $\beta_1 = 5$ $\beta_2 = -8$ $X_1 = 0$ $X_2 = 0$

$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 3$

$\frac{p}{1-p} = e^3$

$p = 0.953$ $\boxed{P(Y=1) = 0.953}$

Odds $= \frac{0.953}{1-0.953}$

$\boxed{\text{Odds} = 20.277}$

b) $X_1 = 2$

$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 3 + (2 \times 5) = 13$

$\frac{p}{1-p} = e^{13}$

$p = 0.99999\ldots \approx 1$

log odds change to 13 (multiplicative change by 10)

odds change by $e^{10} = 22026.466$

$P(Y=1) = 0.99999\ldots \approx 1$

$X_2 = -2$

$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 3 + (-2 \times -8)$

$\frac{p}{1-p} = e^{19}$

$p = 0.99999 \approx 1$

log odds change to 19 (multiplicative change by 16)

odds change by $e^{16} = 8886110.521$

$P(Y=1) = 0.99999\ldots \approx 1$

c) Increasing or decreasing β_0, β_1, or β_2 affects our predictions because they are all coefficients that directly influence parts of our logistic regression function. More specifically, they control how steep our curve is and how we shift it around. Increasing any of these coefficients will increase the odds that $P(y=1)$, and decreasing any of the coefficients will decrease the odds that $P(y=1)$.

?? d) Decision Boundary: $3 + 10X_1 + 16X_2 = 0$

Points: $(X_1, \frac{3}{16} + \frac{5}{8}X_1)$

e) The potential reason for this is that the model senses the absence of the other feature and tries to compensate by adjusting the coefficients of the other features. This could lead to issues as getting rid of either X_1 or X_2 can reduce the complexity of the model that it requires in order to fit the data properly. This may result in underfitting and therefore poorer model performance.

y: weight

x: Frequency

4. a) The intercept coefficient means that a 23-year-old mother will have a correlation of -0.48 between her age, the baby's birth weight, and the frequency of physician visits.

- $\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = -0.48 + 0.06\,Age - 0.45\,Frequency - 0.19\,Age\_Frequency$
- infrequent: $\ln\left(\frac{P(Y=1|X=0)}{P(Y=0|X=0)}\right) = -0.48 + 0.06\,Age$
- frequent: $\ln\left(\frac{P(Y=1|X=1)}{P(Y=0|X=1)}\right) = -0.48 + 0.06\,Age - 0.45 - 0.18\,Age \rightarrow -0.93 - 0.12\,Age$

For frequent physician visits, the odds ratio is $e^{-0.12} = 0.89$ for a one unit increase in age.

For infrequent physician visits, the odds ratio is $e^{0.06} = 1.06$ for a one unit increase in age.

The ratio of the two odds ratios above $\left(\frac{\text{frequent}}{\text{infrequent}}\right)$ is $e^{-0.19}$, which comes from the interaction term Age x Frequency.

b) Frequent: $\ln\left(\frac{P(Y=1|X=1)}{P(Y=0|X=1)}\right) = -0.93 - 0.12\,Age$

$e^{-0.12} = 0.89$

- A one unit increase in age results in 0.89 times the odds of a low weight baby.

Infrequent: $\ln\left(\frac{P(Y=1|X=0)}{P(Y=0|X=0)}\right) = -0.48 + 0.06\,Age$

$e^{0.06} = 1.06$

- A one unit increase in age results in 1.06 times the odds of a low weight baby.

We see that age is a variable in the model - adjusting age will also affect the odds in both situations (Frequent vs. Infrequent physician visits).

c)

Age	Odds Ratio	95% Confidence Interval
18	1.568	(0.705, 4.949)
23	0.638	(0.325, 1.201)
25	0.445	(0.262, 1.036)
28	0.259	(0.206, 0.916)
30	0.181	(0.050, 0.607)

$$\frac{e^{-0.93 - 0.12\,Age}}{e^{-0.48 + 0.06\,Age}}$$

d) The Odds Ratio shows the odds of an outcome occurring from a specific factor (such as age) against the odds of an outcome occurring without that factor. If the odds ratio is greater than 1, the factor increases the odds of that outcome. If the odds ratio is less than 1, the factor decreases the odds of that outcome. If the odds ratio is 1, then the factor has no effect on the odds of the outcome. Based on the table from part C, we can see that it's more likely for frequent physician visits to increase the odds of low baby weight at younger ages (18). As the mother's age increases, the Odds Ratio drops below 1. This means that frequent physician visits will likely decrease the odds of low baby weight in older mothers. All of the calculated odds intervals fall into the confidence interval.

e)

Age	Difference in Probability	95% Confidence Interval
18	0.104	(-0.788, 0.393)
23	-0.099	(-0.197, 0.088)
25	-0.174	(-0.232, 0.046)
28	-0.277	(-0.315, -0.016)
30	-0.339	(-0.540, -0.092)

Difference in probability: $\underbrace{P(Y=1 \mid X=1)}^{\text{Frequent}} - \underbrace{P(Y=1 \mid X=0)}^{\text{Infrequent}}$

Infrequent

$$\ln\left(\frac{P(Y=1 \mid X=1)}{P(Y=0 \mid X=1)}\right) = -0.48 + 0.06\,\text{Age}$$

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{-0.48+0.06\,\text{Age}}$$

$$P(Y=1) = e^{-0.48+0.06\,\text{Age}} - P(Y=1)e$$

$$P(Y=1 \mid X=0) = \frac{e^{-0.48+0.06\,\text{Age}}}{1+e^{-0.48+0.06\,\text{Age}}}$$

Frequent

$$\ln\left(\frac{P(Y=1 \mid X=1)}{P(Y=0 \mid X=1)}\right) = -0.93 - 0.12\,\text{Age}$$

$$P(Y=1 \mid X=1) = \frac{e^{-0.93-0.12\,\text{Age}}}{1+e^{-0.93-0.12\,\text{Age}}}$$

The difference in probability shows the probability of frequent physician visits affecting low weight at different ages versus the probability of infrequent physician visits affecting low weight at different ages. The results show that at age 18, the effect of frequent physician visits has a more significant effect on the odds (because of the positive probability). The baby is actually more likely to have a low weight. At the older ages, frequent physician visits seem to decrease the odds of low baby weight. The older the mother is, the more significant this decrease becomes.

It's worth noting that all of the odds ratios and the probability differences fall within the 95% confidence interval. This means that the reduction in probability of a low-weight baby with frequent physician visits is statistically significant.