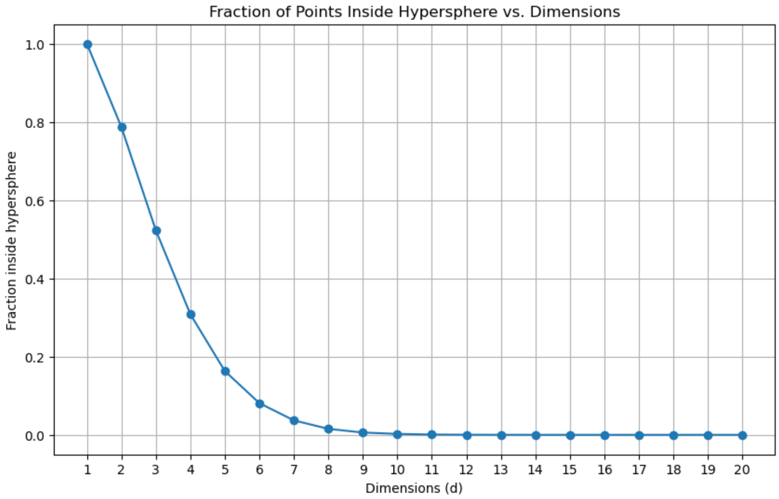
Assignment 1 – Sahil Adwani

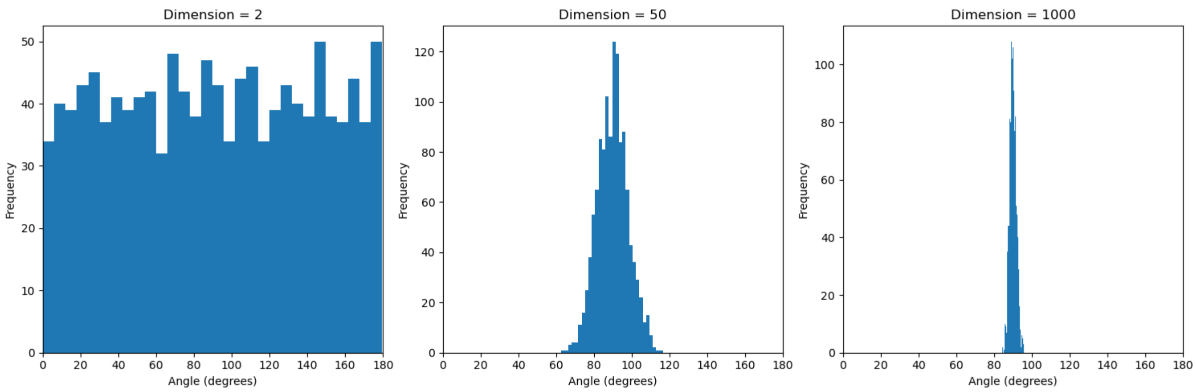
1.1)



The plot shows the fraction of vectors inside a d-dimensional hypersphere of radius 1 as a function of the number of dimensions d. As the number of dimensions increases, the fraction of vectors inside the hypersphere decreases, demonstrating the curse of dimensionality. So, when the dimension is 1, the fraction inside the hypersphere is 1.0. This means that in a one-dimensional space, every vector generated uniformly will always fall within the hypersphere, as the hypersphere and the hypercube, in this case, are the same interval.

In higher dimensions, the space's volume grows much faster than the hypersphere's volume. This means that as dimensions increase, the hypersphere occupies a smaller part of the entire space. Consequently, randomly generated points are less likely to fall inside the hypersphere. As stated, this is known as the curse of dimensionality, where high-dimensional spaces become sparse, making it difficult to find points within certain regions, like the hypersphere in this case.

1.2)



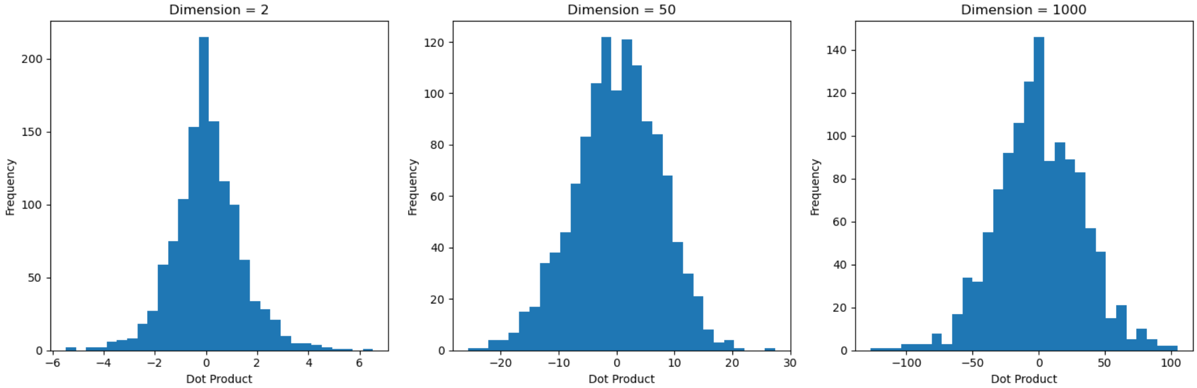
In low-dimensional spaces, such as 2D, vectors have more freedom to point in various directions, leading to a wider range of angles between them. The angles between vectors tend to be more uniformly distributed as vectors can span a larger portion of the space.

As dimensions increase to 50D, the angles between vectors become less spread out and most angles converge towards 90 degrees due to increased orthogonality. However, there is still some variation, with angles ranging from around 60 degrees to about 120 degrees.

By the time dimensions reach 1000D, vectors are almost always nearly orthogonal to each other. This occurs because high-dimensional spaces push vectors towards orthogonality due to their geometric properties. Consequently, in very high dimensions, the histogram of angles shows that most angles are very close to 90 degrees, reflecting the nearly orthogonal nature of vectors in such spaces.

So, as dimensionality increases, vectors tend to become more orthogonal, resulting in a histogram with angles clustered around 90 degrees.

1.3)



Yes, the trend is different.

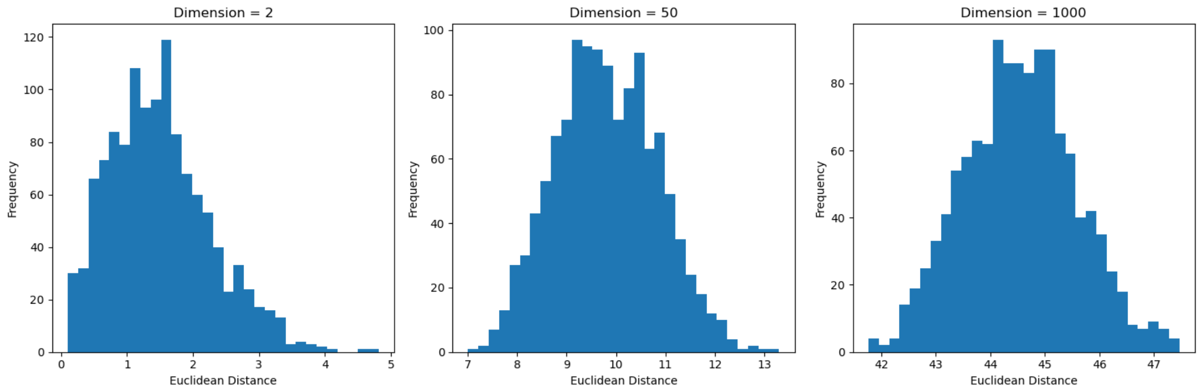
In low dimensions (e.g., 2D), vectors are easier to visualise, and their alignments are straightforward. The dot product reflects how much one vector points in the direction of another, resulting in a more constrained range of values. This limited dimensionality means fewer ways for vectors to align, leading to less variability and a narrower spread of dot product values. Most pairs of vectors are still perpendicular.

As dimensionality increases, such as to 50D, the number of components in vectors grows, allowing for more complex alignments. This results in a broader range of dot product values and increased variability. The vectors have more degrees of freedom, making the dot product more variable and showing a wider spread in the distribution.

In very high dimensions (e.g., 1000D), the dot product values often exhibit a broader range compared to lower dimensions. Despite most random vectors being nearly orthogonal and clustering around zero, the histogram can still show a broader spread with a sharper peak at zero. This occurs because, with many dimensions, even small changes in vector components can cause significant variations in the dot product, though the peak around zero is sharper due to high orthogonality. This reflects the "curse of dimensionality," where vectors become predominantly orthogonal, leading to a sharper peak around zero

As dimensionality increases, the spread of dot product values broadens.

1.4)



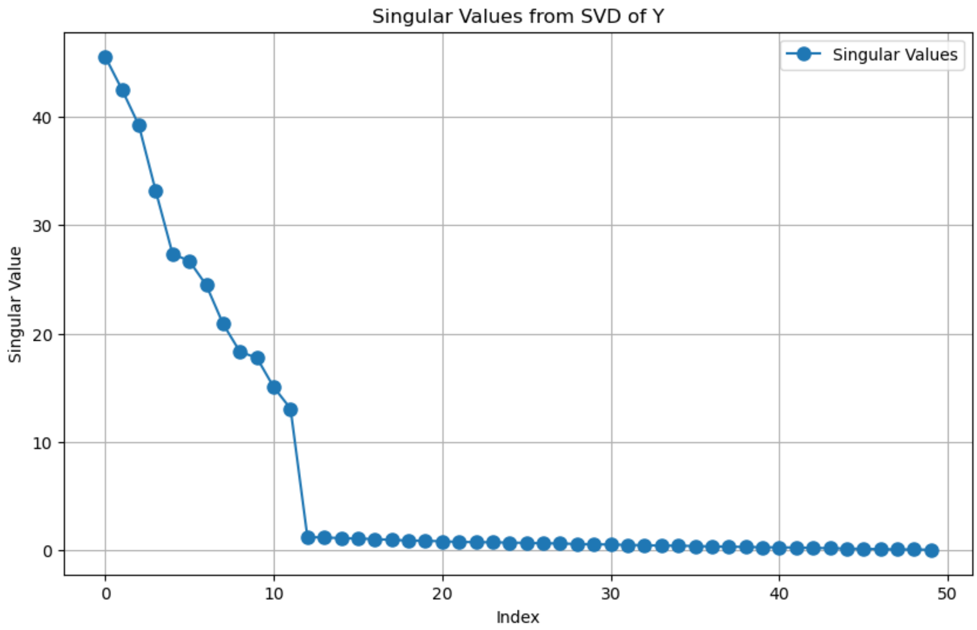
The trend is different. As dimensionality increases, the trends in Euclidean distances between vectors shift significantly. In low dimensions (e.g., 2D/3D), vectors are relatively close to each other due to the compact nature of the space, resulting in a wide range of distances. The distribution of these distances is often slightly right-skewed, with shorter distances being more common and greater variability due to fewer constraints in vector arrangement.

As we move to medium dimensions (e.g., 50D), the distances between vectors increase and start to follow a more normal (Gaussian) distribution. This shift occurs because the space expands, causing vectors to become more uniformly distributed. The central limit theorem suggests that with many dimensions, the sum of squared differences tends to form a normal distribution, concentrating distances around a mean value.

In very high dimensions (e.g., 1000D), distances between vectors become significantly larger and more concentrated around a specific mean value. The distribution becomes even more normal with a narrower range due to the increased likelihood of vectors being nearly orthogonal. This uniform distribution results in reduced variability and a higher average distance. The curse of dimensionality causes distances to converge around a central value, and the central limit theorem applies to the sum of squared differences, leading to distances clustering around the mean. The increased sparsity of data in high dimensions means that vectors are spread out more widely hence the Euclidean distance increases. This is proven by the distances being bigger then 2D and 50D as in 1000D, distance ranges from 42 to 47 roughly.

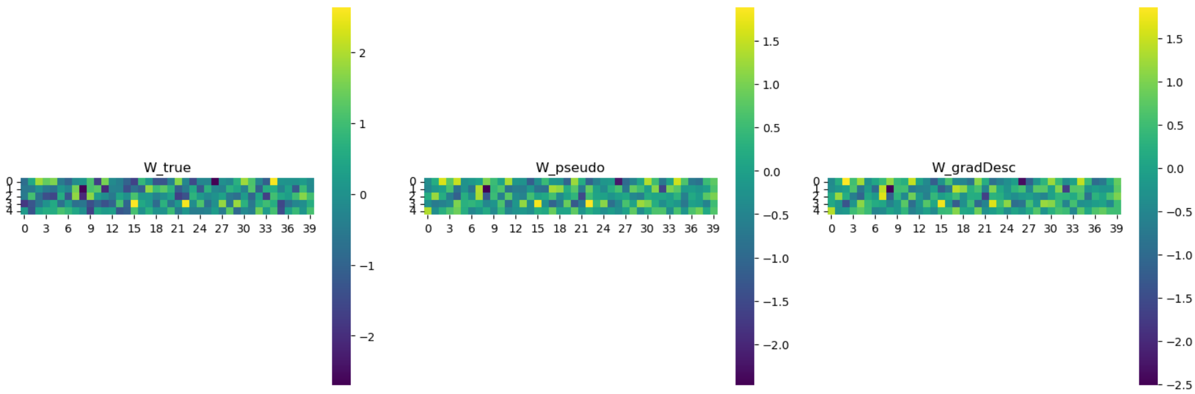
Overall, as dimensionality increases, Euclidean distances generally become larger, more uniformly distributed, and less variable, reflecting the expansive nature of high-dimensional spaces and the effects of the curse of dimensionality.

2.1)



The singular values exhibit a clear pattern where the first 12 values decrease gradually, suggesting that these dimensions capture the majority of the data's variance and important information. These values are much larger compared to the rest, indicating that the data is effectively represented within these 12 dimensions. After these 12 points, there is a sharp drop in singular values, which then levels off into a flat tail. This sharp drop signifies that the remaining dimensions contribute minimally to the overall variance and add little new information. The flat tail that follows reflects that these additional dimensions mainly capture noise or less significant variations rather than meaningful data. Overall, this pattern suggests that the data is well-represented by approximately 12 dimensions, with the additional dimensions providing minimal value and mostly representing noise or minor details.

2.2)



The heatmap of W_true shows the true relationship between input features and output features. I see different patterns of weights across the heatmap symbolising the complexity of the model. Both W_pseudo and W_gradDesc have given similar weights shown by row index 2 and column index 8 have a very dark blue colour (like black) being around -2. And this is the same for W_true meaning it classified weights correctly. But both W-pseudo and W_gradDesc have also have some misclassification as W_true seems to look more darker with more blue and green whilst W_pseudo and W_gradDesc look more yellow and light green indicating a difference in weight distribution. Both W_gradDesc and W_pseudo look similar to each other but differ somewhat from W_true. These differences could be due to the specific way the model was trained such as noise in training data, regularization and other factors. W_pseudo was transposed due to the way the matrices are structured and was inverse and it was necessary to align with the structure of W_true and ensure a proper comparison.

3.1)

1)

X	Y	Z	P(X,Y,Z)
0	0	0	0.0245
0	0	1	0.0105
0	1	0	0.0630
0	1	1	0.2520
1	0	0	0.2730
1	0	1	0.1170
1	1	0	0.0520
1	1	1	0.2080

Working Below:

Marginal Probability P(X):

- P(X=0)=0.35
- P(X=1)=0.65

Conditional Probability P(Y|X):

- P(Y=0|X=0)=0.10
- P(Y=1|X=0)=0.90
- P(Y=0|X=1)=0.60
- P(Y=1|X=1)=0.40

Conditional Probability P(Z|Y):

- P(Z=0|Y=0)=0.70
- P(Z=1|Y=0)=0.30
- P(Z=0|Y=1)=0.20
- P(Z=1|Y=1)=0.80

Independence Assumption: Z and X are conditionally independent given Y.

Product Rule for Joint Probability

The joint probability P(X,Y,Z) is given by:

P(X,Y,Z)=P(X)P(Y|X)×P(Z|Y)

Computation Steps:
For X = 0, Y = 0, Z = 0:
$P(X=0,Y=0,Z=0)=P(X=0)\times P(Y=0|X=0)\times P(Z=0|Y=0) = 0.35 \times 0.10 \times 0.70 = 0.0245$

For X = 0, Y = 0, Z = 1:
$P(X=0,Y=0,Z=1)=P(X=0)\times P(Y=0|X=0)\times P(Z=1|Y=0) = 0.35 \times 0.10 \times 0.30 = 0.0105$

For X = 0, Y = 1, Z = 0:
$P(X=0,Y=1,Z=0)=P(X=0)\times P(Y=1|X=0)\times P(Z=0|Y=1) = 0.35 \times 0.90 \times 0.20 = 0.0630$

For X = 0, Y = 1, Z = 1:
$P(X=0,Y=1,Z=1)=P(X=0)\times P(Y=1|X=0)\times P(Z=1|Y=1) = 0.35 \times 0.90 \times 0.80 = 0.2520$

For X = 1, Y = 0, Z = 0:
$P(X=1,Y=0,Z=0)=P(X=1)\times P(Y=0|X=1)\times P(Z=0|Y=0) = 0.65 \times 0.60 \times 0.70 = 0.2730$

For X = 1, Y = 0, Z = 1:
$P(X=1,Y=0,Z=1)=P(X=1)\times P(Y=0|X=1)\times P(Z=1|Y=0) = 0.65 \times 0.60 \times 0.30 = 0.1170$

For X = 1, Y = 1, Z = 0:
$P(X=1,Y=1,Z=0)=P(X=1)\times P(Y=1|X=1)\times P(Z=0|Y=1) = 0.65 \times 0.40 \times 0.20 = 0.0520$

For X = 1, Y = 1, Z = 1:
$P(X=1,Y=1,Z=1)=P(X=1)\times P(Y=1|X=1)\times P(Z=1|Y=1) = 0.65 \times 0.40 \times 0.80 = 0.2080$

2)

X	Y	P(X,Y)
0	0	0.035
0	1	0.315
1	0	0.390
1	1	0.260

Working Below:
Product Rule Application

- The joint probability P(X,Y) can be computed using:
  $P(X,Y) = P(X) \times P(Y|X)$
- P(X) is the marginal probability of X.
- P(Y|X) is the conditional probability of Y given X.

Given Data

- Marginal Probability P(X):

  $P(X=0)=0.35$

  $P(X=1)=0.65$

- Conditional Probability P(Y|X):

  $P(Y=0|X=0)=0.10$

P(Y=1|X=0)=0.90

- P(Y=0|X=1)=0.60
- P(Y=1|X=1)=0.40

Calculation Steps:

P(X=0,Y=0)=P(X=0)×P(Y=0|X=0) =0.35x0.10=0.035

P(X=0,Y=1)=P(X=0)×P(Y=1|X=0=0.35x0.90=0.315

P(X=1,Y=0)=P(X=1)×P(Y=0|X=1) =0.65x0.60=0.390

P(X=1,Y=1)=P(X=1)×P(Y=1|X=1) =0.65x0.40=0.260

3)

a) Fining P(Z=0), summing joint probabilities over all values of X and Y:

P(Z=0)=$\sum_{X,Y}$P(X,Y,Z=0) //using sum rule

P(Z=0)=P(X=0,Y=0,Z=0)+P(X=0,Y=1,Z=0)+P(X=1,Y=0,Z=0)+P(X=1,Y=1,Z=0)

P(Z=0)=0.0245+0.0630+0.2730+0.0520 = **0.4125**

b) Finding P(X=0,Z=0), summing the joint probabilities over all values of Y:

P(X=0,Z=0)=$\sum_{Y}$P(X=0,Y,Z=0) //using sum rule

P(X=0,Z=0)=P(X=0,Y=0,Z=0)+P(X=0,Y=1,Z=0)

P(X=0,Z=0)=0.0245+0.0630 = **0.0875**

c)

Using the definition of conditional probability:

P(X=1,Y=0|Z=1)= $\frac{P(X=1,Y=0,Z=1)}{P(Z=1)}$

First, getting P(Z=1):

P(Z=1)= $\sum_{X,Y}$P(X,Y,Z=1) // using sum rule

P(Z=1)=P(X=0,Y=0,Z=1)+P(X=0,Y=1,Z=1)+P(X=1,Y=0,Z=1)+P(X=1,Y=1,Z=1)

P(Z=1)=0.0105+0.2520+0.1170+0.2080=0.5875

Getting the conditional probability:
//using the table of the joint distribution P(X,Y,Z) to get P(X=1,Y=0,Z=1) which = 0.1170

P(X=1,Y=0|Z=1)= P(X=1,Y=0,Z=1)/ P(Z=1)

P(X=1,Y=0|Z=1)= 0.1170/0.5875 ≈ **0.199**

d)

Using the definition of conditional probability:

P(X=0|Y=0,Z=0)= $\frac{P(X=0,Y=0,Z=0)}{P(Y=0,Z=0)}$

Finding P(Y=0,Z=0):

P(Y=0,Z=0)=$\sum_X$P(X,Y=0,Z=0) //using sum rule

P(Y=0,Z=0)=P(X=0,Y=0,Z=0)+P(X=1,Y=0,Z=0)

P(Y=0,Z=0)=0.0245+0.2730=0.2975

Now getting the conditional probability:

P(X=0|Y=0,Z=0)= $\frac{P(X=0,Y=0,Z=0)}{P(Y=0,Z=0)}$

P(X=0|Y=0,Z=0)= 0.0245/0.2975 ≈ **0.082**

3.2)

1) P(B=t,C=t)=P(B=t|C=t)×P(C=t) = 0.2x04=0.08 //using product rule

2) P(A=f|B=t) = 1−P(A=t|B=t) = 1-0.3=0.7 //using complement rule

3) //Given that A and B are conditionally independent given C:
P(A=t,B=t|C=t) = P(A=t|C=t)×P(B=t|C=t) = 0.5x 0.2 = 0.1

4) The statement that A and B are conditionally independent given C means:
P(A=t|B=t,C=t)= P(A=t|C=t) = 0.5

5)
//using conditional independence from Q3
P(A=t,B=t|C=t)=P(A=t|C=t)×P(B=t|C=t) = 0.5 x 0.2 = 0.1

P(A=t,B=t,C=t)=P(A=t,B=t|C=t)×P(C=t) = 0.1 x 0.4 = 0.04 //using product rule

Bonus)
P(A=t|B=f): Using the law of total probability. The formula is:

P(A=t|B=f)=P(A=t|B=f,C=t)×P(C=t|B=f)+P(A=t|B=f,C=f)×P(C=f|B=f)

Hence the conditional probabilities being P(A=t|B=f,C=t) and P(A=t|B=f,C=f) as well as marginal probabilities being P(C=t|B=f) and P(C=f|B=f) are needed.

Without the values for P(C=t|B=f) and P(A=t|B=f,C=t), we cannot complete this calculation. Thus, additional information is needed to determine P(A=t|B=f).

3.3)

1)

Counts: Apple = 30, Banana = 10, Orange = 10

- Total counts N: 30+10+10=50
- Number of categories k: 3
- Smoothing parameter α: 1

Applying the formula:

For Apple:

P(Apple)= (30+1)/(50+1x3) = 31/53 ≈ 0.5849
P(Banana) = (10+1)/(50+1x3) = 11/53 ≈ 0.2075
P(Orange) = (10+1)/(50+1x3) = 11/53 ≈ 0.2075

2) $H(X)=-(P(Apple)\times \log_2(P(Apple))+P(Banana)\times \log_2 P(Banana))+(P(Orange)\times \log_2(P(Orange)))$

$H(X)=-(31/53\times\log_2(31/53))+11/53\times\log_2(11/53))+11/53\times\log_2(11/53))) \approx 1.39412$
//using calculator

3) $\text{Log-loss}=-1/50(30\times\log_2(P(Apple))+10\times \log_2(P(Banana))+10\times \log_2(P(Orange)))$

$\text{Log-loss}=-1/50(30\times\log_2(31/53)+10\times \log_2(P(11/53))+10\times \log_2 (P(11/53)))= 1.371630021$
//already calculated probabilities of the fruit above

4)
Entropy and log-loss measure different aspects but are closely related. Entropy quantifies the uncertainty or randomness of the probability distribution itself, measuring the average amount of "information" or "surprise" associated with the distribution.

Log-loss, on the other hand, measures how well a probabilistic model's predictions match the true distribution. It quantifies the average "penalty" for incorrect predictions.

When using the smoothed probabilities as predictions, the entropy of the smoothed distribution closely approximates the log-loss because both metrics use the same smoothed probabilities. Log-loss effectively evaluates the same distribution as entropy but focuses on how the predictions (in this case, the smoothed probabilities) compared to the actual data.

However, log-loss considers the actual counts and their distribution in the dataset, which introduces a slight difference. Entropy assumes a perfect model that knows the true distribution, whereas log-loss incorporates the mismatch between the predictions and the actual data counts. Thus, while they are closely related and often yield similar values, they are not exactly the same due to these differences in focus and calculation.

4) Naïve Bayes

1) Prior to Any Observations:
P(mum in town = true) = 0.01
P(mum in town=false)=0.99

2) after seeing there are many txts on your phone;

P(many txt = true | mum in town = true) = 0.7

P(many txt = true | mum in town = false) = 0.1

P(many txt = true) = 0.106 = 0.11

$$\frac{P(mum\ in\ town = true\mid many\ txt = true)}{P(mum\ in\ town = false\mid many\ txt = true)} = \frac{P(mum\ in\ town = true)}{P(mum\ in\ town = false)} \text{ x } \frac{\text{P(many txt = true | mum in town = true)}}{\text{P(many txt = true | mum in town = false)}}$$

= 0.01/0.99 x 0.7/.1 = 7/99

P(mum in town = true | many txt = true) = 7 / (7 + 99) = 7/106 = 0.066037 = 0.7 = 70%

3)

P(organised sock draw=true |mum in town=true)=0.75

P(organised sock draw=true |mum in town=false)=0.05

$$\frac{P(mum\ in\ town = true\mid many\ txt = true,\ \text{organised sock draw = true})}{P(mum\ in\ town = false\mid many\ txt = true,\ \text{organised sock draw = true})} = \frac{P(mum\ in\ town = true)}{P(mum\ in\ town = false)} \text{ x } \frac{\text{P(many txt = true | mum in town = true)}}{\text{P(many txt = true | mum in town = false)}} \text{ x } \frac{p(organised\ sock\ draw = true\mid mum\ in\ town = true)}{P(organised\ sock\ draw = true\mid mum\ in\ town = false)}$$

= 0.01/0.99 x 0.7/.1 x 0.75/ 0.05 = 35/33

P(mum in town = true | many txt = true, organised sock draw = true) = 35 / (33 + 35) = 35/68 = 0.51470 = 0.51 = 51%

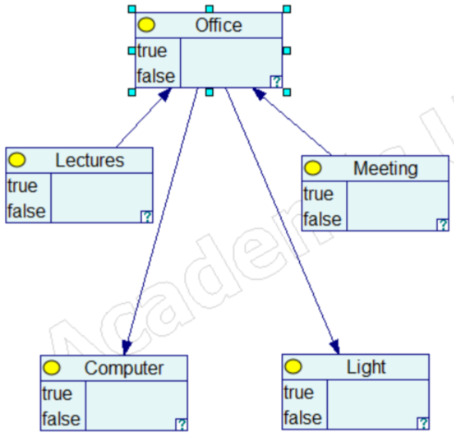
4)

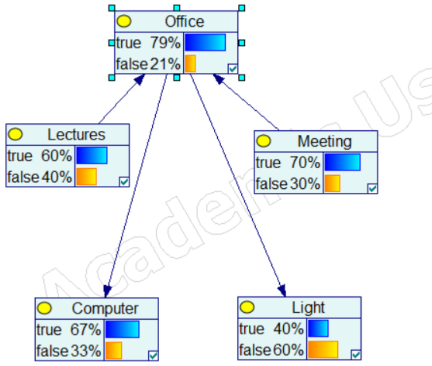
$$\frac{P(mum\ in\ town = true\mid many\ txt = true,\ \text{organised sock draw = true, put away } dishes = true)}{P(mum\ in\ town = false\mid many\ txt = true,\ \text{organised sock draw = true, put away dishes=true})} = \frac{P(mum\ in\ town = true)}{P(mum\ in\ town = false)} \text{ x } \frac{\text{P(many txt = true | mum in town = true)}}{\text{P(many txt = true | mum in town = false)}} \text{ x } \frac{p(organised\ sock\ draw = true\mid mum\ in\ town = true)}{P(organised\ sock\ draw = true\mid mum\ in\ town = false)} \text{ x } \frac{p(put\ away\ dishes = true\mid mum\ in\ town = true)}{P(put\ away\ dishes = true\mid mum\ in\ town = false)}$$

= 0.01/0.99 x 0.7/.1 x 0.75/ 0.05 x 0.1/0.1= 35/33

P(mum in town = true | many txt = true, organised sock draw = true, put away dishes =true) = 35 / (33 + 35) = 35/68 = 0.51470 = 0.51 = 51%

5.1)





In the constructed Bayesian Network, the total number of free parameters is determined by evaluating each node's requirements.

- For the binary nodes Meetings and Lectures, each requires 1 free parameter, representing the probability of one state. The probability of the other state is automatically determined since the probabilities must sum to 1.
- The Office node, which depends on Meetings and Lectures (2 parent nodes), has 4 possible parent state combinations, each requiring 1 parameter to define the probability of being in the office. Thus, Office has 4 free parameters.
- The Light and Computer nodes each depend on Office, and with 2 states for Office, each node requires 1 parameter per state of Office. This results in 2 free parameters for Light and 2 for Computer.

  Total Free Parameters:
  Meetings: 1
  Lectures: 1
  Office: 4
  Light: 2
  Computer: 2
  Total = 1 + 1 + 4 + 2 + 2 = 10

- Summing these, the total number of free parameters in the Bayesian Network is 10. This total encompasses the parameters needed to specify both the prior probabilities for nodes without parents and the conditional probabilities for nodes with parents.

5.2)

// ‘M’ means Meeting, ‘C’ means Computer, ‘Light” means Light, ‘Lectures’ means Lectures and ‘O’ means Office

P(Lectures=True, M=False, O=True, C=True, Light=False) =
P(Lectures=True) × P(M=False) × P(O=True|M=False,Lectures=True) × P(C=True|O=True) × P(Light=False|O=True)
=0.60×0.30×0.90×0.80×0.50 = 0.0648 = 6.48%
//using product rule and seeing what nodes have arrows going in and out

So, the probability that Leo has lectures, has no meetings, is in his office, is logged on his computer, and has the lights off is 0.0648 = 6.48%.

5.3)

Using the prior probability of Leo having meetings.
Not knowing the values of any other features, the probability that Leo has meetings is simply 0.70 = 70%.

5.4)

Given Information:
Condition Probabilities:
P(O=True|M=T, Lectures =T) = 0.95
P(O=True|M=T, Lectures =F) = 0.75
P(O=True|M=F, Lectures =T) = 0.90
P(O=True|M=F, Lectures =F) = 0.16

Prior Probabilities:
P(M=True) = 0.70
P(M=False) = 0.30
P(Lectures=True) = 0.60
P(Lectures=False) = 0.40

Calculating the joint probability for each combination of Meetings and Lectures:

- **Combination 1**: Leo has meetings and lectures
  Probability=P(O=True|M=True,Lectures=True)×P(M=True)×P(Lectures=True)
  = 0.95×0.70×0.60=0.399
- **Combination 2**: Leo has meetings but no lectures
  Probability=P(O=True|M=True,Lectures=False)×P(M=True)×P(Lectures=False)
  =0.75×0.70×0.40=0.21
- **Combination 3**: Leo has lectures but no meetings
  Probability=P(O=True|M=False,Lectures=True)×P(M=False)×P(Lectures=True)
  =0.90×0.30×0.60=0.162
- **Combination 4**: Leo has neither meetings nor lectures
  Probability=P(O=True|M=False,Lectures=False)×P(M=False)×P(Lectures=False)
  =0.16×0.30×0.40=0.0192

P(O=True)=0.399+0.21+0.162+0.0192=0.7902 = 79.02% = 79%

The probability that Leo is in the office, regardless of whether he has meetings or lectures, is 0.7902 = 79.02%. This is calculated by averaging the probability of him being in the office over all possible combinations of meetings and lectures.

5.5)

Conditional Probabilities
$P(C=True|O=True)=0.80$
$P(Light=False|O=True)=0.50$

We assume that whether Leo is logged on his computer and whether his light is off are independent events given that he is in the office

$P(C=True,Light=False|O=True)=P(C=True|O=True)\times P(Light=False|O=True)$
Compute the Joint Probability
$P(C=True,Light=False|O=True)=0.80\times 0.50=0.40$ = 40%

The probability that Leo is logged on his computer (True) and has his light off (False) given that he is in his office is 0.40 = 40%.

5.6)

//Using GeNIe to prove this:

so Prior probability of light
P(Light=True) = 0.40

When a student knows that Leo has no lectures
P(Light=True| Lectures=False) = 0.30

When a student knows that Leo has no Lectures and Logged on (Computer True)
P(Light=True| Lectures=False, Computer = True) = 0.43

Observing that Leo is logged on (Computer=True) has influenced the student's belief about whether Leo's light is on. Initially, the probability of the light being on, given that Leo has no lectures, was 30%. With the new information that Leo is logged in, this probability increases to 43%. This illustrates how additional evidence can significantly alter our beliefs about related events. Also, this updated belief of 43% is also 3% higher than the prior probability of 40% that the light was on, showing an increase in the student's belief based on the new evidence.

More calculations to prove this:
P(Light=True| Lectures=True, Computer = True) = 0.49 = 49%
P(Light=True| Lectures=True, Computer = False) = 0.40 = 40%
P(Light=True| Lectures=False, Computer = True) = 0.43 = 43%
P(Light=True| Lectures=False, Computer = False) = 0.16 = 16%

This demonstrates that if either condition having lectures or being logged on (Computer = True) is true, the probability of the light being on increases significantly from 16% (when neither condition is true) to approximately 40%. If both conditions are true, the probability increases further to 49%. This highlights that the student's confidence in the light being on increases notably with either one of the conditions being true, with the highest confidence achieved when both conditions are met.