CS155 Final

Timothy Liu

March 14, 2018

1 Problem 1

1.1 Question A

True.

1.2 Question B

Option B

1.3 Question C

K1: middle image
K2: rightmost image
K3: leftmost image

1.4 Question D

Option B y = 2.

1.5 Question E

Option B

1.6 Question F

True

1.7 Question G

Option A

1.8 Question H

False

1.9 Question I

True

1.10 Question J

False

1.11 Question K

Option B

1.12 Question L

True

1.13 Question M

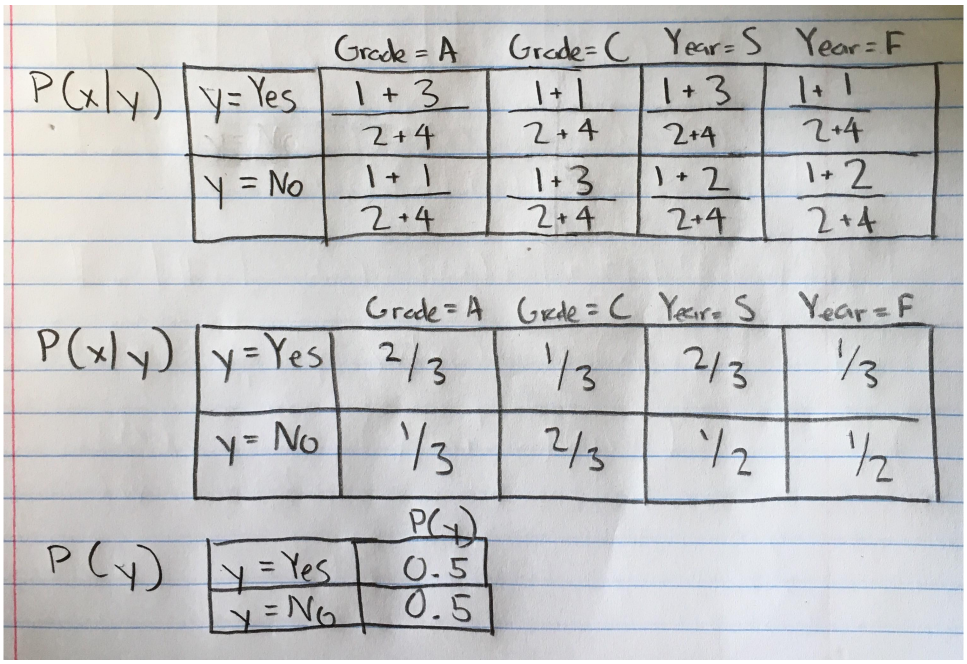
True

1.14 Question N

True

2 Problem 2

2.1 Question 1



P(x|y)

	Grade = A	Grade = C	Year = S	Year = F
y = Yes	$\frac{1+3}{2+4}$	$\frac{1+1}{2+4}$	$\frac{1+3}{2+4}$	$\frac{1+1}{2+4}$
y = No	$\frac{1+1}{2+4}$	$\frac{1+3}{2+4}$	$\frac{1+2}{2+4}$	$\frac{1+2}{2+4}$

P(x|y)

	Grade = A	Grade = C	Year = S	Year = F
y = Yes	2/3	1/3	2/3	1/3
y = No	1/3	2/3	1/2	1/2

P(y)

	P(y)
y = Yes	0.5
y = No	0.5

Figure 1: Naive Bayes model parameters.

2.2 Question 2

P(Year = Freshman, Grade = C, Happy = No) =
$P(Happy = No) \times P(Grade = C|Happy = No) \times P(Year = Freshman|Happy = No) =$
$\frac{1}{2} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{6}$

2.3 Question 3

```
#generate y value
y_seed = random()
if y_seed < P(Happy?):
        y = Yes
else:
        y = No

#generate x values
x_seed1 = random()
if x_seed1 < P(Grade = C|y):
        x1 = C
else:
        x1 = A

x_seed2 = random()
if x_seed2 < P(Year = Senior|y):
        x2 = Senior
else:
        x2 = Freshman

return (y, x1, x2)
```

Figure 2: Pseudocode for drawing from a trained model given the parameters have been calculated..

3 Problem 3

3.1 Question 3

3.1 $\tilde{x} = Ax$ Transformation

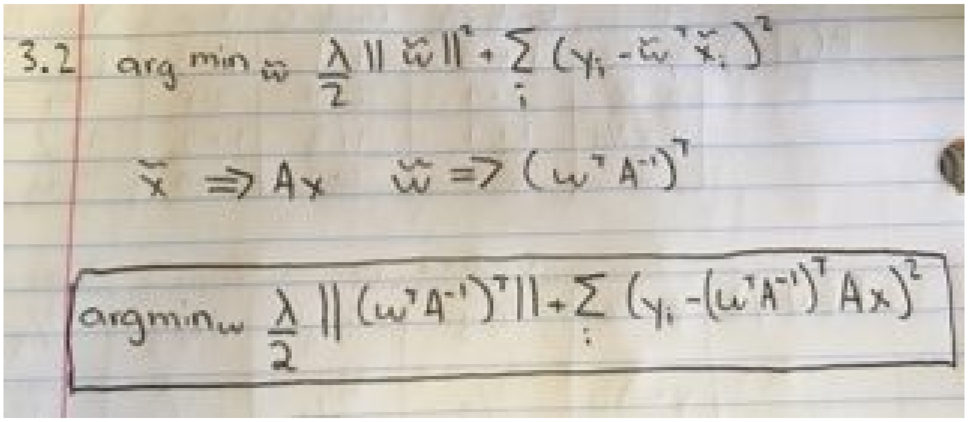
$$w^T x = \tilde{w}^T \tilde{x}$$

$$w^T x = \tilde{w}^T A x$$

$$w^T = \tilde{w}^T A \Rightarrow \tilde{w}^T = w^T A^{-1}$$

$$\boxed{\tilde{w} = (w^T A^{-1})^T}$$

3.2 Question 2



3.2 $\arg\min_{\tilde{w}} \frac{\lambda}{2} \|\tilde{w}\|^2 + \sum_i (y_i - \tilde{w}^T \tilde{x}_i)^2$

$\tilde{x} \Rightarrow Ax \quad \tilde{w} \Rightarrow (w^T A^{-1})^T$

$\boxed{\arg\min_w \frac{\lambda}{2} \|(w^T A^{-1})^T\| + \sum_i (y_i - (w^T A^{-1})^T A x)^2}$

3.3 Question 3

When A is a rescaling matrix, its effect on the weight matrix in Question 1 is to also rescale it. Standard ridge regression is impacted only by the parameter λ which only appears outside of the summation. The λ puts a constrain on the norm of the weight vector. The answer from Question 2 changes the weight vector inside and outside of the summation. Adding A to the inside means the error that we are trying to minimize is also affected by how far the predicted value is from the true value through A. Ridge regression does not have any impact on the accuracy of the weight vector, only the norm. The answer to Question 2 affects both.

4 Problem 4

4.1 Question 1

U and V are selected to maximize P(S). At worse, U and V can both be equal to X, which would make P(S) for the dual point model equivalent to P(S) for the single point model. However, U and V can be selected more optimally, and having two vectors allows greater freedom to choose U and V to maximize P(S). Thus, the data likelihood for the dual-point model is never less than that of the single-point model.

4.2 Question 2

This would imply that U, V, and X are all equivalent to each other and that the dual point and single point models are the same.

5 Problem 5

5.1 Question 1

5.1 $$\frac{\partial}{\partial w_{11}} (y - f(x))^2$$

$$= -2(y - f(x)) \frac{\partial}{\partial w_{11}} f(x)$$

$$= -2(y - f(x)) \frac{\partial}{\partial w_{11}} \sigma(s(x)) \quad s(x) = \sum_{i=1}^{2} u_i h_i(x))$$

$$= -2(y - f(x)) \sigma(s(x))(1 - \sigma(s(x)) \frac{\partial}{\partial w_{11}} s(h(x))$$

$$= -2(y - f(x) \sigma(s(x))(1 - \sigma(s(x)) \sum_{i=1}^{2} u_i \frac{\partial}{\partial w} \sigma\left(\sum_{j=1}^{2} w_{ji} x_j\right)$$

$$\boxed{= -2(y - f(x) \sigma(s(x))(1 - \sigma(s(x)) \cdot \sum_{i=1}^{2} u_i \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\left(1 - \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\right) x_1}$$

5.2 Question 2

5.2 $-2(y - f(x)\sigma(s(x))(1-\sigma(s(x))\cdot$

$$\sum_{i=1}^{2} u_i \, \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\left(1 - \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\right) x_1$$

$$\sum_{j=1}^{2} w_{j1} x_j = 0.1 * 0.25 + 0.05 * 0.5 = 0.05$$

$$s(x) = \sum_{i=1}^{2} u_i h_i(x)$$

$$= 0.5 h_1(x) - 0.1 h_2(x)$$

$$h_1(x) = \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right) = \sigma(0.05) = 0.5125$$

$$h_2(x) = \sigma\left(\sum_{j=1}^{2} w_{j2} x_j\right) = \sigma(-0.75) = 0.321$$

$$s(x) = 0.5(0.5125) - 0.1(0.321) = \underline{0.22415}$$

$$f(x) = \sigma(s(x)) = 0.558$$

$$\frac{\partial}{\partial w_{11}} L(y, f(x)) = -2(0.75 - 0.558(0.558))\cdot (1 - 0.558)\cdot(0.4)(0.5125)(1-0.5125)(0.1)$$

$$\boxed{\frac{\partial}{\partial w_{11}} L(y,(f(x)) = -0.0039}$$

5.3 Question 3

The sigmoid term at the output of each adder results in the vanishing gradient problem. Any value far away from 0 has a small gradient. The problem worsens with multiple layers because the gradient is multiplied by several values that are close to zero, since each layer has its own sigmoid.