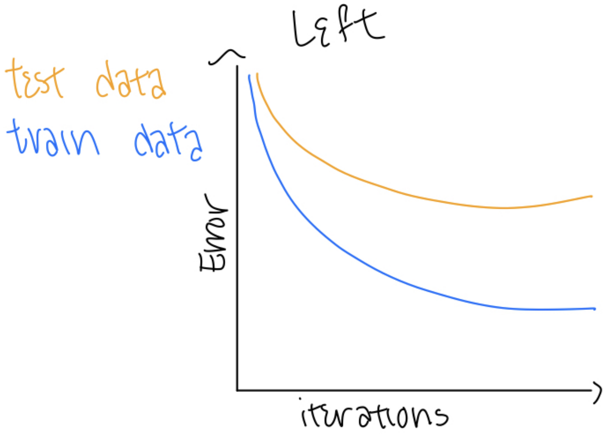
CS M148 Homework 2

Hanna Co

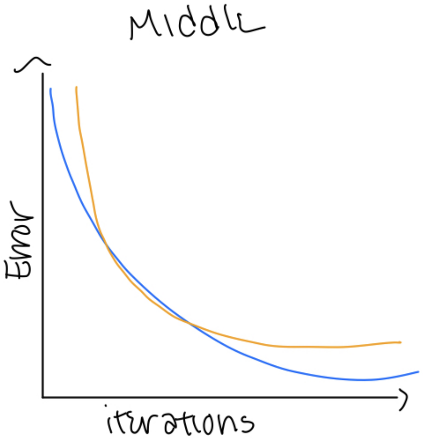
Due: February 16, 2021

1 Bias, Variance and Regularization

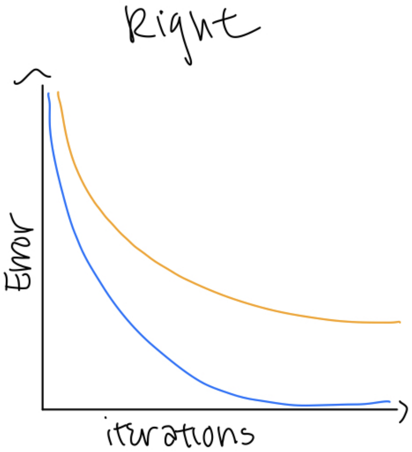
a) The graph on the far left has larger bias, and the graph on the far right has larger variance.



The graph on the left will perform bad on both the train and test data, because it's too generalized.



The graph in the middle will perfrom alright on both train and test data. This is because it's well fitted to the training data, but not overfitted to the point that it performs badly on test data.



The graph on the right will perform really good on the training data, but this is because it is overfitted to the training data. Thus, it will not perform will on the test data.

b) L1 regularization is used on a), because the coefficients are nullified fast, compared to b), where they are not nullified.

2 Maximum Likelihood View of Linear Regression

a) Since the likelihood function represents the probability of producing a particular sample, thus the equation is:
$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-1}{2}(\frac{y_i - \hat{y_i}}{\sigma})^2)$

b) We take the equation in 2a and change it into the log likelihood equation:
$\sum_{i=1}^{n} ln(\frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i-\hat{y_i})^2}{2\sigma^2}))$
$\sum_{i=1}^{n} ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(y_i-\hat{y_i})^2}{2\sigma^2}$
$\sum_{i=1}^{n} -\frac{1}{2} ln(2\pi\sigma^2) - \frac{(y_i-\hat{y_i})^2}{2\sigma^2}$
$-\frac{n}{2} ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i-\hat{y_i})^2}{2\sigma^2}$

We want to find $argmax(L)$:
$argmax(L) = argmax(-\sum_{i=1}^{n} \frac{(y_i-\hat{y_i})^2}{2\sigma^2})$
Since we will end up setting $argmax(L) = 0$, we can simply write
$argmax(L) = argmax(-\sum_{i=1}^{n}(y_i - \hat{y_i})^2)$
Since $argmax(L) = argmin(-L)$,
$argmax(L) = argmin(\sum_{i=1}^{n}(y_i - \hat{y_i})^2)$
Additionally, we end up taking the mean, so we have
$argmax(L) = argmin(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2)$
Notice that the right hand side of the above equation is the equation for MSE. Thus, maximizing log likelihood is equivalent to maximizing MSE.

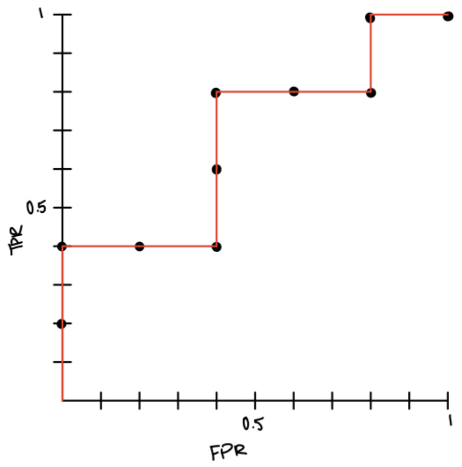
3 Classification Metric

a) ROC Curve

TPR	FPR	threshold
0.2	0	0.98
0.4	0	0.92
0.4	0.2	0.83
0.4	0.4	0.77
0.6	0.4	0.62
0.8	0.4	0.59
0.8	0.6	0.55
0.8	0.8	0.52
1	0.8	0.32
1	1	0.13

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$



b) $(0.4)(1) + (0.4)(0.6) + (0.2)(0.2) = 0.68$

c) Confusion Matrix

		True	
		Pos	Neg
Predicted	Pos	4	4
	Neg	1	1

d) Accuracy $= \frac{TP+TN}{TP+FP+TN+FN} = \frac{4+1}{4+1+4+1} = 0.5$
Precision $= \frac{TP}{TP+FP} = \frac{4}{4+1} = 0.5$
Recall $= \frac{TP}{TP+FN} = \frac{4}{4+1} = 0.8$
F1 Score $= 2 * \frac{precision*recall}{precision+recall} = 2 * \frac{0.8*0.5}{0.8+0.5} = 0.615$

e) No, changing the threshold to improve one score would negatively affect another score. This is because increasing the number of true positive would increase the number of false positive. Conversely, decreasing the number of false positives would decrease the number of true positives.

4 4 K-Nearest Neighbors for Classification

a) The numbers are classes.

b) For the most part, yes,though there are a few classes where it doesn't perform well. For example, class=12 and class=39 do not perform well in terms of the reported metrics.

c) Yes, the results would be different if the lighting or angles of the faces varied more. This gives the model more variety to train on. The results could change with a different background – if the contrast was similar, the results would probably be very similar. However, if it was changed to a low contrast background, it would likely make the model worse, because it becomes harder to differentiate between the face and the background.

5 Logistic Regression

a) We have the equation $ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.
Since $X_1 = X_2 = 0$, this equation becomes $ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0$.
$ln(\frac{P(Y=1)}{1-P(Y=1)}) = 3$.
$\frac{P(Y=1)}{1-P(Y=1)} = e^3$
$P(Y=1) = (1 - P(Y=1))e^3$
$P(Y=1) = e^3 - e^3 P(Y=1)$
$P(Y=1) + e^3 P(Y=1) = e^3$
$P(Y=1) * (1 + e^3) = e^3$
$P(Y=1) = \frac{e^3}{1+e^3}$
$P(Y=1) = 0.953$
$\frac{P(Y=1)}{1-P(Y=1)} = 20.086$
The probability of the event Y=1 is 0.953, and the odds are 20.086.

b) A one unit increase in X_1 changes our probability equation to $ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X_1$.
$ln(\frac{P(Y=1)}{1-P(Y=1)}) = 3 + 2$
$\frac{P(Y=1)}{1-P(Y=1)} = e^5$
$P(Y=1) = (1 - P(Y=1))e^5$
$P(Y=1) = e^5 - e^5 P(Y=1)$
$P(Y=1) + e^5 P(Y=1) = e^5$
$P(Y=1) * (1 + e^5) = e^5$
$P(Y=1) = \frac{e^5}{1+e^5}$
$P(Y=1) = 0.993$
A one unit increase in X_2 changes our probability equation to $ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_2 X_2$.
$ln(\frac{P(Y=1)}{1-P(Y=1)}) = 3 - 5$
$\frac{P(Y=1)}{1-P(Y=1)} = e^{-2}$
$P(Y=1) = (1 - P(Y=1))e^{-2}$
$P(Y=1) = e^{-2} - e^{-2} P(Y=1)$
$P(Y=1) + e^{-2} P(Y=1) = e^{-2}$
$P(Y=1) * (1 + e^{-2}) = e^{-2}$
$P(Y=1) = \frac{e^{-2}}{1+e^{-2}}$

$P(Y = 1) = 0.119$
A one unit increase in X_1 increases the odds and log odds of the event that Y = 1, while a one unit increase in X_2 decreases the odds and log odds of the event that Y = 1.

c) Increasing $\beta_0, \beta_1, \beta_2$ will increase both our odds and log odds, while decreasing them would decrease our odds and log odds.

d) The formulation of our decision boundary is $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 3 + 2X_1 - 5X_2$. Points on the decision boundary are points where P(Y=1) = P(Y=0). For example, the point (1, 1) is on the decision boundary.

e) The coefficients changing is indicative of mutlticolinearity. This is potentially problematic because it undermines the significance of a single variable. It can also give the illusion of statistical significance.

6 Logistic Regression with Interaction Term

a) The intercept indicates that on average, a mother who is 23 and made infrequent visits to the physician during the first trimester has a 0.52 less chance of having a baby with low birth weight. The age coefficient indicates that a one unit increase in age will, on average, produce 0.04 greater odds that a mother will have a baby with low birth weight. The frequency coefficient indicates that a mother who visits the physician frequently during the first trimesster will on average, produce 0.47 less chance of having a baby with low birth weigth. The age x frequency coefficient indicates that a one unit increase in age*frequency will on average result in 0.18 less chance of having a baby with low birth weight.

b) The model is $-0.52 + (0.04)(\text{Age}) + (-0.47)(\text{Frequency}) + (-0.18)(\text{Age x Frequency})$. When a mother visits the physician frequently during the first trimester, the model is $-0.99 + (-0.14)(\text{Age})$. For this model, a one unit increase in the mother's age produces an average of 0.14 less chance of having a baby with a low birth weight. When a mother visits the physician infrequently during the first trimester, the model is $-0.52 + (0.04)(\text{Age})$, where a one unit increase in the mother's age produces an average of 0.04 less change of having a baby with a low birth weight.

c) The odds ratio is calculated as follows:
$ln(\frac{P(Y=1)}{P(Y=0)}) = -0.52 + (0.04)(\text{Age}) + (-0.47)(\text{Frequency}) + (-0.18)(\text{Age x Frequency})$
where we set age to a particular value, and take the quotient of when Frequency = 1 and Frequency = 0.
Age 17: $ln(\frac{P(Y=1)}{P(Y=0)}) = -0.52+(0.04)(-6)+(-0.47)(\text{Frequency})+(-0.18)(-6*\text{Frequency})$
$ln(\frac{P(Y=1)}{P(Y=0)}) = -0.52 + -0.24 + (-0.47)(\text{Frequency}) + (1.08)(\text{Frequency})$
$ln(\frac{P(Y=1)}{P(Y=0)}) = -0.76 + (0.61)(\text{Frequency})$
Odds Ratio: $\frac{e^{-0.15}}{e^{-0.76}} = 1.840$
To compute the odds ratio for the other ages, simply replace Age with 23-age.

Age	**Odds Ratio**	**95% Confidence Interval**
17	1.840	(0.705, 4.949)
23	0.625	(0.325, 1.201)
24	0.522	(0.262, 1.036)
25	0.436	(0.206, 0.916)
30	0.177	(0.050, 0.607)

d) An odds ratio of 1 indicates that both events have an equal probability of occurring. Thus, in our odds ratio table, the column indicates the probability of a mother who makes frequent physician visits during the first trimester have a baby with low birth weight, divided by the probability for a mother who did not make frequent visits. For example, for mothers age 17, the probability of a mother who makes frequent physician visits having a baby with low birth weigh has on average, 1.840 times the probability of a mother who does not make frequent physician visits, also age 17, having a baby with low birth weight. A number under 1 indicates that the event is less likely to occur, while an odds ratio greater than 1 indicates that the event is more likely to occur. An odds ratio that falls within the confidence interval indicates that the odds ratio is statistically significant.

e) The difference in probability is calculated as follows:
$ln(\frac{P(Y=1)}{P(Y=0)}) = -0.52 + (0.04)(\text{Age}) + (-0.47)(\text{Frequency}) + (-0.18)(\text{Age x Frequency})$
where we set age to a particular value, and take the difference of when Frequency = 1 and Frequency = 0.
Age 17: $ln(\frac{P(Y=1)}{P(Y=0)}) = -0.52+(0.04)(-6)+(-0.47)(\text{Frequency})+(-0.18)(-6*\text{Frequency})$
$ln(\frac{P(Y=1)}{P(Y=0)}) = -0.52 + -0.24 + (-0.47)(\text{Frequency}) + (1.08)(\text{Frequency})$
$ln(\frac{P(Y=1)}{P(Y=0)}) = -0.76 + (0.61)(\text{Frequency})$
Difference in Probability: $e^{-0.15} - e^{-0.76} = 0.393$
To compute the difference in probability for the other ages, simply replace Age with 23-age.

Age	**Difference in Probability**	**95% Confidence Interval**
17	0.393	(-0.788,0.393)
23	-0.223	(-0.197,0.088)
24	-0.296	(-0.232,0.046)
25	-0.363	(-0.315,-0.016)
30	-0.647	(-0.540,-0.092)

The difference in probability is a bit more self explanatory. It is the difference between the probabilty of a mother of a certain age that makes frequent physician visits having a baby with low birth weight and a mother of the same age but makes infrequent physician visits having a baby with low birth weight. The results are consistent with those from part c: mothers of age 17 that make frequent physician visits are more likely to have a baby with low birth weight compared of 17 year old mothers who don't make frequent physician visits. For mothers of age 30, the difference in probability is the greatest compared to other ages in the table. However, for part c, all of the results were statistically significant, while not all results for difference in probability are statistically singificant.

7 Multinomial Logistic Regression

a) This is because we are keeping a variable fixed, but at different values, which gives different coefficients.

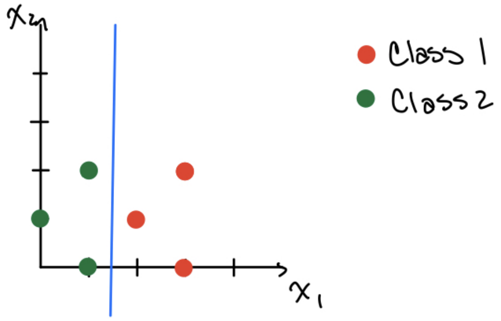
b) No, there are no features that are insignificant across the board.

c) age is only significant for PID = 1, PID = 2, and PID = 5.

d) selfLR is significant for all the classes.

8 Support Vector Machine

a) Plot shown below

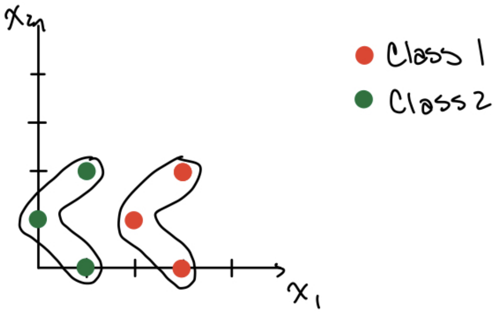


b) Points (1, 0), (1, 1), and (2, 1) are support vectors. If (2, 1) was removed, then the boundary would change.

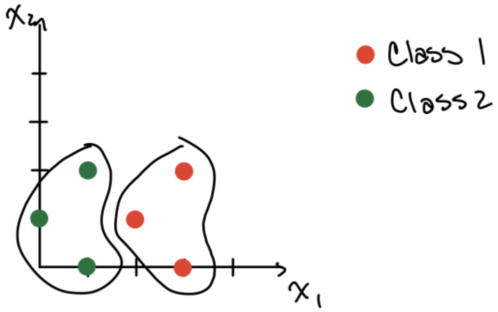
c) A hard margin is used for linearly separable data, and does not allow for misclassifcations. It tries to maxmimize the distance between data points and the boundary. Soft margin allows misclassification in hopes of achieving better generality, so it tries to minimize misclassification error. For this dataset, it doesn't matter whether use hard or soft margin, as they result in the same decision boundary.

d) The left sub-figure corresponds to SVM (linear), the middle sub-figure corresponds to SVM with polynomial kernel, and the right sub-figure corresponds to SVM with RBF kernel with width (γ) equal to 1.

e) Decision boundary for an SVM classifier with RBF kernel with width (γ) equal to 20, without regularization.


x_2
x_1
Class 1
Class 2

f) Decision boundary for an SVM classifier with RBF kernel with $\gamma = 20$ and $C = 0.1$.


x_2
x_1
Class 1
Class 2