Tarun Krishna
11593040
krishnatarun7@gmail.com

Homework Assignment 2
Machine Learning 1, 17/18

2017-09-20

1 First Exercise

$$
\begin{aligned}
y(\mathbf{x}_n, \mathbf{w}) &= \mathbf{w}^T \phi_n \\
p(t_n|\phi, w, \beta) &= \mathcal{N}(t_n|w^T\phi_n, \beta^{-1}) \\
p(w) &= \mathcal{N}(w|\mathbf{0}, \alpha^{-1}\mathbf{I})
\end{aligned}
$$

where **0** is a vector of 0's and **I** identity matrix.

1.1

$$
\begin{aligned}
L(\theta) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} exp^{-(t_n - w^T\phi_n)^2 1/2\beta^{-1}} \\
&= \frac{1}{(2\pi\beta^{-1})^{N/2}} exp^{\sum_{n=1}^{N} -(t_n - w^T\phi_n)^2 1/2\beta^{-1}} \\
\Rightarrow &= \frac{1}{(2\pi\beta^{-1})^{N/2}} exp^{(t-\Phi\mathbf{w})^2 1/2\beta^{-1}} \\
&= \mathcal{N}(t|\Phi\mathbf{w}, \beta^{-1})
\end{aligned}
$$

1.2 Expression for multivariate gaussian distribution for some **X** given by:

$$
p(X|\mu, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} exp^{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)}
$$

$p(w)$ can be written in a similar fashion.

$$
\begin{aligned}
p(w) &= \mathcal{N}(w|\mathbf{0}, \alpha^{-1}\mathbf{I}) \\
p(w) &= \frac{1}{\sqrt{|2\pi\alpha^{-1}I|}} exp^{-1/2(w-\mathbf{0})(\alpha^{-1}\mathbf{I})^{-1}(w-\mathbf{0})}
\end{aligned}
$$

$p(w)$ can be further simplified as:

$$
\begin{aligned}
p(w) &= \frac{1}{\sqrt{|2\pi\alpha^{-1}|}} exp^{-1/2(w^T)\alpha\mathbf{I}(w)} \\
p(w) &= \frac{\sqrt{\alpha}}{\sqrt{2\pi}} exp^{-1/2w^Tw}
\end{aligned}
$$

Inverse of identity matrix is Identity and $det|I| = 1$

ln of $p(w)$ can be written as:

$$
\begin{aligned}
ln(p(w)) &= ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}} exp^{-1/2\alpha w^T w}) \\
&= ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}}) - 1/2\alpha w^T w
\end{aligned}
$$

1.3 Posterior over **w**:

$$
\begin{aligned}
p(w|D) &= \frac{p(D|w)p(w|\alpha)}{p(D)} \\
p(w|x_n, y_n) &= \frac{p(t_n|x_n, w_n)p(w|\alpha)}{p(t|x)} \\
&= \frac{\mathcal{N}(t_n|w^T\phi_n, \beta^{-1})\mathcal{N}(w|\mathbf{0}, \alpha^{-1}\mathbf{I})}{\int_{-\infty}^{\infty} \mathcal{N}(w|\mathbf{0}, \alpha^{-1}\mathbf{I})\mathcal{N}(t_n|w^T\phi_n, \beta^{-1})dw}
\end{aligned}
$$

1.4 Calculate posterior:

$$\text{log-posterior} = ln\big(\mathcal{N}(w|\mathbf{0},\alpha^{-1}\mathbf{I})\mathcal{N}(t_n|w^T\phi_n,\beta^{-1})\big)$$

$ln(p(w))$ already calculated in 1.2 and is already in matrix for i.e w here is a vector

$$\begin{aligned}
&= ln(\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\beta^{-1}}}exp - \frac{(t_n - w^T\phi_n)^2}{2\beta^{-1}}) + ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}}) - 1/2\alpha w^T w \\
&= ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}}) - 1/2\alpha w^T w + ln(\frac{1}{(2\pi\beta^{-1})^{N/2}}) + \sum_{n=1}^{N}\frac{-(t_n - w^T\phi_n)^2}{2\beta^{-1}} \\
Constant \Rightarrow &= ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}}) + ln(\frac{1}{(2\pi\beta^{-1})^{N/2}}) \\
\text{log-posterior} &= -1/2\alpha w^T w + \sum_{n=1}^{N}\frac{-(t_n - w^T\phi_n)^2}{2\beta^{-1}}
\end{aligned}$$

Matix form can be re-written as:

$$\begin{aligned}
&= -1/2\alpha w^T w - \frac{(t-\Phi w)^2}{2\beta^{-1}} \\
&= -1/2\alpha w^T w - \frac{(t-\Phi w)^T(t-\Phi w)}{2\beta^{-1}}
\end{aligned}$$

Also at the same time calculating MAP is easier task than calculating the posterior distribution. In this example we can easily calculate MAP but in case of posterior distribution you may have calculating integral which may not be possible all the time.

1.5 Solve for w_{MAP}

$$\begin{aligned}
p(w|D) &= \frac{p(D|w)p(w)}{p(D)} \\
ln(p(w|D)) &= ln[\frac{p(D|w)p(w)}{p(D)}] \\
&= ln(p(w)) + ln(p(d|w)) - ln(p(D)) \\
{w_{MAP}}^* &= argmax(ln(p(w|D))) \\
{w_{MAP}}^* &= argmax(ln(p(w)) + ln(p(d|w)) - ln(p(D)))
\end{aligned}$$

$p(D)$ is independent of w so can be neglected.Also, $ln(p(w))$ is in matrix form

$$\begin{aligned}
&= argmax[ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}}) - 1/2\alpha w^T w + ln(\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\beta^{-1}}}exp - \frac{(t_n - w^T\phi_n)^2}{2\beta^{-1}})] \\
&= argamx(ln(\frac{\sqrt{\alpha}}{\sqrt{2\pi}}) - 1/2\alpha w^T w + Nln(\frac{1}{\sqrt{2\pi\beta^{-1}}}) + \sum_{n=1}^{N}\frac{-(t_n - w^T\phi_n)^2}{2\beta^{-1}}) \\
w_{MAP} &= \frac{{w_{MAP}}^*}{dw} = 0
\end{aligned}$$

dropping all those terms independent of w

$$\begin{aligned}
&= argmax(-1/2\alpha(w^T + w^T) + \sum_{n=1}^{N} -2(t_n - w^T\phi_n)(\phi_n^T)) \\
&= argmin(+1/2\alpha(w^T + w^T) + \sum_{n=1}^{N}\frac{2(t_n - w^T\phi_n)(-\phi_n^T)}{2\beta^{-1}}) \\
&= \alpha w^T - \frac{1}{\beta^{-1}}\sum_{n=1}^{N} t_n\phi_n^T + \frac{1}{\beta^{-2}}\sum_{n=1}^{N} w^T\phi_n\phi_n^T \\
\frac{1}{\beta^{-2}}\sum_{n=1}^{N} t_n\phi_n^T &= 2\alpha w^T + \frac{w^T}{\beta^{-1}}\sum_{n=1}^{N}\phi_n\phi_n^T \\
\frac{1}{\beta^{-1}}t^T\Phi &= w^T(\alpha I + \frac{\Phi^T\Phi}{\beta^{-1}})
\end{aligned}$$

Taking Transpose both sides

$$\Phi^T t = (\beta^{-1}\alpha I + \Phi^T\Phi)^T w$$
$$\Phi^T t = (\beta^{-1}\alpha I + \frac{\Phi^T\Phi}{\beta^{-1}})w$$
$$(\beta^{-1}\alpha I + \Phi^T\Phi)^{-1}\Phi^T t = w$$
$$w_{MAP} = (\beta^{-1}\alpha I + \Phi^T\Phi)^{-1}\Phi^T t$$

Same can be done from matrix form as well

$$\text{log-posterior} = -1/2\alpha w^T w - \frac{(t-\Phi w)^T(t-\Phi w)}{2\beta^{-1}}$$

Simplify it further:

$$= -1/2\alpha w^T w - \frac{(t^T - w^T\Phi^T)^T(t-\Phi w)}{2\beta^{-1}}$$
$$= -1/2\alpha w^T w - \frac{(t^T t - t^T\Phi w - w^T\Phi^T t + w^T\Phi^T\Phi w)}{2\beta^{-1}}$$

Taking derivative and equating to zero:

$$= -\alpha w^T - \frac{-2t^T\Phi + 2w^T\Phi^T\Phi}{2\beta^{-1}}$$
$$w^T\alpha\beta^{-1} + w^T\Phi^T\Phi = t^T\Phi$$

Taking transpose and inverse

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

6. Imagine, we have a function $f(x) = a.x$. Now, obviously, $f(x)$doesn't have a constant term (the constant term in neural network and ML is called bias). Therefore, $f(x)$ always passes through origin $(0, 0)$. But, what if the line that we are modeling doesn't pass through origin, rather it intersect with the $x - axis$ at c; i.e. $(0, b)$. The bias term is added to give this capability to our function. So in this case, $f(x) = a.x$ and we hope that after training, the bias term b will be close to c rather than forcing it to be zero.Below figure summarizes what I am trying to say.

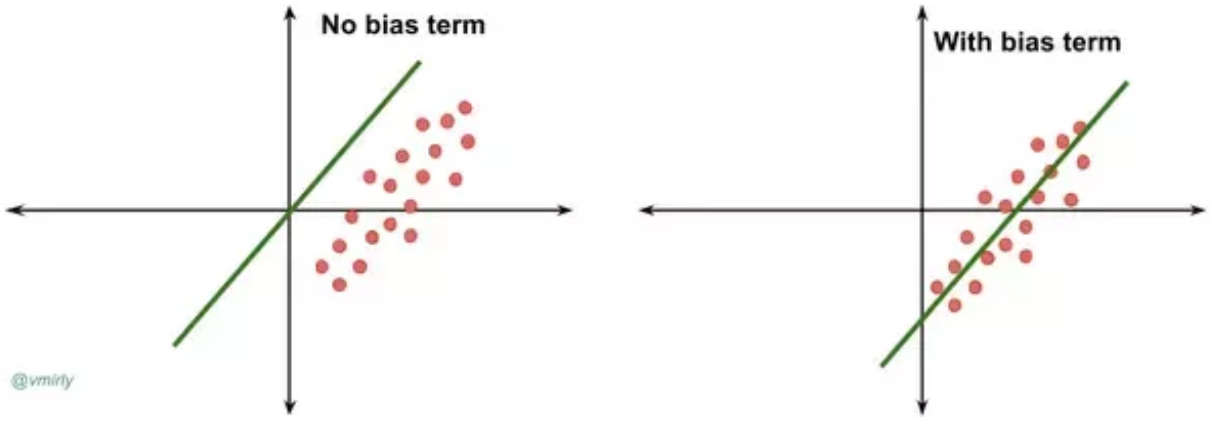

Abbildung 1: fig:taken from **Quora**

2 Probability distributions, likelihoods, and estimators

2.1 First Subtask

1. **Various Distributions** : Definitions below have been taken from **WIKIPEDIA and other online materials.**

a) It is the probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$ i.e., the probability distribution of any single experiment that asks a yes-no question; the question results in a boolean-valued outcome, a single bit of information whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q. It can be used to represent a coin toss where 1 and 0 would represent "headänd "tail"(or vice versa), respectively.

The Bernoulli distribution is a special case of the **binomial distribution** where a single experiment/trial is conducted (n=1). It is also a special case of the two-point distribution, for which the outcome need not be a bit, i.e., the two possible outcomes need not be 0 and 1.

b) In probability theory and statistics, the beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ parametrized by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution.
The beta distribution has been applied to model the behavior of random variables limited to intervals of finite length in a wide variety of disciplines.Beta distribution is conjugate prior to Bernoulli distribution.In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions.

c) The Poisson distribution is a discrete probability distribution that expresses the probability of occurrence of a given number of events in a fixed interval of time, space provided the events occur with an average and are independent events. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space.

d)The Gamma Distribution is a continuous probability distribution that arises naturally in process for which the waiting time between the events are relevant. It can be thought of as a waiting time between Poisson distributed events. Example: the traffic at the gate of science park at any time. It is conjugate prior to Poisson.

e) The normal distribution is a continuous distribution where data tends to be around a central value with no bias left or right like bell curve. Example: The blood pressure closely follows a normal distribution. It is conjugate prior to itself.

f) In probability theory, a log-normal (or log-normal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Example: the number of comment posed in the internet discussion forums follows a ln-normal distribution.

2.2 Second Subtask

1. Likelihood $L(\rho) = f(r_t; \rho)$ where $f(r_t; \rho)$ probability distribution. Likelihood of a single observation r_t is given by *Bernoulli distribution*:

$$L(\rho) = \rho^{r_t=1}(1-\rho)^{r_t=0}$$

For entire observation it will be product of each event as they are i.i.d's

$$L(\rho) = \prod_{t=1}^{N} \rho^{r_t=1}(1-\rho)^{r_t=0}$$

$r_t = 1$ or $r_t = 0$

2. ln-likelihood for the entire set of observations.

$$
\begin{aligned}
ln(L(\rho)) &= ln\left(\prod_{t=1}^{N} \rho^{r_t=1}(1-\rho)^{r_t=0}\right) \\
&= ln\big(\rho^{\sum_{t=1}^{N} r_t=1}\big) + ln\big((1-\rho)^{\sum_{t=1}^{N} r_t=0}\big) \\
\Rightarrow \quad & \underbrace{\sum_{t=1}^{N} r_t = 1}_{\text{total no of day where } r_t == 1} \\
\Rightarrow \quad & \underbrace{\sum_{t=1}^{N} r_t = 0}_{\text{total no of day where } r_t == 0} \\
&= ln(\rho^{n_1}) + ln((1-\rho)^{n_o}) \\
&= n_1 ln(\rho) + n_o ln(1-\rho)
\end{aligned}
$$

3.Solve for the maximum likelihood (ML) estimate of ρ.

$$
\begin{aligned}
\frac{ln(L(\rho))}{d\rho_{ML}} &= \frac{d}{d\rho}\left(n_1 ln(\rho) + n_o ln(1-\rho)\right) = 0 \\
&= \frac{n_1}{\rho} - \frac{n_o}{1-\rho} \\
n_1\rho &= n_o(1-\rho) \\
n_1 &= (n_1+n_o)\rho \\
\rho &= \frac{n_1}{n_1+n_o}
\end{aligned}
$$

$n_1 + n_o = N$

$$
\rho_{ML} = \frac{n_1}{N}
$$

plugging in the numbers

$$
\rho_{ML} = \frac{207}{365}
$$

4. Solve for the MAP estimate for ρ

$$
\begin{aligned}
p(\rho|r_t) &\propto p(r_t|\rho)p(\rho) \\
\rho_{MAP} &\propto argmax(ln(p(r_t|\rho)p(\rho))) \\
\rho_{MAP} &\propto argmax\Big(ln\left(\rho^{n_1}(1-\rho)^{n_o}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}\right)\Big)
\end{aligned}
$$

Adding all power terms

$$
\begin{aligned}
&\propto argmax\Big(ln\left(\rho^{n_1+a-1}(1-\rho)^{n_o+b-1}\right) + ln\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)\Big) \\
&\propto argmax((n_1+a-1)ln(\rho) + (n_o+b-1)ln(1-\rho) + \mathbf{C})
\end{aligned}
$$

C: denotes constant terms based ob a and b.Taking derivative wrt to ρ_{MAP} and equating to zero

$$
\begin{aligned}
0 &= \frac{(n_1+a-1)}{\rho} - \frac{n_o+b-1}{1-\rho} \\
(n_1+a-1)(1-\rho) &= (n_o+b-1)\rho \\
\rho &= \frac{(n_1+a-1)}{n_1+n_o+a+b-2} \\
\Rightarrow \rho_{MAP} &= \frac{n_1+a-1}{n_1+n_o+a+b-2}
\end{aligned}
$$

5. Posterior distribution for ρ.

$$
\begin{aligned}
p(\rho|r_t) &= \frac{p(r_t|\rho)p(\rho)}{p(r_t)} \\
p(r_t|\rho) &= \rho^{n_1}(1-\rho)^{n_o} \\
p(\rho) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1} \\
p(r_t) &= \int p(r_t|\rho)p(\rho)d\rho
\end{aligned}
$$

6. Taking values from above.

$$
\begin{aligned}
p(r_t) &= \int p(r_t|\rho)p(\rho)d\rho \\
&= \int \rho^{n_1}(1-\rho)^{n_o}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1}d\rho \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int \rho^{n_1+a-1}(1-\rho)^{n_o+b-1}d\rho \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(n_o+b)\Gamma(n_1+a)}{\Gamma(n_1+a+n_o+b)}\int \frac{\Gamma(n_1+a+n_o+b)}{\Gamma(n_1+a)\Gamma(n_o+b)}\rho^{n_1+a-1}(1-\rho)^{n_o+b-1}d\rho
\end{aligned}
$$

Integral goes to 1.

$$
\begin{aligned}
p(r_t) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(n_o+b)\Gamma(n_1+a)}{\Gamma(n_1+a+n_o+b)} \\
p(r_t|\rho)p(\rho) &= \rho^{n_1}(1-\rho)^{n_o}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\rho^{a-1}(1-\rho)^{b-1} \\
p(\rho|r_t) &= \frac{p(r_t|\rho)p(\rho)}{p(r_t)} \\
p(\rho|r_t) &= \frac{\Gamma(n_1+a+n_o+b)}{\Gamma(n_1+a)\Gamma(n_o+b)}\rho^{n_1+a-1}(1-\rho)^{n_o+b-1}
\end{aligned}
$$

2.3 Third Subtask

1. Likelihood for a single observation:

$$L(d_t:\lambda) = e^{-\lambda}\frac{\lambda^{d_t}}{\eta}$$

d_t here in this case represents a single observation. But in case if we have more than one observations in our observation set liked $d_1, d_2.....d_N$, in such case likelihood for entire observations can be written as:

$$
\begin{aligned}
L(d_t:\lambda) &= \prod_{t=1}^{T} e^{-\lambda}\frac{\lambda^{d_t}}{d_t!} \\
&= e^{-T\lambda}\frac{\lambda^{\sum_{t=1}^{T} d_t}}{\prod_{t=1}^{T} d_t!} \\
L(d_t:\lambda) &= e^{-T\lambda}\frac{\lambda^{n}}{\prod_{t=1}^{T} d_t!}
\end{aligned}
$$

2.Compute the ln-likelihood

$$
\begin{aligned}
ln(L(d_t:\lambda)) &= ln(e^{-T\lambda}\frac{\lambda^{n}}{\prod_{t=1}^{T} d_t!}) \\
&= ln(e^{-T\lambda}) + ln(\frac{\lambda^{n}}{\prod_{t=1}^{T} d_t!}) \\
&= -T\lambda + nln(\lambda) - ln(\prod_{t=1}^{T} d_t!)
\end{aligned}
$$

3. Solve for the ML estimate of λ.

$$\frac{ln(L(d_t:\lambda))}{d\lambda} = 0$$
$$-T + \frac{n}{\lambda} = 0$$
$$\lambda = \frac{n}{T}$$

4. Compute the MAP estimate of λ the general case:

$$\begin{aligned}
\lambda_{MAP} &\propto argmax(ln(p(d_t|\lambda)) + ln(p(\lambda))) \\
&\propto argmax\bigg(-T\lambda + nln(\lambda) - ln(\prod_{t=1}^{T} d_t!) + ln\big(\frac{a^b}{\Gamma b}\lambda^{b-1}e^{-a\lambda}\big)\bigg) \\
&\propto argmax\bigg(-T\lambda + nln(\lambda) - ln(\prod_{t=1}^{T} d_t!) + ln\big(\frac{a^b}{\Gamma b}\lambda^{b-1}e^{-a\lambda}\big)\bigg) \\
&\propto argmax\bigg(-T\lambda + nln(\lambda) + (b-1)ln(\lambda) - a\lambda + C_{independent(\lambda)}\bigg)
\end{aligned}$$

Take derivative and equate to zero

$$\begin{aligned}
0 &= -T + \frac{n}{\lambda} + \frac{b-1}{\lambda} - a \\
T + a &= \frac{n+b-1}{\lambda} \\
\lambda &= \frac{n+b-1}{T+a}
\end{aligned}$$

5. Posterior distribution for λ.

$$\begin{aligned}
p(\lambda/d_t) &= \frac{p(d_t/\lambda)p(\lambda)}{p(d_t)} \\
p(d_t) &= \int e^{-T\lambda}\frac{\lambda^n}{\prod_{t=1}^{T} d_t!}\frac{a^b}{\Gamma(b)}\lambda^{b-1}e^{-a\lambda}d(\lambda) \\
p(d_t/\lambda)p(\lambda) &= \frac{a^b}{\Gamma(b)\prod_{t=1}^{T} d_t!}e^{-(T+a)\lambda}\lambda^{n+b-1}
\end{aligned}$$

6. Solve analytically.

$p(d_t)$ can written as:

$$p(d_t) = \int e^{-T\lambda}\frac{\lambda^n}{\prod_{t=1}^{T} d_t!}\frac{a^b}{\Gamma(b)}\lambda^{b-1}e^{-a\lambda}d(\lambda)$$

adding powers and taking constant term outside.

$$= \frac{a^b}{\Gamma(b)\prod_{t=1}^{T} d_t!}\int e^{-(T+a)\lambda}\lambda^{n+b-1}d\lambda$$

Looks like Gamma but need to multiply and divide by new params to make integral 1

$$= \frac{\Gamma(n+b)}{(T+a)^{(n+b)}}\frac{a^b}{\Gamma(b)\prod_{t=1}^{T} d_t!}\int \frac{(T+a)^{(n+b)}}{\Gamma(n+b)}e^{-(T+a)\lambda}\lambda^{n+b-1}d\lambda$$

intergral goes to 1, denominator we have constant terms.

$$p(d_t/\lambda)p(\lambda) = \frac{a^b}{\Gamma(b)\prod_{t=1}^{T} d_t!}e^{-(T+a)\lambda}\lambda^{n+b-1}$$

numerator and denominator cancel out terms and we are left with Gamma distribution

$$p(\lambda|d_t) = \frac{(T+a)^{(n+b)}}{\Gamma(n+b)} e^{-(T+a)\lambda} \lambda^{(n+b-1)}$$