1a)

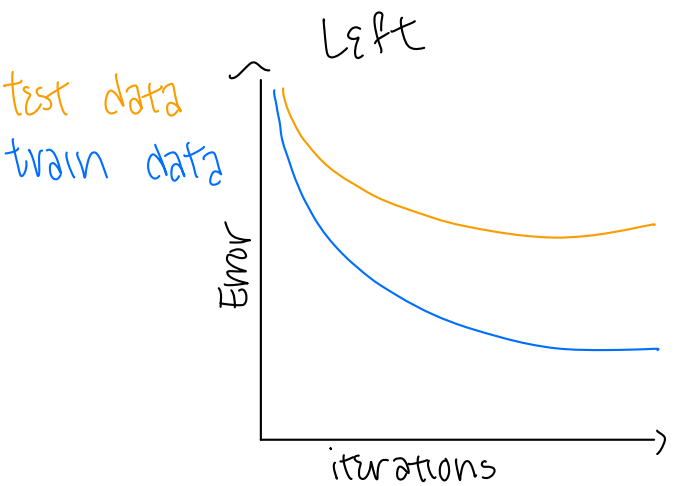
larger bias: 1

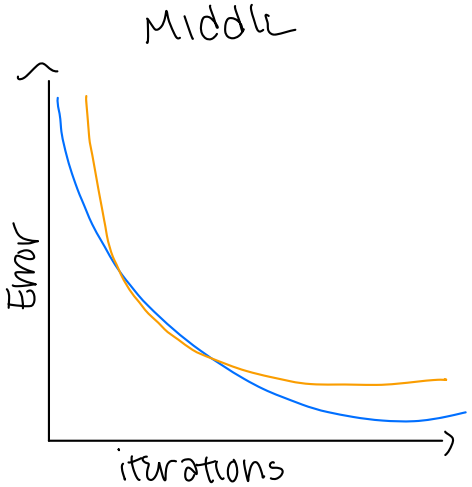
larger variance: 3

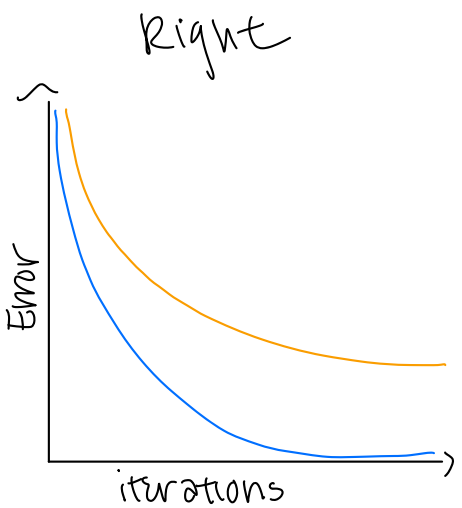
1: bad on train & test

2: alright on train & test

3: really good on train, bad on test







1b) L1 regularization is used on a, b/c the coefficients are nullified fast, compared to b2, when they are not nullified

2a) Normal $\Rightarrow L = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

2b) $L(\beta_0, \beta_1) = \frac{1}{n}\sum (y_i - [\beta_0 + \beta_1 x_i])^2$

$L = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

$= \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - y_i}{\sigma}\right)^2\right)$

$= \sum \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x_i - y_i)^2}{2\sigma^2}$

$= \sum -\frac{1}{2}\ln(\sigma^2 2\pi) - \frac{(x_i - y_i)^2}{2\sigma^2}$

$\frac{-n}{2}\ln(2\pi\sigma^2) - \sum \frac{(x_i - y_i)^2}{2\sigma^2}$

We want argmax (L)

which is just $\text{argmax}\left(-\sum \frac{(x_i - y_i)^2}{2\sigma^2}\right)$

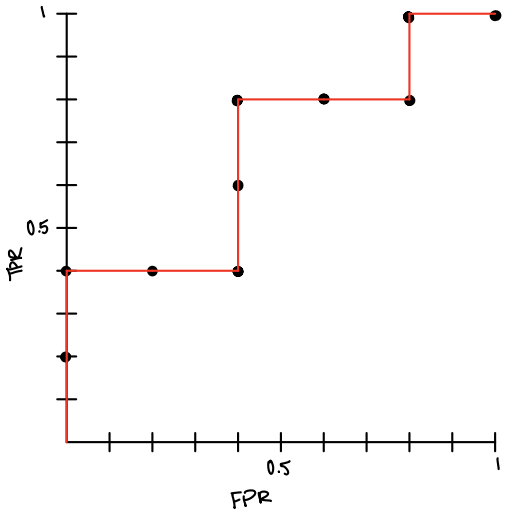
$= \text{argmin}\left(\sum \frac{(x_i - y_i)^2}{2\sigma^2}\right)$

3a)

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

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

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



3b) $(0.4)(1) + (0.4)(0.6) + (0.2)(0.2) = 0.68$

3c)

Predicted \ True	Pos	Neg
Pos	4	4
Neg	1	1

3d) Accuracy: $\frac{(TP+TN)}{ALL} = \frac{5}{10} = 0.5$

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

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

F1 : $(2)(0.8)(0.5)/(0.8+0.5) = 0.615$

3e) No, changing threshold to improve one score would negatively affect another score. This is because, for the thresholds, increasing # of TP will inc. the number of FP.

4a) The numbered lines are classes.

4b) For the most part, yes, though there are a few classes where it doesn't perform well.

4c) Yes & yes - more variation, more to train on

5a) $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ $\quad p: P(Y=1|X=0) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$

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

$P(Y=1) = [1-P(Y=1)]\, e^3$

$P(Y=1) = e^3 - e^3 P(Y=1)$

$e^3 = (e^3+1)P(Y=1)$

$P(Y=1) = 0.953$ (probability) $\quad$ odds: 20.086

5b) $\text{odds} = \frac{P(Y=1)}{1-P(Y=1)}$ $\quad$ inc X_1: inc. odds & log odds

inc X_2: dec. odds & log odds

5c) inc. $\beta_0, \beta_1, \beta_2$ would increase our odds & log odds, while decreasing them would decrease odds & log odds

5d) $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 3 + 2X_1 - 5X_2$

Points on the decision boundary are points where $P(Y=0) = P(Y=1)$

$P(Y=1) = 0.5$

$3 + 2X_1 - 5X_2 = 0$

$(1, 1)$

5e) The coefficients changing is indicative of multicolinearity. This is a problem b/c it undermines the significance of a single variable. Illusion of statistical sig.

6a) The intercept indicates that on avg. mother who is 23 & infrequently visited the physician . . . dec. prob. of low weight baby by 0.52.

on avg. 1yr inc. in age (above 23) corr. w/ 0.04 inc. in chance of low weight baby

on avg. visiting physician freq. results in 0.47 dec in chance of low weight baby

on avg. 1 unit inc. in age x freq results in 0.18 dec in chance of weight baby

6b) $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2)$

freqvent: $-0.52 + 0.04(Age) - 0.47 - 0.18(Age)$

$= -0.99 - 0.14(Age)$

infrequent: $-0.52 + 0.04\ (Age)$

Age will dec. or inc. odds

6c) $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -0.99 - 0.14(Age)$

Age	Odds Ratio	
17	1.840	$-0.76 + 0.61F$
23	0.625	$-0.52 - 0.47F$
24	0.522	$-0.52 + 0.04 - 0.65F$
25	0.436	$-0.52 + 0.08 - 0.47F - 0.36F$
30	0.177	$-0.52 + 0.28 - 0.47F - 1.26F$

6d) Those who visit frequently are less likely to have baby w/ low birth weight w/i CI indicates statistical significance

6e)

Age	diff. in prob	
17	0.393	$e^{-0.15} - e^{0.76}$
23	-0.223	$e^{-0.99} - e^{-0.52}$
24	-0.296	$e^{-1.13} - e^{-0.48}$
25	-0.363	$e^{-1.27} - e^{-0.42}$
30	-0.647	$e^{-1.97} - e^{-0.24}$

7a) This is because they vary the dependent variable, which results in different coefficients

7b) no?

7c/d) PID: 1 : agr

PID: 2 : log popul, agr, educ, income const

PID: 3. const

PID: 4 : log popul, educ, income, const

PID: 5 : log popul, agr, educ, income, const

PID: 6 : log popul, educ, income, const

All: SelfLR

Agr: ||||

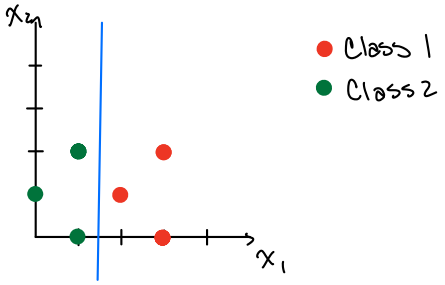
logp : ||||

const : ||||

educ : ||||

income: ||||

8a)



8b) (1,0), (1,1), (2,1)

yes if (2,1) was removed it would shift right

8c) hard - linearly sep, no misclassifications (maximize distance)

soft - allows misclassifications for better generality (minimize misclassification error)

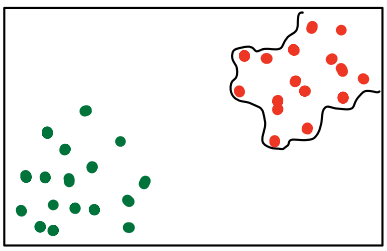
same decision boundary

8d) left - linear

middle - polynomial

right - rbf

8d)



8e)