# Machine Learning 1 - Homework 2

Pascal Mattia Esser

September 20th, 2017

## 1 MAP solution for Linear Regression

### 1.1

given:

$$\mathcal{N}(x|\mu\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{1}$$

a) we can write as product form:

$$p(\boldsymbol{x}|\boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{t}_i|\boldsymbol{w}^T\boldsymbol{\Phi}, \sigma^2) \tag{2}$$

$$= \prod_{i=1}^{N} \frac{1}{(2\pi\beta^{-}1)^{1/2}} exp\left(-\frac{1}{2\beta^{-1}}(t_i - \boldsymbol{w}^T\boldsymbol{\phi_i})^2\right) \tag{3}$$

b) write as matrix form

$$p(\boldsymbol{x}|\boldsymbol{w}) = \prod_{i=1}^{N} \frac{1}{(2\pi\beta^{-}1)^{1/2}} exp\left(-\frac{1}{2\beta^{-}1}(t_i - \boldsymbol{w}^T\boldsymbol{\phi_i})^2\right) \tag{4}$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2} \prod_{i=1}^{N} exp\left(-\frac{\beta}{2}(t_i - \boldsymbol{w}^T\boldsymbol{\phi_i})^2\right) \tag{5}$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2} exp\left(-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - \boldsymbol{w}^T\boldsymbol{\phi_i})^2\right) \tag{6}$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2} exp\left(-\frac{\beta}{2}||\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\Phi}||^2\right) \tag{7}$$

### 1.2

starting from the multivariate Gaussian distribution:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|^2)^{M/2}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{8}$$

with

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \tag{9}$$

you get

$$= \frac{1}{(2\pi|\alpha^{-1}\boldsymbol{I}|^2)^{M/2}} exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{0})^T(\alpha^{-1}\boldsymbol{I})^{-1}(\boldsymbol{w}-\boldsymbol{0})\right) \tag{10}$$

$$= \left(\frac{\alpha}{2\pi}^{M/2}\right) exp\left(-\frac{\alpha \boldsymbol{w}^T\boldsymbol{w}}{2}\right) \tag{11}$$

calculate the log

$$ln(p(\boldsymbol{w})) = \frac{M}{2} log\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha \boldsymbol{w}^T\boldsymbol{w}}{2} \tag{12}$$

## 1.3

starting from a Bayes formula

$$p(\boldsymbol{w}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{x})} \tag{13}$$

and setting:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{t}|\boldsymbol{\Phi}, \boldsymbol{w}, \beta)p(\boldsymbol{w})d\boldsymbol{w} \tag{14}$$

gives us:

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{x}, \beta, \alpha) = \frac{\prod^N \mathcal{N}(t_i|\boldsymbol{w}^T\phi_i, \beta^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{\int p(\boldsymbol{t}|\boldsymbol{\Phi}, \boldsymbol{w}, \beta)p(\boldsymbol{w})d\boldsymbol{w}} \tag{15}$$

rewrite the integral following [Bis06, p. 93]:

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{x}, \beta, \alpha) = \frac{\prod^N \mathcal{N}(t_i|\boldsymbol{w}^T\phi_i, \beta^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{\mathcal{N}(\boldsymbol{t}|\boldsymbol{\Phi 0} + \boldsymbol{0}, \frac{1}{\beta}\boldsymbol{I} + \boldsymbol{\Phi}(\alpha^{-1})^{-1}\boldsymbol{\Phi}^T)} \tag{16}$$

$$= \frac{\prod^N \mathcal{N}(t_i|\boldsymbol{w}^T\phi_i, \beta^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})}{\mathcal{N}(\boldsymbol{0}, \frac{1}{\beta}\boldsymbol{I} + \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^T)} \tag{17}$$

## 1.4

for the logarithm, we know:

$$\ln p(\boldsymbol{w}|\boldsymbol{x}) = \ln p(\boldsymbol{x}|\boldsymbol{w}) + \ln p(\boldsymbol{w}) - \ln p(\boldsymbol{x}) \tag{18}$$

a) writing it in product form (keep $lnp(\boldsymbol{x})$ as term and don't expend it, as it does not depend on $\boldsymbol{w}$:

$$\ln p(\boldsymbol{w}|\boldsymbol{x}) \tag{19}$$

$$= \ln\left(\left(\frac{\beta}{2\pi}\right)^{M/2}\prod_{i=1}^{N} exp\left(-\frac{\beta}{2}(t_i - \boldsymbol{w}^T\boldsymbol{\phi_i})^2\right) + \left(\frac{\alpha}{2\pi}^{M/2}\right) exp\left(-\frac{\alpha \boldsymbol{w}^T\boldsymbol{w}}{2}\right)\right) - \ln p(\boldsymbol{x}) \tag{20}$$

$$= \frac{M}{2}\ln\frac{\beta}{2\pi} + \left(-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - \boldsymbol{w}^T\boldsymbol{\phi_i})^2\right) + \frac{M}{2} log\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha \boldsymbol{w}^T\boldsymbol{w}}{2} - \ln p(\boldsymbol{x}) \tag{21}$$

therms, depending on $\boldsymbol{w}$ and substitute everything else in $C$:

$$\ln p(\boldsymbol{w}|\boldsymbol{x}) = -\frac{\beta}{2}\sum_{i=1}^{N}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}_i)^2 - \frac{\alpha\boldsymbol{w}^T\boldsymbol{w}}{2} + C \tag{22}$$

b) writing it in matrix form:

$$\ln p(\boldsymbol{w}|\boldsymbol{x}) \tag{23}$$

$$= \ln\left(\left(\frac{\beta}{2\pi}\right)^{M/2} exp\left(-\frac{\beta}{2}||\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\Phi}||^2\right) + \left(\frac{\alpha}{2\pi}^{M/2}\right) exp\left(-\frac{\alpha\boldsymbol{w}^T\boldsymbol{w}}{2}\right)\right) - \ln p(\boldsymbol{x}) \tag{24}$$

$$= \frac{M}{2}\ln\frac{\beta}{2\pi} + \left(-\frac{\beta}{2}||\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\Phi}||^2\right) + \frac{M}{2}log\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha\boldsymbol{w}^T\boldsymbol{w}}{2} - \ln p(\boldsymbol{x}) \tag{25}$$

therms, depending on $\boldsymbol{w}$:

$$\ln p(\boldsymbol{w}|\boldsymbol{x}) = \left(-\frac{\beta}{2}||\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\Phi}||^2\right) - \frac{\alpha\boldsymbol{w}^T\boldsymbol{w}}{2} + C \tag{26}$$

were $C$, the terms not depending on w is

$$C = \frac{M}{2}\ln\frac{\beta}{2\pi} + \frac{M}{2}log\left(\frac{\alpha}{2\pi}\right) - \ln p(\boldsymbol{x}) \tag{27}$$

It is easier to calculate the MAP because you don't have to solve the integral in the evidence term $p(\boldsymbol{x})$, which can be hard analytically (or intractable).

## 1.5

take the derivative of the product form:

$$\frac{\partial}{\partial\boldsymbol{w}}\ln p(\boldsymbol{w}|\boldsymbol{x}) = \frac{\partial}{\partial\boldsymbol{w}}\left(-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}_i)^2 - \frac{\alpha\boldsymbol{w}^T\boldsymbol{w}}{2}\right) \tag{28}$$

$$= -\frac{\beta}{2}\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{w}}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}_i)^2 - \alpha\boldsymbol{I}\boldsymbol{w}^T \tag{29}$$

$$= -\beta\sum_{i=1}^{N}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}_i)\boldsymbol{\phi}_i^T - \alpha\boldsymbol{I}\boldsymbol{w}^T \tag{30}$$

$$= \sum_{i=1}^{N}(\boldsymbol{\phi}_i^T t_i) + \boldsymbol{w}^T\sum_{i=1}^{N}(\boldsymbol{\phi}_i\boldsymbol{\phi}_i^T + \lambda\boldsymbol{I}) \tag{31}$$

$$\boldsymbol{w} = \left(\sum_{i=1}^{N}(\boldsymbol{\phi}_i^T\boldsymbol{\phi}_i + \lambda\boldsymbol{I})\right)^{-1}\sum_{i=1}^{N}(\boldsymbol{\phi}_i^T t_i) \tag{32}$$

Take the derivative of the matrix form:

$$\frac{\partial}{\partial \boldsymbol{w}} \ln p(\boldsymbol{w}|\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{w}} \left( -\frac{\beta}{2} ||\boldsymbol{t} - \boldsymbol{w}^T \boldsymbol{\Phi}||^2 \right) + -\frac{\alpha \boldsymbol{w}^T \boldsymbol{w}}{2} + C \tag{33}$$

$$= \left( -\frac{\beta}{2} \frac{\partial}{\partial \boldsymbol{w}} ||\boldsymbol{t} - \boldsymbol{w}^T \boldsymbol{\Phi}||^2 \right) - \frac{\partial}{\partial \boldsymbol{w}} \frac{\alpha \boldsymbol{w}^T \boldsymbol{w}}{2} \tag{34}$$

$$= \left( -\frac{\beta}{2} \frac{\partial}{\partial \boldsymbol{w}} (\boldsymbol{t} - \boldsymbol{w}^T \boldsymbol{\Phi})^T (\boldsymbol{t} - \boldsymbol{w}^T \boldsymbol{\Phi}) \right) - \frac{\partial}{\partial \boldsymbol{w}} \frac{\alpha \boldsymbol{w}^T \boldsymbol{w}}{2} \tag{35}$$

$$= \left( -\frac{\beta}{2} \boldsymbol{\Phi}^T \boldsymbol{I} (\boldsymbol{t} - \boldsymbol{w}^T \boldsymbol{\Phi}) + \boldsymbol{\Phi}^T \boldsymbol{I}^T (\boldsymbol{t} - \boldsymbol{w}^T \boldsymbol{\Phi}) \right) - \alpha \boldsymbol{I} \boldsymbol{w}^T \tag{36}$$

$$= (\beta \boldsymbol{\Phi}^T \boldsymbol{I} \boldsymbol{t}) + (\beta \boldsymbol{\Phi}^T \boldsymbol{I} \boldsymbol{\Phi} + \alpha \boldsymbol{I}) \boldsymbol{w}^T \tag{37}$$

set to 0 and solve for $\boldsymbol{w}$:

$$\boldsymbol{w}^T = (\beta \boldsymbol{\Phi}^T \boldsymbol{I} \boldsymbol{\Phi} + \alpha \boldsymbol{I})^{-1} (\beta \boldsymbol{\Phi}^T \boldsymbol{I} \boldsymbol{t}) \tag{38}$$

$$\boldsymbol{w} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{\Phi}^T \boldsymbol{t}) \tag{39}$$

with $\lambda = \alpha / \beta$

### 1.6

$\phi(\boldsymbol{x}) = 1$ is the offset or bias of the equation:

You should not use the same prior, because otherwise the same prior believe is applied on the offset of the function and the variables, that govern the shape of the function. taking this into account, we can write the new prior as:

$$p(\boldsymbol{w}) = \left( \frac{\alpha}{2\pi} \right)^{(N-1)/2} exp \left( -\frac{\alpha \boldsymbol{w}^T \boldsymbol{w} - \boldsymbol{w}_0^2}{2} \right) \left( \frac{\gamma}{2\pi} \right)^{1/2} exp \left( -\frac{\gamma \boldsymbol{w}_0^2}{2} \right) \tag{40}$$

here we have different distributions for $\boldsymbol{w}$ and $\boldsymbol{w}_0$.

## 2 Probability distributions, likelihoods, and estimators

## 2.1 Describe distributions

### 2.1.1 Bernouilli

- experiments, that ask for yes or now questions.
- it takes value 1 with probability $p$ and 0 with $p - 1$
- coin toss

### 2.1.2 Beta

- contains values in the interval [0,1]
- uses $\alpha$ and $\beta$ as shaping parameters for the from of the distribution
- betting rates in basketball for throwing vs. hitting

#### 2.1.3 Poisson

- probability function over discrete probability values
- probability of a given number of events in a fixed time/space interval.
- known average rate and independently of the time of the last event
- probability of n mails per day

#### 2.1.4 Gamma

- Poisson, where the waiting time between event is relevant
- mails are poisson distributed. now the waiting time till the 4th mail is gamma distributed

#### 2.1.5 Gaussian

- distribution over variables for an sufficient large i.i.d. data set
- mean and standard deviation define the function
- hight distribution of humans

#### 2.1.6 Log-Normal

- random variables, who logs are normal distributed
- distribution of the firing rate across a population of neurons in the hippocampus

### 2.2 Rain in Den Helde

given variables:

$$n_1 = 207 \tag{41}$$
$$N = 365 \tag{42}$$
$$n_0 = 158 \tag{43}$$

Assuming a Bernoulli distribution, we can write:

$$p(\boldsymbol{x}|\rho) = \prod^{N} p(r_t|\rho) \tag{44}$$

using Bernoulli:

$$p(r_t|\rho) = \rho^{[r_t=1]}(1-\rho)^{[r_t=0]} \tag{45}$$

$$p(\boldsymbol{x}|\rho) = \prod^{N} \rho^{[r_t=1]}(1-\rho)^{[r_t=0]} \tag{46}$$

### 2.2.1

we can write:

$$p(\boldsymbol{x}|\rho) = \prod^{N} \rho^{[r_t=1]}(1-\rho)^{[r_t=0]} \tag{47}$$

$$= \rho^{n_1}(1-\rho)^{n_0} \tag{48}$$

### 2.2.2

computing the log

$$\ln(p(\boldsymbol{x}|\rho)) = \ln \rho^{n_1} + \ln(1-\rho)^{n_0} \tag{49}$$

### 2.2.3

Take the derivatives

$$\frac{\partial}{\partial \rho} = \frac{n_1}{\rho} + \frac{n_0}{1-\rho} = 0 \tag{50}$$

solve for $\rho$:

$$\rho = \frac{n_1}{n_0}(1-\rho) \tag{51}$$

$$= \frac{n_1}{n_1+n_0} \tag{52}$$

$$= \frac{n_1}{N} \tag{53}$$

use the numbers:

$$\rho = \frac{207}{365} \approx 0.5671 \tag{54}$$

### 2.2.4

rewrite with a known Beta prior:

$$p(\boldsymbol{x}|\rho) = \frac{\rho^{n_1}(1-\rho)^{n_0}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}\right)}{p(\boldsymbol{x})} \tag{55}$$

compute the ln:

$$\ln p(\boldsymbol{x}|\rho) = n_1 \ln \rho + \ln n_0(1-\rho) + ln\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right) + (a-1)\ln\rho + (b-1)\ln(1-\rho) + C \tag{56}$$

take the derivative

$$\frac{\partial}{\partial \rho} = \frac{n_1}{\rho} - \frac{n_0}{1-\rho} + \frac{a-1}{\rho} - \frac{b-1}{1-\rho} \tag{57}$$

setting to 0 and solving for $\rho$:

$$\rho = \frac{n_1 + a - 1}{n_1 + a + b + n_0 - 2} \tag{58}$$

**2.2.5**

$$p(\rho|\boldsymbol{x}) = \frac{\rho^{n_1}(1-\rho)^{n_0}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}\right)}{\int \rho'^{n_1}(1-\rho')^{n_0}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho'^{a-1}(1-\rho')^{b-1}\right)d\rho'} \tag{59}$$

**2.2.6**

starting from:

$$p(\rho|\boldsymbol{x}) = \frac{\rho^{n_1}(1-\rho)^{n_0}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}\right)}{\int \rho'^{n_1}(1-\rho')^{n_0}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho'^{a-1}(1-\rho')^{b-1}\right)d\rho'} \tag{60}$$

canceling out

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{61}$$

and solving for the denominator

$$\int \rho'^{n_1}(1-\rho')^{n_0}\left(\rho'^{a-1}(1-\rho')^{b-1}\right)d\rho' = 1 \tag{62}$$

rewrite:

$$\int \rho^{a+n_1-1}(1-\rho)^{b+n_0-1} = 1 \tag{63}$$

and compare with

$$\rho'^{a-1}(1-\rho')^{b-1} \tag{64}$$

gives

$$p(\rho) = B(a+n_1, b+n_0) \tag{65}$$

as the Beta distribution

## 2.3 staffing department of a maternity hospital

$$T = 14 \tag{66}$$

$$n = 4+7+3+0+2+2+1+5+4+4+3+3+2+3 = 43 \tag{67}$$

**2.3.1**

likelihood for one observation: set $n = d$ for one observation

$$p(\boldsymbol{x}|\lambda) = \frac{\lambda^{d_i}}{d_i!}e^{-\lambda} \tag{68}$$

for all observations

$$p(\boldsymbol{x}|\lambda) = \prod^{T} \frac{\lambda^{d_i}}{d_i!} e^{-\lambda} \tag{69}$$

$$= \frac{1}{\prod d_i!} \lambda^n e^{-T\lambda} \tag{70}$$

**2.3.2**

likelihood for all observations:

$$\ln p(\boldsymbol{x}|\lambda) = \ln \left( \prod^{T} \frac{\lambda^{d_i}}{d_i!} e^{-\lambda} \right) \tag{71}$$

$$= \sum \ln \frac{\lambda^{d_i}}{d_i!} e^{-\lambda} \tag{72}$$

$$= -T\lambda + nln\lambda - \sum \ln(d_i!) \tag{73}$$

**2.3.3**

take the derivative of the log:

$$\frac{\partial}{\partial \lambda} = \frac{n}{\lambda} - T \tag{74}$$

set the derivative to zero and solve fro $\lambda$

$$\lambda = \frac{n}{T} \tag{75}$$

calculate with numbers:

$$\lambda = \frac{43}{14} \approx 3.0714 \tag{76}$$

**2.3.4**

Assume a Gamma prior for $\lambda$

$$p(\lambda|\boldsymbol{x}) = \frac{\prod \left( \frac{\lambda^{d_i}}{d_i!} e^{-\lambda} \right) \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}{p(\boldsymbol{x})} \tag{77}$$

calculate the ln

$$\ln p(\lambda|\boldsymbol{x}) = \frac{n}{\lambda} - T + \ln \left( \frac{b^a}{\Gamma(a)} \right) + \lambda \ln(a-1) - b\lambda + C \tag{78}$$

calculate the derivative:

$$\frac{\partial}{\partial \lambda} = \frac{n}{\lambda} - T + \lambda \frac{1}{a-1} - b = 0 \tag{79}$$

solving for $\lambda$ gives us:

$$\lambda = \frac{n + a - 1}{T + b} \tag{80}$$

#### 2.3.5

Write down the form of the posterior distribution for $\lambda$

$$p(\lambda|\boldsymbol{x}) = \frac{\prod \left(\frac{\lambda^{d_i}}{d_i!}e^{-\lambda}\right)\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{\int \prod \left(\frac{\lambda'^{d_i}}{d_i!}e^{-\lambda'}\right)\frac{b^a}{\Gamma(a)}\lambda'^{a-1}e^{-b\lambda'}d\lambda'} \tag{81}$$

#### 2.3.6

same approach as in subsubsection 2.2.6: start with

$$p(\lambda|\boldsymbol{x}) = \frac{\prod \left(\frac{\lambda^{d_i}}{d_i!}e^{-\lambda}\right)\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{\int \prod \left(\frac{\lambda'^{d_i}}{d_i!}e^{-\lambda'}\right)\frac{b^a}{\Gamma(a)}\lambda'^{a-1}e^{-b\lambda'}d\lambda'} \tag{82}$$

rewrite as:

$$p(\lambda|\boldsymbol{x}) = \frac{\frac{1}{\prod d_i!}\lambda^T e^{-n\lambda}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}}{\int \frac{1}{\prod d_i!}\lambda^T e^{-n\lambda'}\frac{b^a}{\Gamma(a)}\lambda'^{a-1}e^{-b\lambda'}d\lambda'} \tag{83}$$

after canceling out the constants we get for the denominator:

$$\int \lambda^n e^{-T\lambda'}\lambda'^{a-1}e^{-b\lambda'}d\lambda' = 1 \tag{84}$$

now bring it in the form of the initial prior,

$$\lambda^{a-1}e^{-b\lambda} \tag{85}$$

gives a gamma function of:

$$p(\lambda) = \Gamma(a+n, b+T) \tag{86}$$

## References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.