Andrii Skliar
11636785
andrii.skliar@student.uva.nl

Homework Assignment 1
Machine Learning 1, 17/18

2019-01-08

1 MAP solution for Linear Regression

Note: here the following list of notation specifics is used

- Derivatives taken by vectors (i.e. $\frac{df(\mathbf{w})}{d\mathbf{w}}$) are taken as column vectors.
- $\mathbf{t} = (t_1, t_2 \dots t_N)^T$
- $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2} \dots \mathbf{x_N})^T$
- $\mathbf{w} = (w_1, w_2 \dots w_{M-1})^T$
- $\boldsymbol{\phi_n} = (\phi_0(\mathbf{x_n}), \phi_1(\mathbf{x_n}) \dots \phi_{M-1}(\mathbf{x_n}))^T$, where $\phi_j(\mathbf{x})$ are basis functions with $j = 0, 1 \dots M-1$ and $\phi_0 = 1$
- $\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x_1}) & \phi_1(\mathbf{x_1}) & \cdots & \phi_{M-1}(\mathbf{x_1}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x_N}) & \phi_1(\mathbf{x_N}) & \cdots & \phi_{M-1}(\mathbf{x_N}) \end{pmatrix}$

1.

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{x}, \mathbf{\Phi}, \mathbf{w}, \beta) &= \mathbf{a}) \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, \beta^{-1}) \\
&= \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2} \\
&= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2} \\
\mathbf{b}) &= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w})} \\
&= \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})
\end{aligned}
$$

2.

$$
\begin{aligned}
p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \\
&= \frac{|\alpha\mathbf{I}|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} e^{-\frac{1}{2}\mathbf{w}^T\alpha\mathbf{I}\mathbf{w}} \\
&= \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}} \\
\log p(\mathbf{w}|\alpha) &= \frac{M}{2}\log\frac{\alpha}{2\pi} - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}
\end{aligned}
$$

3.

Note: here $\int_{\mathbf{w}}$ means integrating over set of all possible values of $\mathbf{w}$.

$$\overbrace{p(\mathbf{w}|\mathbf{t},\mathbf{x},\mathbf{\Phi},\alpha,\beta)}^{\text{posterior}} = \frac{\overbrace{\left(\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n,\beta^{-1})\right)}^{\text{likelihood}}\overbrace{\mathcal{N}(\mathbf{w}|0,\alpha^{-1}\mathbf{I})}^{\text{prior}}}{\underbrace{\int_{\mathbf{w}}\left(\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n,\beta^{-1})\right)\mathcal{N}(\mathbf{w}|0,\alpha^{-1}\mathbf{I})d\mathbf{w}}_{\text{evidence}}}$$

$$= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{x},\mathbf{\Phi},\beta)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\mathbf{x},\mathbf{\Phi},\alpha,\beta)}$$

4.

a)

$$\begin{aligned}
\log p(\mathbf{w}|\mathbf{t},\mathbf{x},\mathbf{\Phi},\beta) &= \log p(\mathbf{t}|\mathbf{w},\mathbf{x},\mathbf{\Phi},\beta) + \log p(\mathbf{w}|\alpha)\\
&= \log\prod_{n=1}^{N}\sqrt{\frac{\beta}{2\pi}}e^{-\frac{\beta}{2}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2} + \log\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}\\
&= \sum_{n=1}^{N}\log\sqrt{\frac{\beta}{2\pi}}e^{-\frac{\beta}{2}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2} + \frac{M}{2}\log\frac{\alpha}{2\pi} - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\\
&= N\log\sqrt{\frac{\beta}{2\pi}} - \frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2 + \frac{M}{2}\log\frac{\alpha}{2\pi} - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\\
&= \underbrace{-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}_{\text{depends on }\mathbf{w}} + \underbrace{\frac{N}{2}\log\frac{\beta}{2\pi} + \frac{M}{2}\log\frac{\alpha}{2\pi}}_{\text{does not depend on }\mathbf{w}}\\
&= -\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + C
\end{aligned}$$

b)

$$\begin{aligned}
\log p(\mathbf{w}|\mathbf{t},\mathbf{x},\boldsymbol{\phi},\beta) &= \log p(\mathbf{t}|\mathbf{w},\mathbf{x},\boldsymbol{\phi},\beta) + \log p(\mathbf{w}|\alpha)\\
&= \log\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}e^{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w})} + \log\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}\\
&= \frac{N}{2}\log\frac{\beta}{2\pi} - \frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w}) + \frac{M}{2}\log\frac{\alpha}{2\pi} - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\\
&= \underbrace{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}_{\text{depends on }\mathbf{w}} + \underbrace{\frac{N}{2}\log\frac{\beta}{2\pi} + \frac{M}{2}\log\frac{\alpha}{2\pi}}_{\text{does not depend on }\mathbf{w}}\\
&= -\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + C
\end{aligned}$$

The parts, which are not based on $\mathbf{w}$ are normalizing factors in normal distribution.
Finding MAP might be much easier, because for finding full posterior distribution, in most of the cases of continuous distributions, we would need to find integral in the denominator, which might not always be analytically calculable or just very difficult to calculate. In this case we would have to find multiple integral, which would complicate our calculations quite a lot.

5.

a)

To make calculations less cumbersome, let's calculate $\frac{\partial\left(-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2\right)}{\partial\mathbf{w}}$ first.

$$\begin{aligned}\frac{\partial\left(-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2\right)}{\partial\mathbf{w}} &= \frac{\partial\left(-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2\right)}{\partial\mathbf{w}}\\ &= -\frac{\beta}{2}\sum_{n=1}^{N}\frac{\partial(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2}{\partial\mathbf{w}}\end{aligned}$$

$$\frac{\partial(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2}{\partial\mathbf{w}} = \left(\frac{\partial(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2}{\partial w_0},\frac{\partial(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2}{\partial w_1},\dots\frac{\partial(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2}{\partial w_M-1}\right)^T$$

$$\frac{\partial(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n))^2}{\partial w_i} = 2\phi_i(x_n)\left(t_n-\sum_{i=0}^{M-1}w_i\phi_i(x_n)\right) = 2\phi_i(x_n)\left(t_n-\sum_{i=0}^{M-1}\phi_i(x_n)w_i\right)$$

Generalizing for all i-s, we get following:

$$\begin{aligned}\frac{\partial\left(-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\boldsymbol{\phi}_n)^2\right)}{\partial\mathbf{w}} &= -\frac{\beta}{2}\cdot-2\left(\sum_{n=1}^{N}(\boldsymbol{\phi_n}(t_n-\boldsymbol{\phi}_n^T\mathbf{w}))\right)\\ &= \beta\cdot\left(\sum_{n=1}^{N}(\boldsymbol{\phi_n}(t_n-\boldsymbol{\phi}_n^T\mathbf{w}))\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial\log p(\mathbf{w}|\mathbf{t},\mathbf{x},\boldsymbol{\phi},\beta)}{\partial\mathbf{w}} &= \beta\cdot\left(\sum_{n=1}^{N}(\boldsymbol{\phi_n}(t_n-\boldsymbol{\phi}_n^T\mathbf{w}))\right)-\frac{\cancel{2}\alpha}{\cancel{2}}\mathbf{w}\\ &= \beta\cdot\left(\sum_{n=1}^{N}(\boldsymbol{\phi_n}t_n-\boldsymbol{\phi_n}\boldsymbol{\phi}_n^T\mathbf{w})\right)-\alpha\mathbf{w}\\ &= \beta\cdot(\boldsymbol{\Phi}^T\mathbf{t}-\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w})-\alpha\mathbf{w}=0\end{aligned}$$

If we now divide everything by β, we will get following expression (with $\lambda=\frac{\alpha}{\beta}$):

$$\begin{aligned}\boldsymbol{\Phi}^T\mathbf{t}-\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}-\lambda\mathbf{w}&=0\\ (\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\mathbf{I})\mathbf{w}&=\boldsymbol{\Phi}^T\mathbf{t}\\ \mathbf{w_{MAP}}&=\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\mathbf{I}\right)^{-1}\cdot\boldsymbol{\Phi}^T\mathbf{t}\end{aligned}$$

b)

$$\begin{aligned}\frac{\partial\log p(\mathbf{w}|\mathbf{t},\mathbf{x},\boldsymbol{\phi},\beta)}{\partial\mathbf{w}} &= -\frac{\beta}{2}\cdot\left(-2\boldsymbol{\Phi}^T(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})\right)-\frac{\cancel{2}\alpha}{\cancel{2}}\mathbf{w}\\ &= \beta\boldsymbol{\Phi}^T(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})-\alpha\mathbf{w}=0\\ \beta\boldsymbol{\Phi}^T\mathbf{t}-\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}-\alpha\mathbf{w}&=0\\ (\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\alpha\mathbf{I})\mathbf{w}&=\beta\boldsymbol{\Phi}^T\mathbf{t}\end{aligned}$$

If we now divide everything by β, we will get following expression (with $\lambda=\frac{\alpha}{\beta}$):

$$\begin{aligned}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\mathbf{I})\mathbf{w}&=\boldsymbol{\Phi}^T\mathbf{t}\\ \mathbf{w_{MAP}}&=\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\mathbf{I}\right)^{-1}\cdot\boldsymbol{\Phi}^T\mathbf{t}\end{aligned}$$

6.

First, I would like to explain, what is the role of $\phi_0(\mathbf{x})$ in the regularized regression. It is a bias term, meaning, that it will always be 1, so whenever we multiply it by any number, we just move the function, but don't

change its curvature or complexity. In 2-D space that would mean, that when we change the parameter w_0, which is the parameter, that ϕ_0 gets multiplied by, we move function either upwards or downwards.
Next, to understand why we don't need to regularize w_0, we have to analyze the reason for regularization. It is done so that our weights $\mathbf{w}$ are not too big and don't fit to the data points too much.
Then the reason for not regularizing w_0 comes simple: let's say, we know that a function, that generates our datapoints is following: $f(x) = x^2 + 10000$. In that case, making the bias term weight small won't make any sense, because we won't be able to fit to the $f(0) = 10000$ and therefore won't be able to fit to the data at all. Thus, we don't need to regularize bias weight, because this will make it more difficult or even impossible to fit to the origin point. For these reasons I don't see why we would penalize w_0 term at all.
However, if we indeed intend w_0 to have different penalization term, $p(\mathbf{w})$ would have following form:

$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{A} + \gamma^{-1}\mathbf{B}),$$

$$\text{where A} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \text{ with dimensionality of MxM;}$$

$$\text{and B} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \text{ with dimensionality of MxM.}$$

2 Probability distributions, likelihoods, and estimators

Question 2.1

- Bernoulli distribution can model events which can have only two outcomes (i.e. success or failure). Therefore, variables distributed using this law, can only take on 2 values.
  Also, Bernoulli distribution can be used for modelling series of such events (called Bernoulli process). The most obvious example would be coin toss, where, for example, head means success and tail - failure.
- Poisson distribution can be used for modelling events occurring in a fixed interval of space or time with constant rate independently of when the last event has happened. Variables then take on values of the number of events happening at some fixed period of time/space and probability means how probable it is for this number of events to occur. One of the examples might be the number of calls that will be received in a call-centre during the next hour if usual rate is 50 calls/hour.
- Gaussian (normal) distribution can be used for modelling almost any process, because there is CLT (central limit theorem), proving, that when there are sufficient number of independent observations, their sum has normal distribution. Therefore, properties proven using normal distribution also work for most of the other processes, even if they are not distributed via normal distribution. One of the examples might be modelling population height and, indeed, in most of the cases, the distribution of height of population will be distributed using Gaussian distribution.
- Gamma distribution can be used for modelling processes for which the waiting times between Poisson distributed events are relevant. Because exponential distribution and chi-distribution are specific cases of Gamma Distribution, i.e. it can be used for modelling time of functioning of devices.
- Log-Normal distribution can be used for any processes, which logarithm is distributed using normal distribution. For example, as Wikipedia states, the length of comments posted in Internet discussion forums follows a log-normal distribution.
- Beta distribution can be used for modelling events taking on values from the intervals of finite length. The most obvious example would, probably, be distribution of probabilities when we don't know what the probability is.

Question 2.2

Note: here notation of $\{r_1, r_2, \dots r_N\} = \mathbf{r}$ is used for convenience, where $\{r_1, r_2, \dots r_N\}$ is a set of observations.

1.

$$\begin{aligned}
p(r_t|\rho) &= \rho^{[r_t=1]} \cdot (1-\rho)^{[r_t=0]} \\
&= \{\text{with plugged-in numbers}\} \\
&= \left(\frac{207}{365}\right)^{[r_t=1]} \cdot \left(1-\frac{207}{365}\right)^{[r_t=0]} \\
&= \left(\frac{207}{365}\right)^{[r_t=1]} \cdot \left(\frac{158}{365}\right)^{[r_t=0]}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{r}|\rho, n_1) &= \prod_{t=1}^{N} \rho^{[r_t=1]} \cdot (1-\rho)^{[r_t=0]} \\
&= \rho^{\sum_{t=1}^{N}[r_t=1]} \cdot (1-\rho)^{\sum_{t=1}^{N}[r_t=0]} \\
&= \rho^{n_1} \cdot (1-\rho)^{N-n_1} \\
&= \rho^{n_1} \cdot (1-\rho)^{n_0} \\
&= \{\text{with plugged-in numbers}\} \\
&= \left(\frac{207}{365}\right)^{207} \cdot \left(1-\frac{207}{365}\right)^{158} \\
&= \left(\frac{207}{365}\right)^{207} \cdot \left(\frac{158}{365}\right)^{158}
\end{aligned}$$

2.

$$\begin{aligned}
\log\left(p(\mathbf{r}|\rho, n_1)\right) &= \log\left(\rho^{n_1}\right) + \log\left((1-\rho)^{n_0}\right) \\
&= n_1 \cdot \log\rho + n_0 \cdot \log\left(1-\rho\right)
\end{aligned}$$

3.

$$\begin{aligned}
\frac{\partial \log\left(\rho(\mathbf{r}|\rho, n_1)\right)}{\partial \rho} &= \frac{n_1}{\rho} - \frac{n_0}{1-\rho} \\
&= \frac{n_1 - n_1\rho - n_0\rho}{\rho(1-\rho)} \\
&= \frac{n_1 - N\rho}{\rho - \rho^2} = 0 \\
n_1 - N\rho &= 0 \\
\rho &= \frac{n_1}{N}
\end{aligned}$$

With plugged-in numbers (for a period of year):

$$\rho = \frac{207}{365}$$

4.

$$\begin{aligned}
p(\rho|a,b) &= \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \rho^{b-1}(1-\rho)^{a-1} \\
\rho_{MAP} &= \arg\max_{\rho} p(\mathbf{r}|\rho, n_1) \cdot p(\rho|a,b) \\
&= \arg\max_{\rho} \log\left(p(\mathbf{r}|\rho, n_1) \cdot p(\rho|a,b)\right) \\
&= \arg\max_{\rho} \log\left(\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \rho^{b-1}(1-\rho)^{a-1}\right) \\
&= \arg\max_{\rho} \left(n_1 \log\rho + n_0 \log(1-\rho) + \log\Gamma(b+a) - \log\Gamma(b)\log\Gamma(a) + (b-1)\log\rho + (a-1)\log(1-\rho)\right) \\
&= \arg\max_{\rho} \left((n_1+b-1)\log\rho + (n_0+a-1)\log(1-\rho) - \log\Gamma(b+a) - \log\Gamma(b) - \log\Gamma(a)\right)
\end{aligned}$$

Let's define function f as following for better convenience:

$$f(\rho, n_0, n_1, a, b) = (n_1 + b - 1)\log\rho + (n_0 + a - 1)\log\rho - \log\Gamma(b+a) - \log\Gamma(b) - \log\Gamma(a)$$

Therefore:

$$\begin{aligned}
\rho_{MAP} &= \arg\max_{\rho} f(\rho, n_0, n_1, a, b) \\
\frac{\partial f(\rho, n_0, n_1, a, b)}{\partial \rho} &= \frac{n_1 + b - 1}{\rho} - \frac{n_0 + a - 1}{1 - \rho} \\
&= \frac{n_1 + b - 1 - n_1\rho - b\rho + \rho - n_0\rho - a\rho + \rho}{(1-\rho)\rho} \\
&= \frac{(2 - n_0 - n_1 - a - b)\rho + n_1 + b - 1}{(1-\rho)\rho} = 0 \\
\rho_{MAP} &= \frac{1 - n_1 - b}{2 - n_0 - n_1 - a - b} = \frac{n_1 + b - 1}{N + a + b - 2}
\end{aligned}$$

5.

$$\begin{aligned}
p(\rho|\mathbf{r}) &= \frac{p(\mathbf{r}|\rho) \cdot p(\rho|a,b)}{p(\mathbf{r})} \\
&= \frac{p(\mathbf{r}|\rho) \cdot p(\rho|a,b)}{\int_0^1 p(\mathbf{r}|\rho) \cdot p(\rho|a,b) d\rho} \\
&= \frac{\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \rho^{b-1}(1-\rho)^{a-1}}{\int_0^1 \left(\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \rho^{b-1}(1-\rho)^{a-1}\right) d\rho}
\end{aligned}$$

6.

Note: here $\mathcal{B}(\theta_0, \theta_1)$ means Beta-function with θ_0 and θ_1 as parameters.
First, we should calculate an integral in the denominator to make further calculations easier and less cumbersome.

$$\begin{aligned}
\int_0^1 \left(\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \rho^{b-1}(1-\rho)^{a-1}\right) d\rho &= \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \int_0^1 \left(\rho^{n_1+b-1}(1-\rho)^{n_0+a-1}\right) d\rho \\
&= \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \mathcal{B}(n_1 + b, n_0 + a) \\
&= \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \cdot \frac{\Gamma(n_1+b)\Gamma(n_0+a)}{\Gamma(N+a+b)}
\end{aligned}$$

Using previously calculated integral, we get following:

$$\begin{aligned}
p(\rho|\mathbf{r}) &= \frac{p(\mathbf{r}|\rho) \cdot p(\rho|a,b)}{p(\mathbf{r})} \\
&= \frac{\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \rho^{b-1}(1-\rho)^{a-1}}{\int_0^1 \left(\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)} \rho^{b-1}(1-\rho)^{a-1}\right) d\rho} \\
&= \frac{\rho^{n_1} \cdot (1-\rho)^{n_0} \cdot \cancel{\frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}} \rho^{b-1}(1-\rho)^{a-1}}{\cancel{\frac{\Gamma(b+a)}{\Gamma(b)\Gamma(a)}} \cdot \frac{\Gamma(n_1+b)\Gamma(n_0+a)}{\Gamma(N+a+b)}} \\
&= \frac{\Gamma(N+a+b)}{\Gamma(n_1+b)\Gamma(n_0+a)} \cdot \rho^{n_1+b-1}(1-\rho)^{n_0+a-1} \\
&= Beta(\rho|n_0 + a, n_1 + b)
\end{aligned}$$

Note: here the last $Beta(\rho|n_0 + a, n_1 + b)$ means that $p(\rho|\mathbf{r})$ is distributed by Beta distribution with parameters $n_0 + a \; and \; n_1 + b$.

Question 2.3

Note: here notation of $(d_1, d_2, \ldots d_N) = \mathbf{d}$ is used for convenience.

1.

$$
\begin{aligned}
p(x|\lambda) &= \frac{\lambda^x}{x!}e^{-\lambda} \\
p(\mathbf{d}|\lambda) &= \prod_{t=1}^{T}\frac{\lambda^{d_t}}{d_t!}e^{-\lambda} \\
&= \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t!}e^{-T\lambda} \\
&= \frac{\lambda^{n}}{\prod_{t=1}^{T} d_t!}e^{-T\lambda}
\end{aligned}
$$

2.

$$
\begin{aligned}
\log p(\mathbf{d}|\lambda) &= \log\left(\frac{\lambda^n}{\prod_{t=1}^{T} d_t!}e^{-T\lambda}\right) \\
&= n\log\lambda - \sum_{t=1}^{T}\log d_t! - T\lambda
\end{aligned}
$$

3.

$$
\begin{aligned}
\frac{d\log\left(p(\mathbf{d}|\lambda)\right)}{d\lambda} &= \frac{n}{\lambda} - T = 0 \\
\frac{n}{\lambda} &= T \\
n &= \lambda T \\
\lambda &= \frac{n}{T} = \frac{1}{T}\sum_{t=1}^{T} d_t
\end{aligned}
$$

With plugged in numbers:

$$
\lambda = \frac{1}{14}\sum_{t=1}^{14} d_t = \frac{1}{14}\cdot 43 = \frac{43}{14} \approx 3.07
$$

4.

$$
p(\lambda|a,b) = \Gamma(a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}
$$

$$
\begin{aligned}
\lambda_{MAP} &= \arg\max_{\lambda} p(\mathbf{d}|\lambda)\cdot p(\lambda|a,b) \\
&= \arg\max_{\lambda} \prod_{t=1}^{T}\frac{\lambda^{d_t}}{d_t}e^{-\lambda}\cdot\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} \\
&= \arg\max_{\lambda} \log\left(\prod_{t=1}^{T}\frac{\lambda^{d_t}}{d_t}e^{-\lambda}\cdot\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}\right) \\
&= \arg\max_{\lambda} \log\left(\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t}e^{-T\lambda}\cdot\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}\right) \\
&= \arg\max_{\lambda} \left((\sum_{t=1}^{T} d_t)\log\lambda - \sum_{t=1}^{T}\log d_t - T\lambda + a\log b - \log\Gamma(a) + (a-1)\log\lambda - b\lambda\right) \\
&= \arg\max_{\lambda} \left(n\log\lambda - \sum_{t=1}^{T}\log d_t - T\lambda + a\log b - \log\Gamma(a) + (a-1)\log\lambda - b\lambda\right)
\end{aligned}
$$

$$\begin{aligned}\frac{\partial \log\left(p(\mathbf{d}|\lambda)\cdot p(\lambda|a,b)\right)}{\partial\lambda} &= \left(\frac{n}{\lambda} - T + \frac{a-1}{\lambda} - b\right)\\ &= \frac{n+a-1}{\lambda} - T - b = 0\\ \frac{n+a-1}{\lambda} &= T + b\\ \lambda &= \frac{n+a-1}{T+b}\end{aligned}$$

5.

Note: here $\int_\lambda$ means integrating over set of all possible values of λ.

$$\begin{aligned}p(\lambda|\mathbf{d}) &= \frac{p(\mathbf{d}|\lambda)p(\lambda|a,b)}{\int_\lambda p(\mathbf{d}|\lambda)p(\lambda|a,b)d\lambda}\\ &= \frac{\prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)}{\int_\lambda \prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)d\lambda}\\ &= \frac{\prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)}{\int_0^{+\infty} \prod_{t=1}^{T} p(d_t|\lambda)p(\lambda|a,b)d\lambda}\\ &= \frac{\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t}e^{-T\lambda}\cdot\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{\int_0^{+\infty}\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t}e^{-T\lambda}\cdot\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}d\lambda}\end{aligned}$$

6.

Note: here $\Gamma(\theta)$ means Gamma-function with θ as a parameter.
First, we should calculate an integral in the denominator to make further calculations easier and less cumbersome.

$$\begin{aligned}\int_0^{+\infty}\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t}e^{-T\lambda}\cdot\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}d\lambda &= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!}\int_0^{+\infty}\lambda^{\sum_{t=1}^{T} d_t}e^{-T\lambda-b\lambda}\lambda^{a-1}d\lambda\\ &= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!}\int_0^{+\infty}\lambda^{n+a-1}e^{-(T+b)\lambda}d\lambda\\ = \left\{\lambda' = (T+b)\lambda;\ d\lambda' = (T+b)d\lambda;\ \lambda = \frac{\lambda'}{T+b};\ d\lambda = \frac{d\lambda'}{T+b}\right\} &= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!}\int_0^{+\infty}\left(\frac{\lambda'}{T+b}\right)^{n+a-1}e^{-\lambda'}\frac{1}{T+b}d\lambda'\\ &= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!}\cdot\frac{1}{(T+b)^{n+a}}\int_0^{+\infty}{\lambda'}^{n+a-1}e^{-\lambda'}d\lambda'\\ &= \frac{b^a}{\Gamma(a)\prod_{t=1}^{T} d_t!}\cdot\frac{1}{(T+b)^{n+a}}\Gamma(n+a)\end{aligned}$$

Using previously calculated integral, we get following:

$$
\begin{aligned}
p(\lambda|\mathbf{d}) &= \frac{\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}{\int_0^{+\infty} \frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda} \\
&= \frac{\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t} e^{-T\lambda} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}{\frac{b^a}{\Gamma(a) \prod_{t=1}^{T} d_t!} \cdot \frac{1}{(T+b)^{n+a}} \Gamma(n+a)} \\
&= \frac{\cancel{\frac{\lambda^n}{\prod_{t=1}^{T} d_t}} e^{-T\lambda} \cdot \cancel{\frac{b^a}{\Gamma(a)}} \lambda^{a-1} e^{-b\lambda}}{\cancel{\frac{b^a}{\Gamma(a) \prod_{t=1}^{T} d_t!}} \cdot \frac{1}{(T+b)^{n+a}} \Gamma(n+a)} \\
&= \frac{\lambda^n e^{-T\lambda} \lambda^{a-1} e^{-b\lambda} (T-b)^{n+a}}{\Gamma(n+a)} \\
&= \frac{\lambda^{n+a-1} e^{-(T+b)\lambda} (T+b)^{n+a}}{\Gamma(n+a)} \\
&= \frac{(T+b)^{n+a}}{\Gamma(n+a)} \lambda^{(n+a)-1} e^{-(T+b)\lambda} \\
&= Gamma(\lambda|n+a, T+b)
\end{aligned}
$$

Note: here the last $Gamma(\lambda|n+a, T+b)$ means that $p(\lambda|\mathbf{d})$ is distributed by Gamma distribution with parameters $n+a$ *and* $T+b$.