CMSI 3300 – Classwork 6

Instructions:
This worksheet will not only provide you with practice problems for your upcoming exam, but will add to your toolset as initiate data scientists taking various machine learning problems from start to finish. Specific notes:

- Provide answers to each of the following questions and write your responses in the blanks. If you are expected to show your work in arriving at a particular solution, space will be provided for you.
- Place the names of your group members below:

Group Members:

1. Milo Fritzen
2. ________________
3. ________________

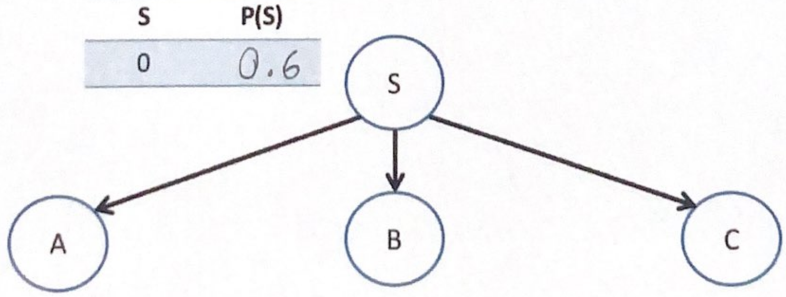
Problem 1 – Naïve Bayes Classifiers

Forney Industries' FornGene group has continued work developing a classifier to identify cases of Schistoforneymiosis [S] based on indicators of three genetic markers [A, B, C]. In this section, we will attempt to craft a Naïve Bayes Classifier that explains the data collected by the FornGene group, and use that to answer classification queries related to the disease. Note how this problem is different than the "Bag of Words" example from class in that each feature's CPT has its own parameters.

S	A	B	C	# of datum
0	0	0	0	48
0	0	0	1	144
0	0	1	0	12
0	0	1	1	36
0	1	0	0	72
0	1	0	1	216
0	1	1	0	18
0	1	1	1	54
1	0	0	0	21
1	0	0	1	189
1	0	1	0	7
1	0	1	1	63
1	1	0	0	9
1	1	0	1	81
1	1	1	0	3
1	1	1	1	27

1.1: In the table above, we track each combination of A, B, C, and S witnessed from the group's study. From this data, construct the *Maximum Likelihood* CPTs over each feature below. *Place your final answers directly in the tables below.*

S	P(S)
0	0.6



S	A	P(A | S)
0	0	0.4
1	0	0.7

S	B	P(B | S)
0	0	0.8
1	0	0.75

S	C	P(C | S)
0	0	0.25
1	0	0.1

1.2. Using the data from the previous section, suppose we instead *smooth* our estimates for the feature likelihoods using Laplacian Smoothing with $k = 10$. Compute $P_{LAP,10}(C = 0|S = 1)$

$$P_{lap,10}(C=0|S=1) = \frac{40+10}{400+(10\cdot 2)} = \frac{50}{420} \approx 0.119$$

1.3: Using the NBC you learned in Part 1.1, determine (showing your work) the most likely class for each of the following data points.

1. $\{A = 0, B = 1, C = 0\}$

$$\sum_s P(S=s \mid A=0, B=1, C=0) = \frac{\sum_s P(S=s, A=0, B=1, C=0)}{P(A=0, B=1, C=0)}$$

$$P(S=0, A=0, B=1, C=0) = P(S=0)P(A=0|S=0)P(B=1|S=0)P(C=0|S=0) = 0.6(0.4)(0.2)(0.25) = 0.012$$

$$P(S=1, A=0, B=1, C=0) = \ldots = 0.4(0.7)(0.25)(0.1) = 0.007$$

$$P(A=0, B=1, C=0) = 0.012 + 0.007 = 0.019$$

$$P(S=0|A=0, B=1, C=0) = \frac{0.012}{0.019} \approx 0.632 > 0.5 \quad \therefore \hat{y} = (S=0)$$

2. $\{A = 1, C = 1\}$ [Hint: Remember that NBCs are still Bayesian Networks, so B as a missing feature shouldn't be a problem – make sure your answer derives why]

$$\sum_s P(S=s \mid A=1, C=1) = \frac{\sum_s \sum_b P(S=s, A=1, B=b, C=1)}{\sum_b P(A=1, B=b, C=c)}$$

$$\sum_b P(S=0, A=1, B=b, C=1) = P(S=0)P(A=1|S=0)P(C=1|S=0)\underbrace{\sum_b P(B=b|S=0)}_{(P(B=0|S=0)+P(B=1|S=0))}$$

$$= 0.6(0.6)(0.75)((0.8)+(0.2)) = 0.27$$

$$\sum_b P(S=1, A=1, B=b, C=1) = 0.4(0.3)(0.9)((0.75)+(0.25)) = 0.108$$

$$P(A=1, C=1) = 0.27 + 0.108 = 0.378$$

$$P(S=0|A=1, C=1) = \frac{0.27}{0.378} \approx 0.714 \quad \therefore \hat{y} = (S=0)$$

Problem 2 – Linear Perceptrons

Returning to our email classification supervised learning task, consider that we have a trinary class variable $Y \in \{0,1,2\} = \{Spam, Ham, \overset{\text{Honeybaked}}{\cancel{Important}}\}$. Moreover, we've decided on a simple feature extractor f that takes an email x as input and returns a vector of features indexed as:

1. $f_0(x) = \#\ of\ ALL\ CAPS\ words$
2. $f_1(x) = \#\ of\ occurrences\ of\ word\ \text{“free”}$
3. $f_2(x) = whether\ or\ not\ email\ is\ in\ known\ contacts$ *(0 = not in contacts, 1 = in contacts)*

2.1. What feature vector would the above feature extractor return for the following email?

Email (x)	$f_0(x)$	$f_1(x)$	$f_2(x)$
From: Ray.Toal@lmu.edu *Message:* Hi all, There is free pizza in the Keck Lab, COME GET IT!	3	1	1

Consider that the above is in a training set with label $y = 2$ (this is a very important email). If, during learning, our class weight vectors are as follows…

$w_0 = \langle 2, 2, -3 \rangle$ $\qquad w_1 = \langle -1, 1, 2 \rangle$ $\qquad w_2 = \langle 2, -3, 1 \rangle$

2.2. First, determine which class y our perceptron would currently assign this email.

$y=i$	$activation_{w_i}$
$y=0$	$3 \cdot 2 + 1 \cdot 2 + 1 \cdot -3 = 5$ } highest $\therefore \hat{y} = 0$
$y=1$	$3 \cdot -1 + 1 \cdot 1 + 1 \cdot 2 = 0$
$y=2$	$3 \cdot 2 + 1 \cdot -3 + 1 \cdot 1 = 4$

2.3. Did our Perceptron make a mistake? If so, calculate the updated weights that would amount from its misclassification. If not, draw BlindBot on vacation (or draw him in either case if you want, I'm a classwork, not a cop).

w_0 too high, $w_0 = \langle 2, 2, -3 \rangle - \langle 3, 1, 1 \rangle = \langle -1, 1, -4 \rangle$

w_2 too low, $w_2 = \langle 2, -3, 1 \rangle + \langle 3, 1, 1 \rangle = \langle 5, -2, 2 \rangle$

Problem 3 – Logistic Regression

Suppose now we pivot to a multiclass logistic regression model with $Y \in \{0,1,2\}$ and the following class-specific weights.

$w_0 = \langle 2, 2, -3 \rangle \qquad w_1 = \langle -1, 1, 2 \rangle \qquad w_2 = \langle 2, -3, 1 \rangle$

3.1. Compute the likelihoods of each class $P\left(Y^{(i)} \middle| x^{(i)}; w\right)$ that a Logistic Regression classifier would give for the sample features $f\left(x^{(i)}\right) = \langle 1,1,1 \rangle$. Show your work / steps in the box below.

$P(y^{(i)} \mid x^{(i)}; w) = \dfrac{e^{z_y}}{\sum_{y \in Y} e^{z_y}} \qquad z_y = w_y \cdot f(x)$

$z_0 = 2\cdot1 + 2\cdot1 - 3\cdot1 = 1$

$z_1 = -1\cdot1 + 1\cdot1 + 2\cdot1 = 2 \qquad z = \langle 1, 2, 0 \rangle$

$z_2 = 2\cdot1 - 3\cdot1 + 1\cdot1 = 0$

	$P(y^{(i)} \mid x^{(i)}; w)$
$y=0$	$e^1/(e^1+e^2+e^0) \approx 0.24$
$y=1$	$e^2/(e^1+e^2+e^0) \approx 0.67$
$y=2$	$e^0/(e^1+e^2+e^0) \approx 0.09$

$P\left(Y^{(i)} = 0 \middle\| x^{(i)}; w\right) =$	$P\left(Y^{(i)} = 1 \middle\| x^{(i)}; w\right) =$	$P\left(Y^{(i)} = 2 \middle\| x^{(i)}; w\right) =$
0.24	0.67	0.09

3.2. Suppose our sample $x^{(i)}$ was meant to have label $y^{(i)} = 0$. Determine the new weights that would be associated with each class in a single step of stochastic gradient descent with learning rate $\eta = 0.1$ and cross entropy loss. Show your work in the box below.

$$\frac{\partial L_{CE}}{\partial w_{k,i}} = -\left[\mathbb{1}(k=y^{(i)}) - P(\hat{y}^{(i)}=k \mid x^{(i)}; w)\right] \cdot f_i(x)$$

$$\frac{\partial L_{CE}}{\partial w_{0,i}} = -\left[1-0.24\right]\cdot\langle 1,1,1\rangle = \langle -0.76, -0.76, -0.76\rangle$$

$$\frac{\partial L_{CE}}{\partial w_{1,i}} = -\left[0-0.67\right]\cdot\langle 1,1,1\rangle = \langle 0.67, 0.67, 0.67\rangle$$

$$\frac{\partial L_{CE}}{\partial w_{2,i}} = -\left[0-0.09\right]\cdot\langle 1,1,1\rangle = \langle 0.09, 0.09, 0.09\rangle$$

$$w_{k,i} = w_{k,i} - \eta \cdot \frac{\partial L_{CE}}{\partial w_{k,i}}$$

$$w_{0,i} = \langle 2,2,-3\rangle - 0.1\cdot\langle -0.76, -0.76, -0.76\rangle = \langle 2.076, 2.076, -2.924\rangle$$

$$w_{1,i} = \langle -1,1,2\rangle - 0.1\cdot\langle 0.67, 0.67, 0.67\rangle = \langle -1.067, 2.076, 1.933\rangle$$

$$w_{2,i} = \langle 2,-3,1\rangle - 0.1\langle 0.09, 0.09, 0.09\rangle = \langle 1.991, -3.009, 0.991\rangle$$

$w_0 =$	$w_1 =$	$w_2 =$
$\langle 2.076, 2.076, -2.924\rangle$	$\langle -1.067, 2.076, 1.933\rangle$	$\langle 1.991, -3.009, 0.991\rangle$