

Clustering

Exercise 1

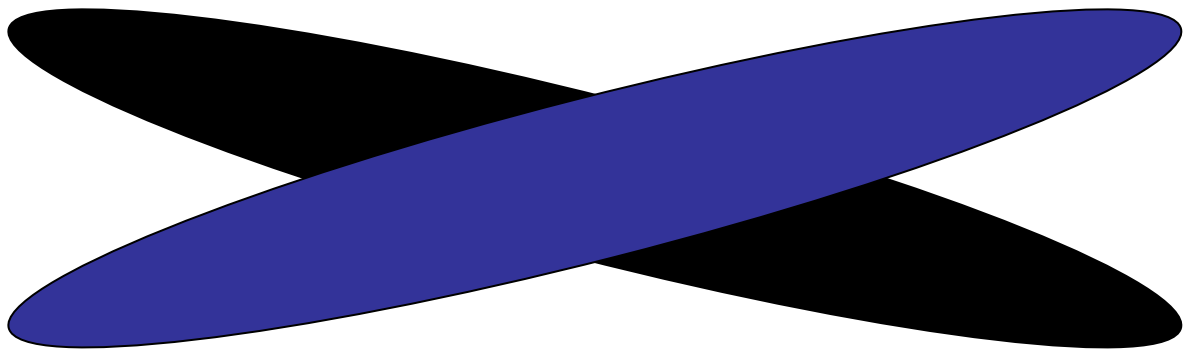
The main difference is that GMMs are probabilistic whereas k-means are not. This gives GMMs a greater flexibility since the covariance matrices Σ_k of the Gaussians may be anything we want and hence the clusters may have any desired elliptical shapes.

It can be shown that k-means are a special case of GMMs where a) the covariance matrix Σ is spherical (ie $\Sigma = \lambda \mathbf{I}$ with \mathbf{I} the identity matrix), b) λ tends to infinity, and c) Σ is shared between clusters. Note that for our purposes λ does not matter.

Sharing the covariance matrix between clusters means that only the location of the cluster counts, so that the boundaries between clusters are linear, whereas in GMMs they can be quadratic.

Having a spherical covariance matrix means that the clusters are round.

A good dataset would be 2 elongated Gaussians forming a cross, ie overlapping, such as



Such a dataset should not be difficult to generate using the function `rnorm` in R. The function to apply k-means is `kmeans`. The function to apply the EM algorithm is the function `Mclust` from the `mclust` package (see slides).

Linear Models

Exercise 1

Question 1

The problem using least squares is that we try to minimise the distance between $y(x)$ and our target $t \in \{0,1\}$. However, in the dataset we have, although the data is linearly separable, there is a cluster that is far away from the rest.

If we draw the real decision boundary, we can see that for any x in the outlier cluster, $|y(x)|$ is going to be larger than for points of other clusters. Hence least squares is going to try to compensate by pushing the decision boundary towards the outlier cluster.

Question 2

This is an ugly solution (no colours, legends or anything) but it gives you an idea.

```
x1=10+10*rnorm(200); # 1st dimension for points of class 1
y1=2*x1+20+10*rnorm(200); # 2nd dimension for points of class 1
x2=10+10*rnorm(300); # 1st dimension for points of class 2
y2=x2-15+10*rnorm(300); # 2nd dimension for points of class 2
outliers=matrix(0,nrow=60,ncol=2); # class 2 outliers
outliers[,1]=35+5*rnorm(60); # 1st dimension for outliers
outliers[,2]=-80+10*rnorm(60); # 2nd dimension for outliers

X=matrix(0,nrow=length(x1)+length(x2)+nrow(outliers),ncol=2)
X[,1]=c(x1,x2,outliers[,1]); # complete data – 1st dimension
X[,2]=c(y1,y2,outliers[,2]); # complete data – 2nd dimension
t=vector()
for(i in 1:length(x1)) t=c(t,1); # labels for class 1
for(i in 1:(length(x2)+nrow(outliers))) t=c(t,0); # labels for class 2
tls= vector()
for(i in 1:length(x1)) tls=c(tls,1); # labels for class 1
for(i in 1:(length(x2)+nrow(outliers))) tls= c(tls,-1); # labels for class 2
```

```
# least squares
```

```
lsmode1=lm(tls~X)
abs=seq(-20,50,by=0.5)
ord=(-lsmode1$coefficients[1]-abs*lsmode1$coefficients[2])/
lsmode1$coefficients[3]
windows(); plot(X[,1],X[,2])
lines(abs,ord)
```

```
# logistic regression
```

```
lrmodel=glm(t~X,family=binomial)
abs=seq(-20,50,by=0.5)
ord=(-lrmodel$coefficients[1]-abs*lrmodel$coefficients[2])/
lrmodel$coefficients[3]
windows(); plot(X[,1],X[,2])
lines(abs,ord)
```

Exercise 2

1) Dual formulation

When we saw kernel regression, we rewrote \mathbf{w} as a linear combination of the training points $\mathbf{X} = \{\mathbf{x}_n\}$, so that $\mathbf{w}^T \mathbf{y}$ became a linear combination of $\mathbf{x}_n^T \mathbf{y}$. To kernelise PCA, you have to find where the kernel may replace the dot product $\mathbf{x}_n^T \mathbf{y}$ in the equations.

In PCA, we want to find a new vector \mathbf{u} , the principal component, and project our data on it

$$y = \text{transformed}(\mathbf{x}) = \mathbf{x}^T \mathbf{u}$$

So now imagine that we write \mathbf{u} as a linear combination of the training points \mathbf{X}

$$\mathbf{u} = \sum_{n=1}^N a_n \mathbf{x}_n = \mathbf{X}^T \mathbf{a}$$

This gives us:

$$y = \text{transformed}(\mathbf{x}) = \mathbf{x}^T \left(\sum_{n=1}^N a_n \mathbf{x}_n \right) = \sum_{n=1}^N a_n \mathbf{x}^T \mathbf{x}_n$$

We can now replace the dot product with any kernel we like ☺:

$$y = \text{transformed}(\mathbf{x}) = \sum_{n=1}^N a_n K(\mathbf{x}, \mathbf{x}_n)$$

Now instead of learning the principal component \mathbf{u} , we learn the coefficients \mathbf{a} . Remember that \mathbf{u} is learnt to maximise the variance and we showed that \mathbf{u} has to be an eigenvector of the correlation matrix.

2) First point of view

The variance is

$$\begin{aligned} \sum_{n=1}^N y_n^T y_n &= \sum_{n=1}^N (\mathbf{x}_n^T \mathbf{u})^T (\mathbf{x}_n^T \mathbf{u}) = \sum_{n=1}^N \mathbf{u}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{u} \\ &= \mathbf{u}^T \left(\sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{u} = \mathbf{u}^T (\mathbf{X}^T \mathbf{X}) \mathbf{u} = (\mathbf{X}^T \mathbf{a})^T (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{a}) \\ &= \mathbf{a}^T \mathbf{X} \mathbf{X}^T \mathbf{X} \mathbf{X}^T \mathbf{a} = \mathbf{a}^T (\mathbf{X} \mathbf{X}^T) (\mathbf{X} \mathbf{X}^T) \mathbf{a} = \mathbf{a}^T (\mathbf{X} \mathbf{X}^T)^2 \mathbf{a} \\ &= \mathbf{a}^T \mathbf{K}^2 \mathbf{a} \end{aligned}$$

We want the resulting \mathbf{u} to be normalised

$$\mathbf{u}^T \mathbf{u} = 1 \Leftrightarrow (\mathbf{X}^T \mathbf{a})^T (\mathbf{X}^T \mathbf{a}) = 1 \Leftrightarrow \mathbf{a}^T \mathbf{X} \mathbf{X}^T \mathbf{a} = 1 \Leftrightarrow \mathbf{a}^T (\mathbf{X} \mathbf{X}^T) \mathbf{a} = 1 \Leftrightarrow \mathbf{a}^T \mathbf{K} \mathbf{a} = 1$$

Maximising the variance with respect to \mathbf{a} therefore means

$$\frac{d}{d\mathbf{a}} (\mathbf{a}^T \mathbf{K}^2 \mathbf{a} + \lambda (1 - \mathbf{a}^T \mathbf{K} \mathbf{a})) = 0 \Leftrightarrow 2\mathbf{K}^2 \mathbf{a} - 2\lambda \mathbf{K} \mathbf{a} = 0 \Leftrightarrow \mathbf{K}^2 \mathbf{a} = \lambda \mathbf{K} \mathbf{a}$$

$$\Leftrightarrow \mathbf{K} \mathbf{a} = \lambda \mathbf{a}$$

So, much in the same way, \mathbf{a} will be an eigenvector of the kernel matrix \mathbf{K} .

3) second point of view

If C is the covariance matrix, then

$$\mathbf{C}\mathbf{u} = \lambda\mathbf{u}$$

$$(\mathbf{X}^T\mathbf{X})\mathbf{u} = \lambda\mathbf{u}$$

$$(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{a}) = \lambda(\mathbf{X}^T\mathbf{a})$$

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{a}) = \lambda\mathbf{X}(\mathbf{X}^T\mathbf{a})$$

$$(\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T)\mathbf{a} = \lambda(\mathbf{X}\mathbf{X}^T)\mathbf{a}$$

$$(\mathbf{X}\mathbf{X}^T)^2\mathbf{a} = \lambda(\mathbf{X}\mathbf{X}^T)\mathbf{a}$$

$$(\mathbf{X}\mathbf{X}^T)\mathbf{a} = \lambda\mathbf{a}$$

$$\mathbf{K}\mathbf{a} = \lambda\mathbf{a}$$

So, much in the same way, \mathbf{a} will be an eigenvector of the kernel matrix \mathbf{K} .

4) Which eigen vector?

The variance will then be

$$\sum_{n=1}^N y_n^T y_n = \mathbf{a}^T \mathbf{K}^2 \mathbf{a} = \mathbf{a}^T \mathbf{K} (\mathbf{K}\mathbf{a}) = \lambda \mathbf{a}^T \mathbf{K}\mathbf{a} = \lambda$$

So, again, \mathbf{u} must be the eigen vector corresponding to the largest eigen value.

Probability theory

Exercise 1

Decipher the text

What are the random variables (ie the events)?

S: the ship sinks – then possible values for S are {0,1}

R: the route of the ship – possible values for R are {s,l} for short and long

What are the probabilities that are known?

“The danger of sinking is 10% with the short route” means that

$$p(S = 1 | R = s) = 0.1$$

“The danger of sinking is 5% with the long route” means that

$$p(S = 1 | R = l) = 0.05$$

“About 20% of the ships decided to take the shorter route” means that $p(R = s) = 0.2$

Question 1

What is the probability of a ship to sink?

$$\begin{aligned} p(S = 1) &= \sum_R p(S = 1, R) = \sum_R p(S = 1 | R) p(R) \\ &= p(S = 1 | R = s) p(R = s) + p(S = 1 | R = l) p(R = l) \\ &= 0.1 \times 0.2 + 0.05 \times (1 - 0.2) = 0.02 + 0.04 = 0.06 \end{aligned}$$

The expected number of sunken ships is then

$$p(S = 1) N_{\text{ships}} = 0.06 \times 200 = 12$$

Question 2

$$p(R = s | S = 1) = \frac{p(S = 1 | R = s) p(R = s)}{p(S = 1)} = \frac{0.1 \times 0.2}{0.06} = \frac{1}{3}$$

$$p(R = l | S = 1) = 1 - p(R = s | S = 1) = \frac{2}{3}$$

So we should check the long route!!

Exercise 2

Decipher the text

What are the random variables (ie the events)?

T: the test is positive – then possible values for T are {0,1}

D: you have the disease – possible values for D are {0,1}

What are the probabilities that are known?

“The test is 99% accurate” means that $p(T = 1 | D = 1) = 0.99$ and

$$p(T = 0 | D = 0) = 0.99$$

“This disease usually strikes only 1 in 10000 people” means that

$$p(D = 1) = 0.0001$$

Question 1

What you are concerned about is

$$\begin{aligned} p(D = 1 | T = 1) &= \frac{p(T = 1 | D = 1)p(D = 1)}{p(T = 1)} \\ &= \frac{p(T = 1 | D = 1)p(D = 1)}{p(T = 1 | D = 0)p(D = 0) + p(T = 1 | D = 1)p(D = 1)} \\ &= \frac{p(T = 1 | D = 1)p(D = 1)}{(1 - p(T = 0 | D = 0))(1 - p(D = 1)) + p(T = 1 | D = 1)p(D = 1)} \\ &= \frac{0.99 \times 0.0001}{0.01 \times 0.9999 + 0.99 \times 0.0001} = 0.009804 \end{aligned}$$

The test is essentially useless.

Question 2

The test would be much more useful if the accuracy was higher or if the disease was more common.

Exercise 3

Question 1

$$p(A, B | E) = \frac{p(A, B, E)}{p(E)} = \frac{p(A | B, E)p(B | E)p(E)}{p(E)} = p(A | B, E)p(B | E)$$

Question 2

$$p(A | B, E) = \frac{p(A, B, E)}{p(B, E)} = \frac{p(B | A, E)p(A | E)p(E)}{p(B | E)p(E)} = \frac{p(B | A, E)p(A | E)}{p(B | E)}$$

Exercise 4

Question 1

$$p(H | E_1, E_2) = \frac{p(E_1, E_2 | H)p(H)}{p(E_1, E_2)} \text{ so only the second set is sufficient.}$$

Question 2

“ E_1 and E_2 are conditionally independent given H ” means that

$$p(E_1, E_2 | H) = p(E_1 | H)p(E_2 | H)$$

So $p(H | E_1, E_2) = \frac{p(E_1, E_2 | H)p(H)}{p(E_1, E_2)} = \frac{p(E_1 | H)p(E_2 | H)p(H)}{p(E_1, E_2)}$ and the first and the second sets are now enough.

Moreover we can add that

$$p(H | E_1, E_2) = \frac{p(E_1 | H)p(E_2 | H)p(H)}{\sum_H p(E_1, E_2 | H)p(H)} = \frac{p(E_1 | H)p(E_2 | H)p(H)}{\sum_H p(E | H)p(E_2 | H)p(H)}^s$$

so the third set is also enough!

Bayesian networks

Exercise 1

Question 1

The joint distribution is the product of the distribution of each variable separately, conditioned on its parents.

Question 2

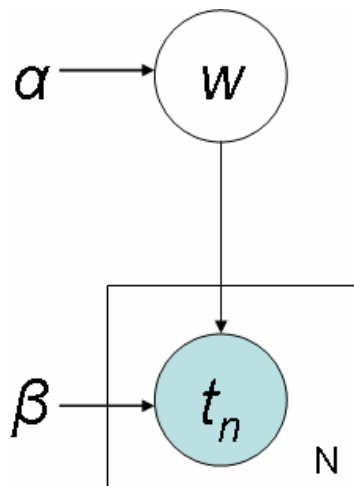
- represents the conditional independencies
- simplifies the expression of the joint distribution => compact representation, memory and time savings
- the graph is also helpful to “see” the model

Question 3

Learning is the process of computing the value of the parameters. Inference is the computation of the conditional distribution of the hidden variables given the observed ones.

Question 4

$$\begin{aligned} p(\mathbf{t}, \mathbf{w} \mid \alpha, \beta) &= p(\mathbf{w} \mid \alpha) \prod_{n=1}^N p(t_n \mid \mathbf{w}, \beta) \\ &= \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1} \mathbf{I}) \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^T \mathbf{x}_n, \beta^{-1}) \end{aligned}$$



Note here that, to really be Bayesian, β should also be a random variable and have a prior distribution with some hyperparameter b for example. β would then be circled (in white), we would add a non-circled parameter b , and there would be an arrow going from b to β , exactly like we have for α and \mathbf{w} . The joint distribution would be written

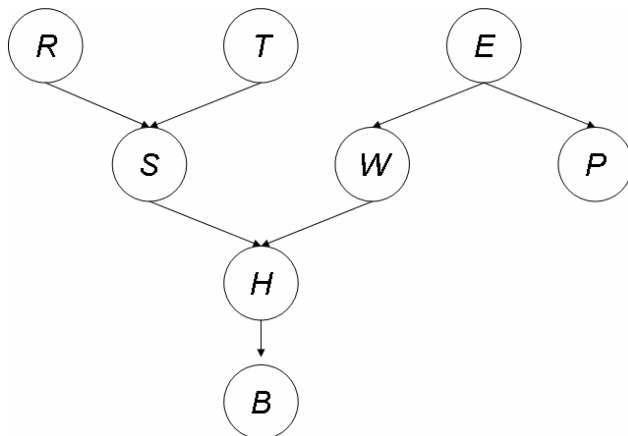
$$p(\mathbf{t}, \mathbf{w}, \beta | \alpha, b) = p(\mathbf{w} | \alpha) p(\beta | b) \prod_{n=1}^N p(t_n | \mathbf{w}, \beta)$$

However in the lecture we only considered \mathbf{w} for simplicity, so it is the network I am showing.

Exercise 2

See slides! ☺

Exercise 7



- R: rain $\in \{0,1\}$
- T: temperature $\in \{\text{cold,warm}\}$
- S: sick $\in \{0,1\}$
- E: exam $\in \{0,1\}$
- P: party $\in \{0,1\}$
- W: must work $\in \{0,1\}$
- H: must stay home $\in \{0,1\}$
- B: bored $\in \{0,1\}$

Question 1

The joint distribution is the product of the distribution of each variable separately, conditioned on its parents.

$$p(R, T, E, S, W, P, H, B) = p(R)p(T)p(E)p(S | R, T)p(W | E)p(P | E)p(H | S, W)p(B | H)$$

Question 2

$$p(R, T, E) = p(R)p(T)p(E)$$

You can convince yourself of this by writing:

$$p(R, T, E) = \sum_{S, W, P, H, B} p(R, T, E, S, W, P, H, B)$$

rewrite using the joint distribution given by the network

$$= \sum_{S, W, P, H, B} p(R)p(T)p(E)p(S | R, T)p(W | E)p(P | E)p(H | S, W)p(B | H)$$

take $p(R)p(T)p(E)$ out of the sum

$$= p(R)p(T)p(E) \sum_{S, W, P, H, B} p(S | R, T)p(W | E)p(P | E)p(H | S, W)p(B | H)$$

B is not used as a condition and is summed over \Rightarrow cancels

$$= p(R)p(T)p(E) \sum_{S, W, P, H, B} p(S | R, T)p(W | E)p(P | E)p(H | S, W) \underbrace{p(B | H)}_1$$

H is not used as a condition and is summed over \Rightarrow cancels

$$= p(R)p(T)p(E) \sum_{S, W, P, H} p(S | R, T)p(W | E)p(P | E) \underbrace{p(H | S, W)}_1$$

P is not used as a condition and is summed over \Rightarrow cancels

$$= p(R)p(T)p(E) \sum_{S, W, P} p(S | R, T)p(W | E) \underbrace{p(P | E)}_1$$

W is not used as a condition and is summed over \Rightarrow cancels

$$= p(R)p(T)p(E) \sum_{S, W} p(S | R, T) \underbrace{p(W | E)}_1$$

S is not used as a condition and is summed over \Rightarrow cancels

$$= p(R)p(T)p(E) \underbrace{\sum_S p(S | R, T)}_1$$

$$= p(R)p(T)p(E)$$

This means that R , T and E are marginally independent of each other.

$$p(W, P | E) = p(W | E)p(P | E)$$

You can convince yourself of this by writing:

$$p(W, P | E) = \frac{\sum_{R,T,S,H,B} p(R, T, E, S, W, P, H, B)}{p(E)}$$

rewrite using the joint distribution given by the network

$$= \frac{\sum_{R,T,S,H,B} p(R)p(T)p(E)p(S | R, T)p(W | E)p(P | E)p(H | S, W)p(B | H)}{p(E)}$$

take $p(E)p(W | E)p(P | E)$ out of the sum

$$= \frac{p(E)p(W | E)p(P | E) \sum_{R,T,S,H,B} p(R)p(T)p(S | R, T)p(H | S, W)p(B | H)}{p(E)}$$

cancel $p(E)$

$$= p(W | E)p(P | E) \sum_{R,T,S,H,B} p(R)p(T)p(S | R, T)p(H | S, W)p(B | H)$$

B is not used as a condition and is summed over => cancels

$$= p(W | E)p(P | E) \sum_{R,T,S,H,B} p(R)p(T)p(S | R, T)p(H | S, W) \underbrace{p(B | H)}_1$$

H is not used as a condition and is summed over => cancels

$$= p(W | E)p(P | E) \sum_{R,T,S,H} p(R)p(T)p(S | R, T) \underbrace{p(H | S, W)}_1$$

S is not used as a condition and is summed over => cancels

$$= p(W | E)p(P | E) \sum_{R,T,S} p(R)p(T) \underbrace{p(S | R, T)}_1$$

T is not used as a condition and is summed over => cancels

$$= p(W | E)p(P | E) \sum_{R,T} p(R) \underbrace{p(T)}_1$$

R is not used as a condition and is summed over => cancels

$$= p(W | E)p(P | E) \underbrace{\sum_R p(R)}_1$$

$$= p(W | E)p(P | E)$$

This means that W and P are conditionally independent given E .

Question 3

$$p(B=1|T=c) = \frac{\sum_{R,E,S,W,P,H} p(R, T=c, E, S, W, P, H, B=1)}{p(T=c)}$$

rewrite using the joint distribution given by the network

$$= \frac{\sum_{R,E,S,W,P,H} p(R)p(T=c)p(E)p(S|R, T=c)p(W|E)p(P|E)p(H|S, W)p(B=1|H)}{p(T=c)}$$

take $p(T=C)$ out of the sum

$$= \frac{p(T=c) \sum_{R,E,S,W,P,H} p(R)p(E)p(S|R, T=c)p(W|E)p(P|E)p(H|S, W)p(B=1|H)}{p(T=c)}$$

cancel $p(T=c)$

$$= \sum_{R,E,S,W,P,H} p(R)p(E)p(S|R, T=c)p(W|E)p(P|E)p(H|S, W)p(B=1|H)$$

P is not used as a condition and it is summed over => cancels

$$= \sum_{R,E,S,W,H} p(R)p(E)p(S|R, T=c)p(W|E)p(H|S, W)p(B=1|H) \underbrace{\sum_P p(P|E)}_1$$

$$= \sum_{R,E,S,W,H} p(R)p(E)p(S|R, T=c)p(W|E)p(H|S, W)p(B=1|H)$$

$$p(T=c|B=1) = \frac{p(B=1|T=c)p(T=c)}{p(B=1)}$$

$$= \frac{p(B=1|T=c)p(T=c)}{p(B=1|T=c)p(T=c) + p(B=1|T=w)p(T=w)}$$

$$\begin{aligned}
p(W=1|S=1) &= \frac{\sum_{R,T,E,P,H,B} p(R,T,E,S=1,W=1,P,H,B)}{\sum_{R,T,E,W,P,H,B} p(R,T,E,S=1,W=1,P,H,B)} \\
&\text{rewrite using the joint distribution given by the network} \\
&= \frac{\sum_{R,T,E,P,H,B} p(R)p(T)p(E)p(S=1|R,T)p(W=1|E)p(P|E)p(H|S=1,W=1)p(B|H)}{\sum_{R,T,E,W,P,H,B} p(R)p(T)p(E)p(S=1|R,T)p(W|E)p(P|E)p(H|S=1,W)p(B|H)} \\
&\text{B is not used as a condition and it is summed over } \Rightarrow \text{cancels} \\
&= \frac{\sum_{R,T,E,P,H} p(R)p(T)p(E)p(S=1|R,T)p(W=1|E)p(P|E)p(H|S=1,W=1) \underbrace{\sum_B p(B|H)}_1}{\sum_{R,T,E,W,P,H} p(R)p(T)p(E)p(S=1|R,T)p(W|E)p(P|E)p(H|S=1,W) \underbrace{\sum_B p(B|H)}_1} \\
&\text{H is not used as a condition and it is summed over } \Rightarrow \text{cancels} \\
&= \frac{\sum_{R,T,E,P} p(R)p(T)p(E)p(S=1|R,T)p(W=1|E)p(P|E) \underbrace{\sum_H p(H|S=1,W=1)}_1}{\sum_{R,T,E,W,P} p(R)p(T)p(E)p(S=1|R,T)p(W|E)p(P|E) \underbrace{\sum_H p(H|S=1,W)}_1} \\
&\text{P is not used as a condition and it is summed over } \Rightarrow \text{cancels} \\
&= \frac{\sum_{R,T,E} p(R)p(T)p(E)p(S=1|R,T)p(W=1|E) \underbrace{\sum_P p(P|E)}_1}{\sum_{R,T,E,W} p(R)p(T)p(E)p(S=1|R,T)p(W|E) \underbrace{\sum_P p(P|E)}_1} \\
&\text{group probabilities} \\
&= \frac{\sum_E p(E)p(W=1|E) \sum_{R,T} p(R)p(T)p(S=1|R,T)}{\sum_{E,W} p(E)p(W|E) \sum_{R,T} p(R)p(T)p(S=1|R,T)} \\
&= \frac{\sum_E p(E)p(W=1|E)}{\sum_{E,W} p(E)p(W|E)} = \frac{\sum_E p(W=1,E)}{\sum_{E,W} p(W,E)} = \frac{p(W=1)}{1} \\
&= p(W=1)
\end{aligned}$$

This means that W and S are independent.

Again, you can convince yourself by doing the computation:

$$\begin{aligned}
p(W, S) &= \sum_{R, T, E, P, H, B} p(R, T, E, S, W, P, H, B) \\
&\quad \text{rewrite using the joint distribution given by the network} \\
&= \sum_{R, T, E, P, H, B} p(R)p(T)p(E)p(S | R, T)p(W | E)p(P | E)p(H | S, W)p(B | H) \\
&\quad B \text{ is not used as a condition and is summed over } \Rightarrow \text{cancels} \\
&= \sum_{R, T, E, P, H, B} p(R)p(T)p(E)p(S | R, T)p(W | E)p(P | E)p(H | S, W) \underbrace{p(B | H)}_1 \\
&\quad H \text{ is not used as a condition and is summed over } \Rightarrow \text{cancels} \\
&= \sum_{R, T, E, P, H} p(R)p(T)p(E)p(S | R, T)p(W | E)p(P | E) \underbrace{p(H | S, W)}_1 \\
&\quad P \text{ is not used as a condition and is summed over } \Rightarrow \text{cancels} \\
&= \sum_{R, T, E, P} p(R)p(T)p(E)p(S | R, T)p(W | E) \underbrace{p(P | E)}_1 \\
&= \sum_{R, T, E} p(R)p(T)p(E)p(S | R, T)p(W | E) \\
&\quad \text{factorise } E \\
&= \sum_E p(E)p(W | E) \sum_{R, T} p(R)p(T)p(S | R, T) \\
&\quad \text{factorise } R \text{ and } T \\
&= \left(\sum_E p(E)p(W | E) \right) \left(\sum_{R, T} p(R)p(T)p(S | R, T) \right) \\
&= p(W)p(S)
\end{aligned}$$

Exercise 4

To make it Bayesian, π , μ and Σ must be made random variables with prior distributions of their own. We would then have:

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Sigma | \alpha, \beta, \gamma) = p(\pi | \alpha) p(\mu | \beta) p(\Sigma | \gamma) \prod_{n=1}^N p(\mathbf{z}_n | \pi) p(\mathbf{x}_n | \mathbf{z}_n, \mu, \Sigma)$$