Brandon Peck
Machine Learning
Homework 5

(Q1) Preprocessing
Implemented 1st Choice

(Q5) MNBC Theta

```
thetaPos = [  9.61921189e-04   2.07747044e-04   1.01778594e-03   9.39226134e-04
   9.07802211e-05   6.98309393e-04   2.53137155e-04   4.64375746e-04
   5.93562984e-05   5.76105249e-05   2.79323757e-05   6.80851658e-05
   1.92733392e-03   7.78614973e-04   9.92447784e-01]
```

```
thetaNeg = [  7.42773032e-04   7.77772809e-05   6.80551208e-04   4.99719030e-04
   2.52776163e-05   6.41662567e-04   1.34165809e-04   1.34943582e-03
   3.49997764e-04   2.83887075e-04   1.61387858e-04   3.32497876e-04
   2.96720327e-03   1.46999061e-03   9.90283673e-01]
```

(Q6) MNBC Accuracy
MNBC classification accuracy = 0.6583333333333333

(Q7) MNBC SKLearn Accuracy
Sklearn MultinomialNB accuracy = 0.676666666667

(Q11) BNBC Theta

```
ThetaPosTrue = [0.34045584045584043, 0.05270655270655271, 0.3504273504273504,
0.26638176638176636, 0.018518518518518517, 0.3247863247863248, 0.08974358974358974,
0.4886039886039886, 0.19230769230769232, 0.15954415954415954, 0.10541310541310542,
0.18518518518518517, 0.688034188034188, 0.39886039886039887, 0.9985754985754985]
```

```
thetaNegTrue = [0.40883190883190884, 0.12393162393162394, 0.4928774928774929,
0.41452991452991456, 0.06552706552706553, 0.37037037037037035, 0.1467236467236467,
0.2621082621082621, 0.045584045584045586, 0.038461538461538464,
0.022792022792022793, 0.05128205128205128, 0.5427350427350427, 0.2777777777777778,
0.9985754985754985]
```

(Q12) BNBC Accuracy
BNBC classification accuracy = 0.6583333333333333

Sample Questions:
Question 1:
(a)
1/3

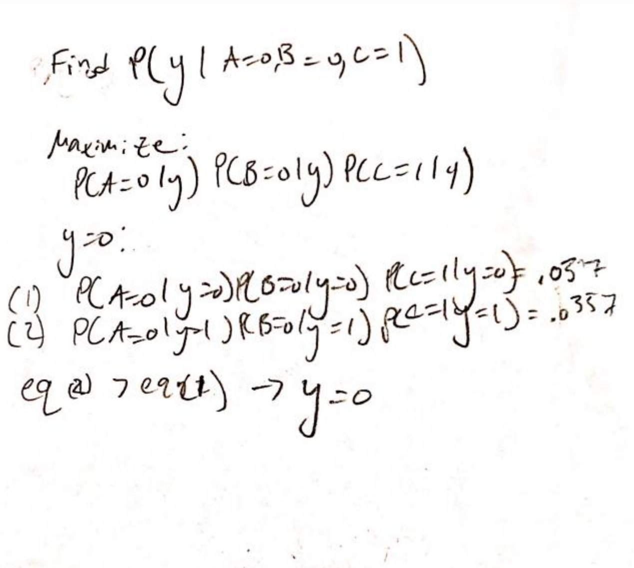
$$P(G|a,b) = \frac{P(G,a,b)}{P(a,b)} = \frac{P(G)P(a|G)P(b|G)}{P(a)P(b)} = \frac{(\frac{1}{2}) * (\frac{1}{2}) * (\frac{1}{2})}{(\frac{1}{2}) * (\frac{3}{8})} = (\frac{1}{3})$$

(b)
False - logistic regression follows the model p(C|X) whereas the Naïve Bayes models p(X|C). Here C is a discrete value and X is discrete or continuous. So, with logistic regression we are calculating some variable having been given the data of X. Naïve Bayes computes p(C) and p(C|X) first.

(c)
True – they both are prefixed Gaussian to detail how the data is expected to be distributed.

(d)
False – The classifier will have a quadratic decision boundary.

Question 2:
(a)



Find $P(y \mid A=0, B=0, C=1)$

Maximize:
$P(A=0 \mid y)\, P(B=0 \mid y)\, P(C=1 \mid y)$

$y=0$:
(1) $P(A=0 \mid y=0) P(B=0 \mid y=0)\, P(C=1 \mid y=0) = .037$
(2) $P(A=0 \mid y=1) P(B=0 \mid y=1)\, P(C=1 \mid y=1) = .0357$

eq (2) > eq (1) → $y=0$

(b)

Yes because variables A, B, and C may not be independent when conditioned on the dependent variable. The example gives y = A XOR B in which case A may not depend on B but y depends on how they interact with oneanother.