CS 461 Homework 1

Due September 27th

Student: Matthew McCaughan, matthew.mccaughan@rutgers.edu

Problem 1

(a) If precision rate is used to measure the empirical error on the training set, we can define precision rate as:
True Positive / (True Positive + False Positive)
Based on this relation, we want to design the circle classifier to minimize the number of false positives we get from our data. We can consider this by making sure these classifiers have a bias to be smaller rather than larger, in order to be more sure in true positives.
(b) If recall rate is used to measure empirical error on the training set, we can define recall rate as:
True Positive/(True Positive + False Negative)
Based on this relation, we want to design this circle classifier to minimize the number of false negatives we get from our data. We can consider this by making sure these classifiers have a bias to be larger and overfit the data rather than smaller, to prioritize the cases of positives that are true.
(c) Problems that are expected for cases 1.1 and 1.2 are the opposite of the problems they are solving, Circle classifiers based on precision rate may run into the problem of underfitting the data, and classifiers based on recall rate may run into the problem of overfitting the data. A mathematical combination of the precision rate and recall rate would likely minimize some of the issues presented

Problem 2

(a) The sample spaces $\Omega_2, W_3, ..., W_M$ are defined as follows:
$\Omega_2 : (LL), (LR), (RL), (RR)$
$\Omega_3 : (LLL), (LLR)...(RRL), (RRR)$
$\Omega_M : (L...L), (L...R)...(R...L), (R...R)$

(b) The Sample space $(\Omega = \Omega_1 * \Omega_2 * ... * \Omega_M)$
is the product of sample spaces for each level of the Galton board. Ω is the set of all paths up and to the bottom level M.

(c) The final location at L_g where the ball finally arrives is the result of the ball taking a specified path of lefts and right, and the specific combination of left and right moves results in different position on the final location.For example, specifying the blue star from the green star, the former is the result of all left moves as it is the first and leftmost position on the final location, where the green star has taken 1 right turn at some point in its journey, and all left turns otherwise

(d) A random variable to model this final position could be defined as the number of left turns made after reaching the final position, where Random Variable X is the sum of the number of left turns made after reaching the final position, X = $\Omega(0, 1, 2...M)$

(e) Because our random variable is designed as the probability an event happens a certain amount of time in a certain sequence of moves, we can represent our random variable with the binomial distribution. Given P(X=x) = (M choose L)*$(1/2)^L$.

M is the number of levels the pass has to pass through so that the ball can make from 0 to M left moves based on the depth. Because of the Galton board design, balls have an equally likely chance to go left or right and are more likely to have a seemingly random oscillation left and right than all left or all right. This will leave us with something closer to a normal distribution of the balls as the levels increase, allowing for a more defined trans to occur. As explained in the central limit theorem, as the sample size increases for the distribution, the distribution begins to approximate to a normal distribution

Problem 3

(a) The chance that the plant will survive the week is a combination of both instances in which the plant lives. The probability the plant lives is the combined probability that it lives and it was watered, and it lives without being watered:
P(Lives) = P(Lives$\cap Watered$) + $P(Lives \cap NotWatered)$)
$P(Lives) = P(L|W) * P(W) + P(L|W') * P(W')$
$= 0.80 * 0.70 + 0.20 * 0.30$
$= 0.56 + 0.06$
$= 0.62$

The probability that the plant will survive the week is 0.62

(b) The probability that the plant will be dead when I return after my friend forgot to water it is the probability that the plant will die without water over the week:
P(Dies given no water)
= 1-P(D|W)
$= 1 - 0.2$
$= 0.8$
The probability the plant will die given my friend forgot to water it is 0.8

(c) If the plant is dead when I return, the probability that my friend forgot to water it requires implementing Bayes rule:
P(Friend Forgot |Dead) = $(P(DF) * P(F))/P(D)$
$= (0.8 * 0.3)/P(D|W) * P(W) + P(D|W') * P(W')$
$(0.8 * 0.3)/0.2 * 0.7 + 0.8 * 0.3$
$= 0.24/0.38$
$= 0.62$
The probability that my friend forgot to water (given the plant died) is 0.62

Problem 4

Assuming conditional independence of G and B given the Naive Bayes assumption both,
P(G=g,B=b $|D = +$)

$P(D = -|G = g, B = b)$

can be split, and the formula becomes:
$P(D = -|G = g, B = b) = (P(G = g|D = -) * P(B = b|D = -)) * P(D = -)/P(G = g, B = b)$

(b c) The classifier and trained data are written in the provided **matthewmccaughan-nb-trainerandclassifier.py** file

(d) The classifier outputs an accuracy of about 92.31 percent

(e) The standardization of the data is not completely necessary as the data provided does not indicate any biases or data skewing that would require the model to be standardized.

(f) The data seems to reflect reality reasonably. This data would be more reflective of reality if its features were less limited. More data points and inputs would likely be required in reality to determine a diabetes diagnosis, also if

Problem 5

(a) Let Y be a random vector defined by $\vec{Y} = A\vec{X} + \vec{b}$. Express E[Y] and COV [Y,Y] in terms of E[X] and COV [X, X]:

for expressing E[Y] in terms of E[X]:
$\vec{Y} = A\vec{X} + \vec{b}$.
$E[Y] = E[AX + b]$.

Because of the linearity of expectation, A becomes a coefficient of the expectation and b becomes a constant:
$E[Y] = A * E[X] + b$

for expressing COV[Y,Y] in terms of COV[X,X]:
COV[Y,Y] = COV[AX+b,AX+b]
Because of the linearity of variance, the constant b can be dropped and not considered for the covariance:
COV[Y,Y] = COV[AX,AX]
(b) Design A and b to whiten Y. i.e. E[Y] = 0 and COV [Y, Y] = I
For b:
$E[Y] = 0$
$A * E[X] + b = 0$
Solving for b:
b = -A*E[X]

For A:
$COV[Y, Y] = I$
$COV[AX, AX] = I$
after decomp:
$A * COV[X, X] * A^T = I$
$A = I/(COV[X, X] * A^T)$