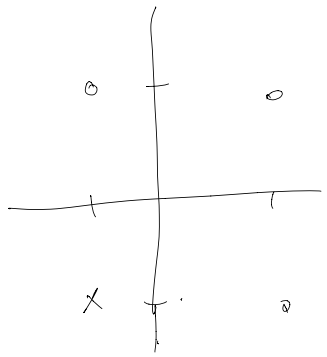
1. Perceptron

1a) OR

X_1	X_2	Y
-1	-1	-1
-1	1	1
1	-1	1
1	1	1



Valid perceptron

$a = w^T x + b$ where $w \in \mathbb{R}^2$

$w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $b = 1/2$

$a = x_1 + x_2 + 1/2$

X_1	X_2	$a(x)$	Y
-1	-1	$-3/2$	-1
-1	1	$1/2$	1
1	-1	$1/2$	1
1	1	$5/2$	1

$sign(a(x)) = Y$ for all data points.

another valid perceptron

$a = w^T x + b$, $w_1 = 1$, $w_2 = 1$, $b = 1/3$

X_1	X_2	$a(x)$	Y
-1	-1	$-5/3$	-1
-1	1	$1/3$	1
1	-1	$1/3$	1
1	1	$7/3$	1

$sign(a(x)) = Y$ for all data points.

16) XOR

x_1	x_2	y
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

valid perceptron does not exist since this data set is not linearly separable.

proof.

$$a = w^T x + b = x_1 w_1 + x_2 w_2 + b$$

$$-x_1 - x_2 + b < 0$$
$$-x_1 + x_2 + b > 0$$
$$x_1 - x_2 + b > 0$$
$$x_1 + x_2 + b < 0$$

contradict.

$$-x_1 - x_2 < -b$$
$$x_1 + x_2 < -b$$

contradict.

2. Logistic Regression

$$J(\theta) = -\sum_{n=1}^{N} \left[y_n \log h_\theta(x_n) + (1-y_n) \log (1-h_\theta(x_n)) \right]$$

* $\frac{\partial}{\partial \theta_j} h_\theta(x_n) = \frac{\partial}{\partial \theta_j} \sigma(\theta^T x_n) = (1-\sigma(\theta^T x_n)) \sigma(\theta^T x_n) \frac{\partial}{\partial \theta_j} (\theta^T x_n)$

  * $\frac{\partial}{\partial \theta_j} (\theta^T x_n) = x_{nj}$

(a) $\frac{\partial J}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} -\sum_{n=1}^{N} \left[y_n \log h_\theta(x_n) + (1-y_n) \log (1-h_\theta(x_n)) \right]$

$= -\sum_{n=1}^{N} \left[\frac{\partial}{\partial \theta_j} y_n \log h_\theta(x_n) + \frac{\partial}{\partial \theta_j} (1-y_n) \log (1-h_\theta(x_n)) \right]$

$= -\sum_{n=1}^{N} \left[y_n \frac{(1-\sigma(\theta^T x_n)) \sigma(\theta^T x_n)}{\sigma(\theta^T x_n)} \frac{\partial}{\partial \theta_j} (\theta^T x_n) - (1-y_n) \frac{(1-\sigma(\theta^T x_n)) \sigma(\theta^T x_n)}{1-\sigma(\theta^T x_n)} \frac{\partial}{\partial \theta_j} (\theta^T x_n) \right]$

$= -\sum_{n=1}^{N} \left[y_n (1-\sigma(\theta^T x_n)) \frac{\partial}{\partial \theta_j} (\theta^T x_n) - (1-y_n) \sigma(\theta^T x_n) \frac{\partial}{\partial \theta_j} (\theta^T x_n) \right]$

$= -\sum_{n=1}^{N} \left[y_n - y_n \sigma(\theta^T x_n) - \sigma(\theta^T x_n) + y_n \sigma(\theta^T x_n) \right] \frac{\partial}{\partial \theta_j} (\theta^T x_n)$

$= -\sum_{n=1}^{N} \left[y_n - \sigma(\theta^T x_n) \right] \frac{\partial}{\partial \theta_j} (\theta^T x_n)$

$= -\sum_{n=1}^{N} \left[y_n - h_\theta(x_n) \right] x_{nj}$

$= \sum_{n=1}^{N} \left[h_\theta(x_n) - y_n \right] x_{nj}$

2b) $$\frac{\partial^2 J}{\partial\theta_j \partial\theta_k} = \frac{\partial}{\partial\theta_j} \sum_{n=1}^{N} [h_\theta(x_n) - y_n] x_{nk}$$

$$= \frac{\partial}{\partial\theta_j} \sum_{n=1}^{N} [\sigma(\theta^T x_n) - y_n] x_{nk} \quad (y_n \to 0)$$

$$= \frac{\partial}{\partial\theta_j} \sum_{n=1}^{N} \sigma(\theta^T x_n) x_{nk}$$

$$= \sum_{n=1}^{N} [1 - \sigma(\theta^T x_n)]\, \sigma(\theta^T x_n) \cdot x_{nj} x_{nk}$$

$$= \sum_{n=1}^{N} [1 - h_\theta(x_n)]\, h_\theta(x_n) x_{nj} x_{nk}$$

$$\Rightarrow H = \sum_{n=1}^{N} [1 - h_\theta(x_n)]\, h_\theta(x_n) x_n^T x_n$$

2c) $z^T H z \equiv \sum_{j,k} z_j z_k H_{jk} \geq 0$

$$H_{jk} = \sum_{n=1}^{N} (1 - h_\theta(x_n)) h_\theta(x_n) x_{nj} x_{nk}$$

$$z^T H z = \sum_{j,k} z_j z_k \sum_{n=1}^{N} (1 - h_\theta(x_n)) h_\theta(x_n) x_{nj} x_{nk}$$

$$= \sum_{n=1}^{N} (1 - h_\theta(x_n)) h_\theta(x_n) \sum_{j,k} z_j z_k x_{nj} x_{nk}$$

$$= \sum_{n=1}^{N} (1 - h_\theta(x_n)) h_\theta(x_n) (z^T z)(x_n^T x_n)$$

$$z^T z = \sum_{z} z^2 \geq 0$$

$$x_n^T x_n = \sum_{x_n} x_n^2 \geq 0$$

$$(1 - h_\theta(x_n)) h_\theta(x_n) \geq 0$$

since $0 \leq h_\theta(x_n) = \sigma(\theta^T x_n) \leq 1$

Since the Hessian is positive semi-definite, function J is convex.

3 a) $L(\theta) = p(x_1, \ldots, x_n \mid \theta)$

assume iid $= \prod_{i=1}^{n} p(x_i \mid \theta) = \boxed{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}}$

$p(x_i = 1 \mid \theta) = \theta$

$p(x_i = 0 \mid \theta) = 1 - \theta$

$p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$

The likelihood does not depend on order of data because. likelihood is the product of all $p(x_i \mid \theta)$ where x_i is the independent random variable that is drawn from Bernoulli distribution. Changing the order would not change the product.

3b) $\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i}$

$$= \sum_{i=1}^{N} \log \theta^{x_i} + \log (1-\theta)^{1-x_i}$$

$$= \sum_{i=1}^{N} x_i \log \theta + (1-x_i) \log (1-\theta)$$

$$\ell'(\theta) = \sum_{i=1}^{N} \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}$$

$$= \sum_{i=1}^{N} \frac{x_i - \cancel{x_i\theta} - \theta + \cancel{x_i\theta}}{\theta(1-\theta)} = \sum_{i=1}^{N} \frac{x_i - \theta}{\theta(1-\theta)}$$

$$= \frac{1}{\theta(1-\theta)} \left(\sum_{i=1}^{N} x_i \right) - N\theta$$

$$\ell''(\theta) = \sum_{i=1}^{N} -\frac{x_i}{\theta^2} - \frac{1-x_i}{(1-\theta)^2}$$

$\ell'(\hat{\theta}) = 0$ $\quad \sum_{i=1}^{N} \frac{x_i - \hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = 0$

$$\sum_{i=1}^{N} \frac{x_i}{\cancel{\hat{\theta}(1-\hat{\theta})}} = \frac{N\hat{\theta}}{\cancel{\hat{\theta}(1-\hat{\theta})}}$$

$$\sum_{i=1}^{N} x_i = N\hat{\theta}$$

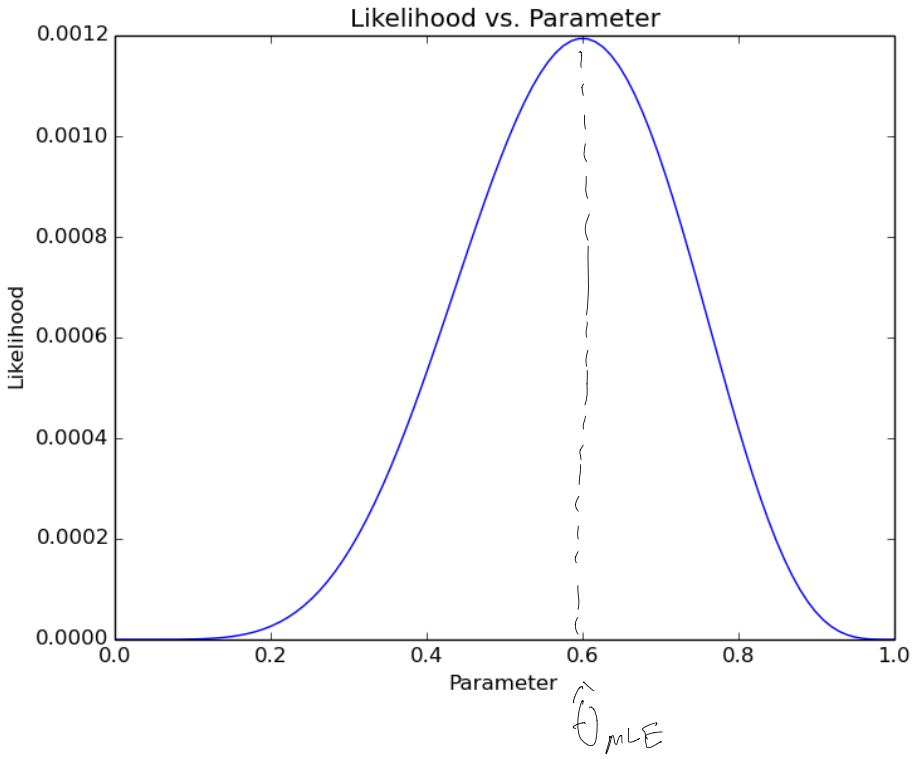
$$\boxed{\hat{\theta} = \sum_{i=1}^{N} \frac{x_i}{N}}$$

$$\ell''(\hat{\theta}) = \sum_{i=1}^{N} -\frac{x_i}{\hat{\theta}^2} - \frac{1-x_i}{(1-\hat{\theta})^2}$$

$\ell''(\hat{\theta}) \leq 0$

Therefore, $\hat{\theta}$ is the maximum likelihood estimator

3C)

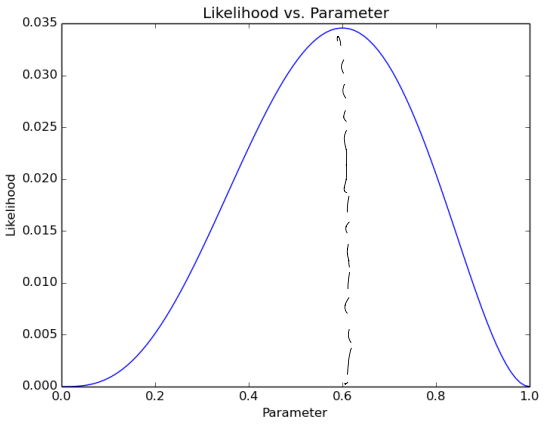


$\hat{\theta}_{MLE}$

The closed form answer, $\hat{\theta}_{MLE} = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{3}{5} = 0.6$

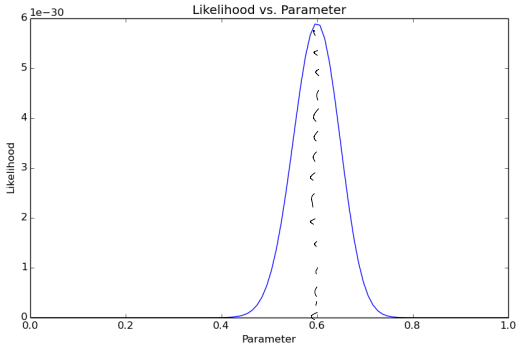
agrees with the above plot.

3d)

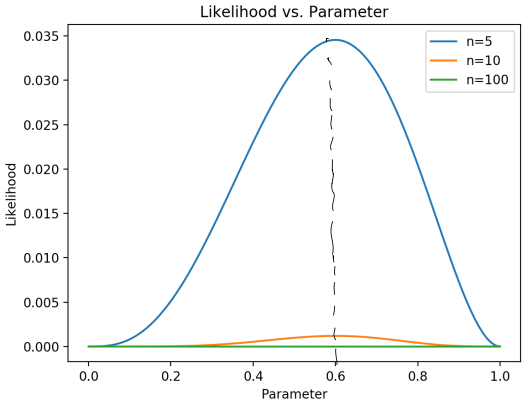


N=5

MLE for each size of N are all same at 0.6. The difference in the likelihood function is the value of the likelihood. For lower values of N, the maximum value of likelihood is higher than higher values of N.



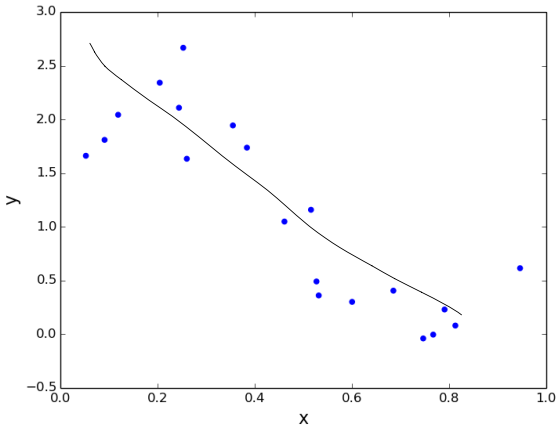
N=100



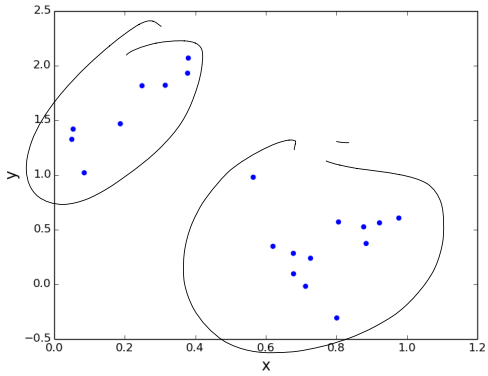
All combined

On the same scale, it is more apparent that likelihood values are higher for lower n values.

4 a)



Training data



Test data

I observe general trend of decreasing y values over decreasing x values in the training data. However, test data shows two groups that are clustered together with weak correlation. Linear regression may be effective for training and building a model, but the models would not be effective at predicting for test data.

4.)

Step size	Iteration	Cost	Coefficients
0.0001	10000	4.09	[2.27, -2.46]
0.001	7676	3.91	[2.45, -2.82]
0.01	765	3.91	[2.45, -2.82]
0.0407	10000	2.71×10^{29}	$[-9.4 \times 10^{18}, -4.65 \cdot 10^{18}]$

Learning rate of 0.0001 hits the max iteration limit and ends with the cost of 4.09, This is due to small learningrate and slow updates of weights.

learning rate of 0.001 and 0.01 both converge at cost of 3.91 before max iteration limit. 0.01 converges first because of bigger step size than 0.001.

Learning rate of 0.0407 does not converge before max iteration limit with extremely large cost. This indicates that the step size is too large to be used for gradient descent.

Coefficients are same for converged models with learning rates 0.001 and 0.01. Step size 0.0001 results in similar coefficients. Step size 0.0407 results in vastly different coefficients

4e closed form solution results in the objective function value (cost) of 3.91. and coefficient of $[2.45, -2.82]$, which is the same as coefficients resulted from gradient descent fittings with step size 0.01 and 0.001, which converged. The algorithm runs much faster than gradient descent. Gradient descent with learning rate = 0.01 took 0.09 seconds to fit the data, this was the fastest computing time for GD. For closed form, it took 0.0004 seconds to fit the data.

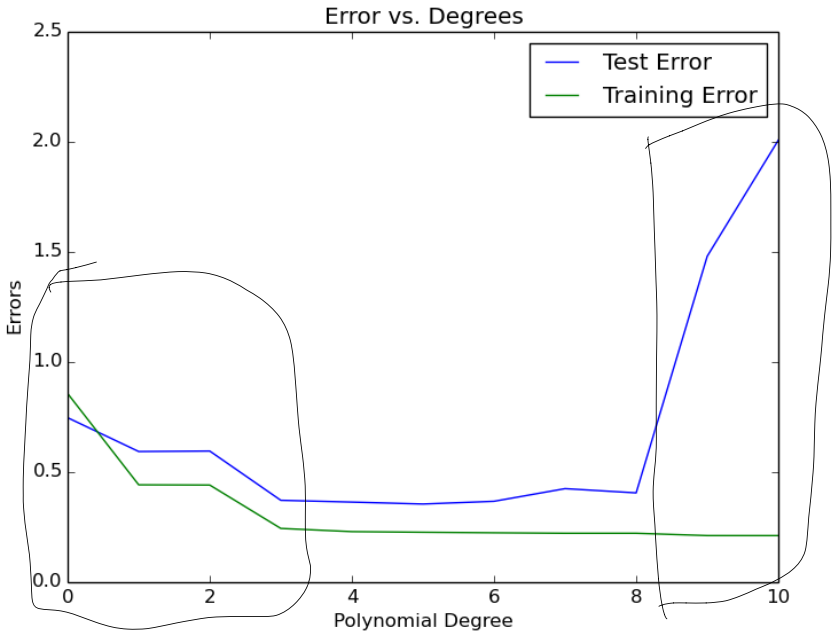
4f). GD with $\eta = \frac{1}{1+k}$

Iteration: 1408 to converge.

cost: 3.91

It takes longer than $\eta = 0.01$ but shorter than $\eta = 0.001$ to converge.

4i)



polynomial degree of 5 is the best since it seems to minimize both test error and training error for the data.

Overfitting is apparent for polynomial degree of higher than 8. Test error increases drastically.

Test error and training error seems to decrease until degree of 3. Below degree of 3 would be underfitting.