NTB
Interstaatliche Hochschule
für Technik Buchs

FHO Fachhochschule Ostschweiz

# Lab 9a (solution)

## Probabilistic Reasoning - Gaussian distribution and Bayes Theorem

TSM MachLe MSE
FS 2019

Machine Learning
WÜRC

The central *paradigm of probabilistic reasoning* is to identify all relevant variables $x_1, \ldots, x_N$ in the environment,and make a probabilistic model $p(x_1, \ldots, x_N)$ of their interaction. Reasoning (*inference*) is then performed by introducing *evidence* that sets variables in known states, and subsequently computing probabilities of interest, *conditioned on this evidence*. The rules of probability, combined with Bayes' rule make a reasoning system complete.

After this unit, ...

### Lernziele/Kompetenzen

- you have repeated the *basic rules of probability theory*.
- you know the difference between a *joint* and a *conditional probability* distribution.
- you know how to apply *Bayes Theorem* to calculate the *posterior* probability distribution for simple discrete examples. You can name the *prior* probability distribution, the *likelihood function*, the *evidence*, and you know how to *marginalize* over a joint probability distribution.
- you know the basic properties of a *multivariate Gaussian* probability distribution. You can plot a 2D Gaussian probability distribution given the *mean vector* $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.
- you can *sample* data points from a given multivariate gaussian distribution.
- you can explain the *naïve Bayes classifier* to your classmates and to your teacher.

## 1. Supervised Bayesian Learining [M, II]

The table below contains the result of a market survey for a promotion for different items such as a magazine, a watch and a life insurance and a credit card insurance. 10 people were interviewed and asked whether they would by such items.

Use this count table for supervised Bayesian learning. The output attribute is *sex* with possible values `male` and `female`. Consider an individual who has said *no* to the life insurance promotion, *yes* to the magazine promotion, *yes* to the watch promotion and *yes* to the credit card insurance. Use the values in the table together with the Naive Bayes classifier to determine which of a,b,c or d represents the probability that this individual is male. $p(E)$ is the marginal distribution.

| | Magazine | | Watch | | Life Insurance | | Credit Card | |
|---|---|---|---|---|---|---|---|---|
| | `male` | `female` | `male` | `female` | `male` | `female` | `male` | `female` |
| yes | 4 | 3 | 2 | 2 | 2 | 3 | 2 | 1 |
| no | 2 | 1 | 4 | 2 | 4 | 1 | 4 | 3 |

| Welche der folgenden Aussagen sind wahr und welche falsch? | **wahr** | **falsch** |
|---|---|---|
| **a)** $p(\texttt{sex} = \texttt{male}\|...) = \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{6}{10} \cdot \frac{1}{p(E)}$ | ☐ | ■ |
| **b)** $p(\texttt{sex} = \texttt{male}\|...) = \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{3}{4} \cdot \frac{2}{6} \cdot \frac{3}{4} \cdot \frac{1}{p(E)}$ | ☐ | ■ |
| **c)** $p(\texttt{sex} = \texttt{male}\|...) = \frac{4}{6} \cdot \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{6}{10} \cdot \frac{1}{p(E)}$ | ■ | ☐ |
| **d)** $p(\texttt{sex} = \texttt{male}\|...) = \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{4}{10} \cdot \frac{1}{p(E)}$ | ☐ | ■ |

## 2. Hamburger and Bayes Rule [A, II]

**a)** The probability that a hamburger eater HE will have Kreuzfeld-Jacob disease given the prior $p(\mathsf{KJ})$ and the marginal $p(\mathsf{HE}) = \sum_{\mathsf{KJ}} p(\mathsf{HE}|\mathsf{KJ}) \cdot p(\mathsf{KJ})$ is:

$$p(\mathsf{KJ}|\mathsf{HE}) = \frac{p(\mathsf{HE}, \mathsf{KJ})}{p(\mathsf{HE})} = \frac{p(\mathsf{HE}|\mathsf{KJ}) \cdot p(\mathsf{KJ})}{p(\mathsf{HE})} \tag{1}$$

$$= \frac{p(\mathsf{HE}|\mathsf{KJ}) \cdot p(\mathsf{KJ})}{\sum_{\mathsf{KJ}} p(\mathsf{HE}|\mathsf{KJ}) \cdot p(\mathsf{KJ})} = \frac{\frac{9}{10} \cdot \frac{1}{100000}}{\frac{1}{2}} \approx \underline{\underline{1.8 \cdot 10^{-5}}} \tag{2}$$

**b)** If the fraction of people eating hamburgers was rather small, $p(\mathsf{HE}) = 0,001$, what is the probability that a regular hamburger eater will have Kreuzfeld-Jacob disease?

## 3. Naïve Bayes Classifier [A,II]

**a)** Python Code: Naive Bayes prior and likelihoods

```
p_y = 4.0/10; # p(y) = 4/10
# p(xi=1 | y=-1)
p_x1_y0 = 3.0/6;
p_x2_y0 = 5.0/6;
p_x3_y0 = 4.0/6;
p_x4_y0 = 5.0/6;
p_x5_y0 = 2.0/6;

# p(xi=1 | y=+1)
p_x1_y1 = 3.0/4;
p_x2_y1 = 0.0/4;
p_x3_y1 = 3.0/4;
p_x4_y1 = 2.0/4;
p_x5_y1 = 1.0/4;
```

**b)** Python Code: Naive Bayes classification decisions

```
f_y1_00000 = p_y*(1-p_x1_y1)*(1-p_x2_y1)*(1-p_x3_y1)*
  (1-p_x4_y1)*(1-p_x5_y1)
print("f_y1_00000 = ",f_y1_00000)

f_y0_00000 = (1-p_y)*(1-p_x1_y0)*(1-p_x2_y0)*(1-p_x3_y0)*
  (1-p_x4_y0)*(1-p_x5_y0)
print("f_y0_00000 = " , f_y0_00000)

if (f_y1_00000 > f_y0_00000):
  print("Predict class +1")
else:
  print ("Predict class -1")
print("\n")

f_y1_11010 = p_y*(p_x1_y1)*(p_x2_y1)*(1-p_x3_y1)*
  (p_x4_y1)*(1-p_x5_y1)
print ("f_y1_11010 = ",f_y1_11010)

f_y0_11010 = (1-p_y)*(p_x1_y0)*(p_x2_y0)*(1-p_x3_y0)*
  (p_x4_y0)*(1-p_x5_y0)
print("f_y0_11010 = ",f_y0_11010)

if (f_y1_11010 > f_y0_11010):
  print("Predict class +1")
else:
  print ("Predict class -1")
```

The numerical solution $x = \{00000\}$ is:

$$f(y = +1|00000) = 0.009375000000000001$$
$$f(y = -1|00000) = 0.0018518518518518515$$
$$f(y = +1|00000) > f(y = -1|00000) \implies \hat{y} = +1$$

The numerical solution for $x = \{11010\}$ is:

$$f(y = +1|11010) = 0.0$$
$$f(y = -1|11010) = 0.046296296296296315$$
$$f(y = +1|11010) < f(y = -1|11010) \Longrightarrow \hat{y} = -1$$

**c)** Python Code: Naive Bayes posterior probabilities

```
# p(y=1|00000) =
print ("p(y=1|00000) =", f_y1_00000 / (f_y1_00000 + f_y0_00000))

# p(y=1|11010) =
print ("p(y=1|11010) =", f_y1_11010 / (f_y1_11010 + f_y0_11010))
```

The Naive Bayes posterior probabilities are:

$$p(y = 1|00000) = 0.8350515463917526$$
$$p(y = 1|11010) = 0.0$$

**d)** A *Bayes classifier using a joint distribution model* for $p(x_1, \ldots, x_5|y = c)$ would have $2^5 - 1 = 31$ degrees of freedom (independent probabilities) to estimate. But here, we have only 6 data points from class $y = -1$, and 4 data points from class $y = +1$. Thus these models would assign zero probability to many feature combinations, and *would probably not generalize well* to new data.

**e)** No, we do not need to re-train the model by estimating new probabilities. Due to the conditional independence assumptions of naïve Bayes, it is optimal to simply ignore $p(x_1|y)$ , and use the previously estimated probabilities of the other four features when computing $p(y|x_2, x_3, x_4, x_5)$. If you did recompute $p(x_i|y)$ using the formulas above, it is easy to verify that the values would not change.

## 4. Passenger Scanner [A,II]

The detector is such tat 95% of all terrorists are identified as terrorists

$$p(\mathsf{label} = true \mid \mathsf{terr} = true) = 0.95 \tag{3}$$

We can infer from the former that only 5% of terrorists got mislabeled as good people (false negative):

$$p(\mathsf{label} = false \mid \mathsf{terr} = true) = 0.05 \tag{4}$$

Furthermore 95% of all upstanding citizens are identified as such (non terrorists). Therefore, the false positive rate is:

$$p(\mathsf{label} = true \mid \mathsf{terr} = false) = 0.05 \tag{5}$$

Assuming that the informant is correct:

$$p(\mathsf{terr} = true) = \frac{1}{100} \tag{6}$$

We can then infer that $\frac{99}{100}$ are not terrorists:

$$p(\mathsf{terr} = false) = \frac{99}{100} \tag{7}$$

We can find out the probability that the person picked is a terrorist using *Bayes' Rule*:

$$p(\mathsf{terr} = true \mid \mathsf{label} = true) = \frac{p(\mathsf{label} = true \mid \mathsf{terr} = true)p(\mathsf{terr} = true)}{p(\mathsf{label} = true)} \tag{8}$$

To find out the denominator, we can *marginalize* over the joint distribution of $p(\mathsf{label} = true)$ and $p(\mathsf{terr})$:

$$\begin{aligned} p(label = true) &= \sum_t p(\mathsf{label} = true, \mathsf{terr} = t) \\ &= \sum_t p(\mathsf{label} = true \mid \mathsf{terr} = t) \cdot p(\mathsf{terr} = t) \end{aligned}$$

We make the summation explicit:

$$\begin{aligned} p(label = true) = p(\mathsf{label} = true \mid \mathsf{terr} = true)p(\mathsf{terr} = true) + \\ p(\mathsf{label} = true \mid \mathsf{terr} = false)p(\mathsf{terr} = false) \\ = 0.95 \cdot 0.01 + 0.05 \cdot 0.99 = \underline{\underline{0.059}} \end{aligned}$$

Substituting all the values into Bayes' Rule, we find out that even though the detector is 'reliable' we still get a low posterior probability that the person suspected of being a terrorist is actually a terrorist:

$$\begin{aligned} p(\mathsf{terr} = true \mid \mathsf{label} = true) &= \frac{p(\mathsf{label} = true \mid \mathsf{terr} = true)p(\mathsf{terr} = true)}{p(\mathsf{label} = true)} \\ &= \frac{0.95(0.01)}{0.059} = \underline{\underline{0.1610169492}} \end{aligned}$$

**5. Weather in London** [A, II]

**a)** Assuming that the prior probability it rained yesterday is 0.5, what is the probability that it was raining yesterday given that it's sunny today?

$$p(\mathsf{yesterday} = rain) = 50\%$$

I infer from this that the probability of being sunny yesterday is also 50%:

$$p(\mathsf{yesterday} = sun) = 50\%$$

$$\begin{aligned} &p(\mathsf{yesterday} = rain \mid \mathsf{today} = sun) \\ &\qquad = \frac{p(\mathsf{today} = sun \mid \mathsf{yesterday} = rain)p(\mathsf{yesterday} = rain)}{p(\mathsf{today} = sun)} \end{aligned}$$

$$
\begin{aligned}
p(\mathsf{today} = sun) &= \sum_y p(\mathsf{today} = sun, \mathsf{yesterday} = y) \\
&= \sum_y p(\mathsf{today} = sun \mid \mathsf{yesterday} = y) p(\mathsf{yesterday} = y) \\
&= p(\mathsf{today} = sun \mid \mathsf{yesterday} = sun) \cdot p(\mathsf{yesterday} = sun) \\
&+ p(\mathsf{today} = sun \mid \mathsf{yesterday} = rain) \cdot p(\mathsf{yesterday} = rain) \\
&= 0.40 \cdot 0.5 + 0.30 \cdot 0.50 = 0.20 + 0.15 = \underline{\underline{0.35}}
\end{aligned}
$$

Thus, the probability of raining yesterday given that today is sunny:

$$
p(\mathsf{yesterday} = rain \mid \mathsf{today} = sun) = \frac{0.15}{0.35} = \underline{\underline{42.86\%}}
$$

**b)** If the weather follows the same pattern as above, day after day, what is the probability that it will rain on any day (based on an effectively infinite number of days of observing the weather)?
On any day, not considering whether it rained or not the day before, the probability of raining is:

$$
\begin{aligned}
p(\mathsf{today} = rain) &= \sum_y p(\mathsf{today} = rain, \mathsf{yesterday} = y) \\
&= \sum_y p(\mathsf{today} = rain \mid \mathsf{yesterday} = y) p(\mathsf{yesterday} = y) \\
&= p(\mathsf{today} = rain \mid \mathsf{yesterday} = rain) p(\mathsf{yesterday} = rain) \\
&+ p(\mathsf{today} = rain \mid \mathsf{yesterday} = sun) p(\mathsf{yesterday} = sun) \\
&= 0.7 \cdot 0.5 + 0.6 \cdot 0.5 = \underline{\underline{0.65}}
\end{aligned}
$$

**c)** Use the result from b) above as a new prior probability of rain yesterday and recompute the probability that it was raining yesterday given that it's sunny today.

$$
p(\mathsf{yesterday} = rain) = 0.65
$$

Therefore:

$$
p(\mathsf{yesterday} = sun) = 0.35
$$

$$
p(\mathsf{yesterday} = rain \mid \mathsf{today} = sun) = \frac{p(\mathsf{today} = sun \mid \mathsf{yesterday} = rain) p(\mathsf{yesterday} = rain)}{p(\mathsf{today} = sun)}
$$

$$
\begin{aligned}
p(\mathsf{today} = sun) &= \sum_y p(\mathsf{today} = sun, \mathsf{yesterday} = y) \\
&= \sum_y p(\mathsf{today} = sun \mid \mathsf{yesterday} = y) p(\mathsf{yesterday} = y) \\
&= p(\mathsf{today} = sun \mid \mathsf{yesterday} = sun) p(\mathsf{yesterday} = sun) \\
&+ p(\mathsf{today} = sun \mid \mathsf{yesterday} = rain) p(\mathsf{yesterday} = rain) \\
&= 0.40 \cdot 0.35 + 0.30 \cdot 0.65 \\
&= 0.14 + 0.195 = \underline{\underline{0.335}}
\end{aligned}
$$

$$p(\text{yesterday} = rain \mid \text{today} = sun) = \frac{0.30 \cdot 0.65}{0.335} = \underline{\underline{58.21\%}}$$

Very interesting. The probability of raining has increased a little bit after updating the prior probabilities. Since it was likely that it rained yesterday, it's slightly more likely that it will rain today.

## 6. Inspector Clouseau [A,II]

**a)** Using $b$ for the two states of $B$ and $m$ for the two states of $M$,

$$p(B|K) = \sum_m p(B, m|K) = \sum_m \frac{p(B, m, K)}{p(K)} = \frac{p(B) \sum_m p(K|B, m)p(m)}{\sum_b p(b) \sum_m p(K|b, m)p(m)}$$

Plugging in the values we have

$$\begin{aligned} p(B = \text{murderer}|\text{knife used}) &= \frac{\frac{6}{10}\left(\frac{2}{10} \cdot \frac{1}{10} + \frac{8}{10} \cdot \frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10} \cdot \frac{1}{10} + \frac{8}{10} \cdot \frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10} \cdot \frac{2}{10} + \frac{8}{10} \cdot \frac{3}{10}\right)} \\ &= \frac{300}{412} \approx \underline{\underline{0.73}} \end{aligned}$$

Remark: The role of $p(\text{knife used})$ in the Inspector Clouseau example can cause some confusion. In the above, $p(\text{knife used})$ is computed to be 0.412.

$$p(\text{knife used}) = \sum_b p(b) \cdot \sum_m p(\text{knife used}|b, m) \cdot p(m) \approx \underline{\underline{0.412}} \tag{9}$$

But surely, $p(\text{knife used}) = 1$, since this is given in the question! Note that the quantity $p(\text{knife used})$ relates to the *prior probability* the model assigns to the knife being used (in the absence of any other information). If we know that the knife is used, then the posterior $p(\text{knife used}) = 1$.

## 7. Factorization of a multivariate probability distribution [A,II]

A belief network (BN) is a distribution of the form

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i|\text{pa}(x_i)) \tag{10}$$

where $\text{pa}(x_i)$ represent the *parental* variables of variable $x_i$. Represented as a directed graph, with an arrow pointing from a parent variable to child variable, a *belief network* corresponds to a *Directed Acyclic Graph* (DAG), with the $i^{th}$ node in the graph corresponding to the factor $p(x_i|\text{pa}(x_i))$. In general, two different graphs may represent the same independence assumptions. If one wishes to make independence assumptions, then the choice of factorisation becomes significant.

The observation that any distribution may be written in the *cascade form*, gives an algorithm for constructing a BN on variables $x_1, \ldots, x_n$: write down the $n-$node cascade graph; label the nodes with the variables in any order; now each successive independence statement corresponds

to deleting one of the edges. More formally, this corresponds to an ordering of the variables which, without loss of generality, we may write as $x_1, \ldots, x_n$. Then, from Bayes' rule, we have

$$p(x_1, \ldots, x_N) = p(x_1|x_2, \ldots, x_N) \cdot p(x_2, \ldots, x_N) \tag{11}$$
$$= p(x_1|x_2, \ldots, x_N) \cdot p(x_2|x_3, \ldots, x_N) \cdot p(x_3, \ldots, x_N) \tag{12}$$
$$= p(x_n) \cdot \prod_{i=1}^{N-1} p(x_i|x_{i+1}, \ldots, x_N) \tag{13}$$

The representation of any BN is therefore a Directed Acyclic Graph (DAG). Every probability distribution can be written as a BN, even though it may correspond to a fully connected 'cascade' DAG. The particular role of a BN is that the structure of the DAG corresponds to a *set of conditional independence assumptions*, namely which ancestral parental variables are sufficient to specify each conditional probability table. Note that this does not mean that non-parental variables have no influence.

## 8. Conditional distribution of a bivariate Gaussian distribution [A, II]

- We start by calculating the precision matrix $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$, the inverse of the covariance matrix $\mathbf{\Sigma}$. We use the fact that the inverse of a $2 \times 2$-matrix $\mathbf{A}$ is given by:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{14}$$
$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{15}$$

  In our case, we have

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{16}$$
$$\mathbf{\Lambda} = \mathbf{\Sigma}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} \tag{17}$$

- Now, we concentrate only on the *quadratic term* $\square$ in the exponential without the factor $-\frac{1}{2}$ which is given by:

$$\square = \frac{1}{1-\rho^2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \tag{18}$$
$$= \frac{1}{1-\rho^2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \cdot \begin{pmatrix} \frac{1}{\sigma_1^2}(x_1 - \mu_1) - \frac{\rho}{\sigma_1\sigma_2}(x_2 - \mu_2) \\ \frac{1}{\sigma_2^2}(x_2 - \mu_2) - \frac{\rho}{\sigma_1\sigma_2}(x_1 - \mu_1) \end{pmatrix} \tag{19}$$
$$= \frac{1}{1-\rho^2} \left\{ \frac{1}{\sigma_1^2}(x_1^2 - 2\mu_1 x_1 + \mu_1^2) - \frac{\rho}{\sigma_1\sigma_2}(x_1x_2 - x_1\mu_2 - x_2\mu_1 + \mu_1\mu_2) \right\} \tag{20}$$
$$+ \frac{1}{1-\rho^2} \left\{ \frac{1}{\sigma_2^2}(x_2^2 - 2\mu_2 x_2 + \mu_2^2) - \frac{\rho}{\sigma_1\sigma_2}(x_1x_2 - x_1\mu_2 - x_2\mu_1 + \mu_1\mu_2) \right\} \tag{21}$$

- Now, we only consider terms quadratic and linear in $x_1$ and get:

$$\square = \frac{1}{1-\rho^2} \left\{ \frac{x_1^2}{\sigma_1^2} - x_1 \cdot \left( \frac{2\mu_1}{\sigma_1^2} + \frac{2\rho}{\sigma_1\sigma_2}(x_2 - \mu_2) \right) + \ldots \right\} \tag{22}$$

$$= \frac{1}{1-\rho^2}\frac{1}{\sigma_1^2}\left\{x_1^2 - x_1 \cdot \left(2\mu_1 + \frac{2\rho\sigma_1^2}{\sigma_1\sigma_2}(x_2-\mu_2)\right) + \dots\right\} \tag{23}$$

- By completing the square, we finally get:

$$\square = \frac{1}{1-\rho^2}\frac{1}{\sigma_1^2}\left\{\left(x_1 - (\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-\mu_2))\right)^2 + \dots\right\} \tag{24}$$

- By identifying the coefficient before the quadratic term as the inverse of the variance $\tilde{sigma}$ of the conditional probability $p(x_1|x_2)$ and the shift as the mean $\tilde{\mu}$ of the conditional probability density, we find:

$$\tilde{\mu} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\cdot(x_2-\mu_2) \tag{25}$$

$$\tilde{\sigma} = (1-\rho^2)\cdot\sigma_1^2 \tag{26}$$

## 9. Bayesian inference for the Gaussian [A, II]

**a)** We consider only the terms in the exponentials and neglect the normalization factor. In this case, we have:

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}\cdot\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \tag{27}$$

Now, we only look at the quadratic term $\square$. The posterior probability distribution is a function of $\mu$. So we are interested only in the linear and quadratic terms in $\mu$.

$$\square = \left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \tag{28}$$

$$= \left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n^2+2x_n\mu+\mu^2) - \frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2)\right\} \tag{29}$$

$$= \left\{-\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N}x_n^2 + 2\mu\sum_{n=1}^{N}x_n + \mu^2\right) - \frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2)\right\} \tag{30}$$

$$= \left\{-\mu^2\left(\frac{1}{2\sigma_0^2}+\frac{N}{2\sigma^2}\right) + 2\mu\left(\frac{\sum_{n=1}^{N}x_n}{2\sigma_0^2}+\frac{\mu_0}{2\sigma_0^2}\right) + \dots\right\} \tag{31}$$

$$-\frac{1}{2}\left\{\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\left[\mu - \frac{\frac{\sum_n x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}}\right]^2 + \dots\right\} \tag{32}$$

The term in front of the square bracket is the inverse variance:

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{33}$$

The shift term in the quare bracket is the mean $\mu_N$:

$$\mu_N = \frac{\frac{\sum_n x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} \tag{34}$$

$$= \frac{\sigma_0^2 \sum_n x_n + \sigma_0^2 \mu_0}{\sigma^2 + N\sigma_0^2} \tag{35}$$

$$= \frac{\sigma^2}{\sigma^2 + N\sigma_0^2} \cdot \mu_0 + \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2} \cdot \frac{1}{N} \sum_n x_n \tag{36}$$

$$= \frac{\sigma^2}{\sigma^2 + N\sigma_0^2} \cdot \mu_0 + \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2} \cdot \mu_{\mathrm{ML}} \tag{37}$$

**b)** It is worth spending a moment studying the form of the posterior mean and variance. Using simple manipulation involving completing the square in the exponent, you could show that the posterior distribution is given by:

$$p(\mu \mid \mathbf{X}) = \mathcal{N}\left(\mu \mid \mu_N, \sigma_N^2\right) \tag{38}$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{39}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \cdot \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \cdot \mu_{\mathrm{ML}} \tag{40}$$

in which $\mu_{\mathrm{ML}}$ is the maximum likelihood solution for $\mu$ given by the sample mean

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{41}$$

First of all, we note that the mean of the posterior distribution given by 40 is a compromise between the prior mean $\mu_0$ and the maximum likelihood solution $\mu_{\mathrm{ML}}$. If the number of observed data points $N = 0$, then 40 reduces to the prior mean as expected. For $N \to \infty$, the posterior mean is given by the maximum likelihood solution.
Similarly, consider the result refVariancePosteriorA for the variance of the posterior distribution. We see that this is most naturally expressed in terms of the inverse variance, which is called the *precision*. Furthermore, the precisions are additive, so that the precision of the posterior is given by the precision of the prior plus one contribution of the data *precision* from each of the observed data points. As we increase the number of observed data points, the precision steadily increases, corresponding to a posterior distribution with steadily decreasing variance. With no observed data points, we have the prior variance, whereas if the number of data points $N \to \infty$, the variance $\sigma_N^2$ goes to zero and the posterior distribution becomes infinitely peaked around the maximum likelihood solution. We therefore see that the maximum likelihood result of a point estimate for $\mu$ given by 41 is recovered precisely from the Bayesian formalism in the limit of an infinite number of observations. Note also that for finite $N$, if we take the limit $\sigma_0^2 \to \infty$ which the prior has infinite variance then the posterior mean 40 reduces to the maximum likelihood result, while the posterior variance is given by $\sigma_N^2 = \sigma^2/N$.