Machine Learning:Data to Models
Assignment 1a: Bayesian Linear Regression

Qun Gao, JHED ID qgao6:
Li-Yi Lin, JHED ID: llin34

Problem
(a) Answer:
$P(\mathbf{w}|\mathbf{X}, y, \sigma^2) \propto P(y|\mathbf{X}, \mathbf{w}, \sigma^2)P(\mathbf{w})$
$= \mathcal{N}(y|\mathbf{X}\mathbf{w}, \sigma^2 I_n)\mathcal{N}(\mathbf{w}|\mu_0, \Sigma_0)$
$\propto \exp[-\frac{1}{2}(y - \mathbf{X}\mathbf{w})^T(\sigma^2 I_n)^{-1}(y - \mathbf{X}\mathbf{w})] \exp[-\frac{1}{2}(\mathbf{w} - \mu_0)^T\Sigma_0^{-1}(\mathbf{w} - \mu_0)]$
$= \exp[-\frac{1}{2}(y - \mathbf{X}\mathbf{w})^T(\sigma^2 I_n)^{-1}(y - \mathbf{X}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mu_0)^T\Sigma_0^{-1}(\mathbf{w} - \mu_0)]$
$\propto \exp[-\frac{1}{2\sigma^2}(y^T y - \mathbf{w}^T\mathbf{X}^T y - y^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) - \frac{1}{2}\mathbf{w}^T\Sigma_0^{-1}\mathbf{w}) + \frac{1}{2}2\mu_0\Sigma_0^{-1}\mathbf{w}]$
$= \exp[-\frac{1}{2\sigma^2}y^T y + \frac{1}{\sigma^2}y^T\mathbf{X}\mathbf{w} - \frac{1}{2\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - \frac{1}{2}\mathbf{w}^T\Sigma_0^{-1}\mathbf{w} + \mu_0\Sigma_0^{-1}\mathbf{w}]$
$= \exp[-\frac{1}{2}(\frac{1}{\sigma^2}y^T y - (\frac{2}{\sigma^2}y^T\mathbf{X}\mathbf{w} + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\Sigma_0^{-1}\mathbf{w}))]$
$= e^{-\frac{1}{2}[\mathbf{w}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\Sigma_0^{-1})\mathbf{w}-(\frac{2}{\sigma^2}y^T\mathbf{X}\mathbf{w}+\frac{1}{\sigma^2}y^T y+2\mu_0^T\Sigma_0^{-1}\mathbf{w})]}$
$= e^{-\frac{1}{2}[\mathbf{w}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\Sigma_0^{-1})\mathbf{w}-(\frac{2}{\sigma^2}y^T\mathbf{X}+2\mu_0^T\Sigma_0^{-1})\mathbf{w}+\frac{1}{\sigma^2}y^T y]}$
$\propto \mathcal{N}(\mathbf{w}|\mu_\mathbf{w}, \Sigma_\mathbf{w})$
Where $\mu_\mathbf{w} = \Sigma_\mathbf{w}(\frac{1}{\sigma^2}\mathbf{X}^T y + \Sigma_0^{-1}\mu_0)$ and $\Sigma_\mathbf{w} = \sigma^2(\sigma^2\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}$
Next, we are going to prove it.

By examining the power of exponent shown above, we have
$(\mathbf{w} - \mu_\mathbf{w})^T\Sigma_\mathbf{w}^{-1}(\mathbf{w} - \mu_\mathbf{w})$
$= \mathbf{w}^T\Sigma_\mathbf{w}^{-1}\mathbf{w} - \mu_\mathbf{w}^T\Sigma_\mathbf{w}^{-1}\mathbf{w} - \mathbf{w}^T\Sigma_\mathbf{w}^{-1}\mu_\mathbf{w} + \mu_\mathbf{w}^T\Sigma_\mathbf{w}^{-1}\mu_\mathbf{w}$
$= \mathbf{w}^T\Sigma_\mathbf{w}^{-1}\mathbf{w} - 2\mu_\mathbf{w}^T\Sigma_\mathbf{w}^{-1}\mathbf{w} + \mu^T\Sigma_\mathbf{w}^{-1}\mu_\mathbf{w}$
Compared with above equation, we have
$\Sigma_\mathbf{w}^{-1} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \Sigma_0^{-1}$
$\Sigma_\mathbf{w} = \sigma^2(\sigma^2\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}$

$\mu_\mathbf{w}^T\Sigma_\mathbf{w}^{-1} = \frac{1}{\sigma^2}y^T\mathbf{X} + \mu_0^T\Sigma_0^{-1}$
$\mu_\mathbf{w}^T = (\frac{1}{\sigma^2}y^T\mathbf{X} + \mu_0^T\Sigma_0^{-1})\Sigma_\mathbf{w}$
$\mu_\mathbf{w} = \Sigma_\mathbf{w}(\frac{1}{\sigma^2}\mathbf{X}^T y + \Sigma_0^{-1}\mu_0)$

Thus, we have prove that $\mu_\mathbf{w} = \Sigma_\mathbf{w}(\frac{1}{\sigma^2}\mathbf{X}^T y + \Sigma_0^{-1}\mu_0)$ and $\Sigma_\mathbf{w} = \sigma^2(\sigma^2\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}$

(b) Answer:
$P(y_{n+1}|x_{n+1}, X, \sigma^2)$
$= \int N(y_{n+1}|\mathbf{x_{n+1}^T}\mathbf{w}, \sigma^2)N(\mathbf{w}|\mu_\mathbf{w}, \Sigma_w)d\mathbf{w}$
$\propto \int exp[-\frac{1}{2}(y_{n+1} - \mathbf{x_{n+1}^T}\mathbf{w})^T\frac{1}{\sigma^2}(y_{n+1} - \mathbf{x_{n+1}^T}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mu_\mathbf{w})^T\Sigma_w^{-1}(\mathbf{w} - \mu_\mathbf{w})]d\mathbf{w}$
$\propto \int exp[-\frac{1}{2\sigma^2}(y_{n+1}^T y_{n+1} - 2y_{n+1}^T\mathbf{x_{n+1}^T}\mathbf{w} + \mathbf{w^T}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T}\mathbf{w})$
$-\frac{1}{2}(\mathbf{w^T}\Sigma_w^{-1}\mathbf{w} - 2\mu_\mathbf{w}^\mathbf{T}\Sigma_w^{-1}\mathbf{w})]d\mathbf{w}$
$\propto exp[-\frac{1}{2\sigma^2}y_{n+1}^T y_{n+1} + \frac{1}{2}\mu_\mathbf{1}^\mathbf{T}\Sigma_1^{-1}\mu_\mathbf{1}] \int exp[-\frac{1}{2}(\mathbf{w} - \mu_\mathbf{1})^T\Sigma_1^{-1}(\mathbf{w} - \mu_\mathbf{1})]d\mathbf{w}$
where $\mu_\mathbf{1} = \Sigma_1(\frac{1}{\sigma^2}\mathbf{x_{n+1}}y_{n+1} + \Sigma_w^{-1}\mu_\mathbf{w})$, $\Sigma_1^{-1} = \frac{1}{\sigma^2}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T} + \Sigma_w^{-1}$.
Since $exp[-\frac{1}{2}(\mathbf{w} - \mu_\mathbf{1})^T\Sigma_1^{-1}(\mathbf{w} - \mu_\mathbf{1})] \propto N(\mathbf{w}|\mu_\mathbf{1}, \Sigma_1)$, we can make the integral in the last step to be 1 over $(-\infty, +\infty)$ with some constant we ignored. So we have
$\propto exp[-\frac{1}{2\sigma^2}y_{n+1}^T y_{n+1} + \frac{1}{2}\mu_\mathbf{1}^\mathbf{T}\Sigma_1^{-1}\mu_\mathbf{1}]$ (1)

Here
$\mu_{\mathbf{1}}^{\mathbf{T}}\Sigma_1^{-1}\mu_{\mathbf{1}}$
$= (\frac{1}{\sigma^2}\mathbf{x_{n+1}}y_{n+1} + \Sigma_w^{-1}\mu_{\mathbf{w}})^T\Sigma_1\Sigma_1^{-1}\Sigma_1(\frac{1}{\sigma^2}\mathbf{x_{n+1}}y_{n+1} + \Sigma_w^{-1}\mu_{\mathbf{w}})$
$= (\frac{1}{\sigma^2}y_{n+1}^T\mathbf{x_{n+1}^T} + \mu_{\mathbf{w}}^{\mathbf{T}}\Sigma_w^{-1})\Sigma_1(\frac{1}{\sigma^2}\mathbf{x_{n+1}}y_{n+1} + \Sigma_w^{-1}\mu_{\mathbf{w}})$
$\propto \frac{1}{\sigma^4}y_{n+1}^T\mathbf{x_{n+1}^T}\Sigma_1\mathbf{x_{n+1}}y_{n+1} + \frac{2}{\sigma^2}\mu_{\mathbf{w}}^{\mathbf{T}}\Sigma_w^{-1}\Sigma_1\mathbf{x_{n+1}}y_{n+1}$
Then (1) would be
$\propto exp[-\frac{1}{2}(\frac{1}{\sigma^2}y_{n+1}^Ty_{n+1} - \frac{1}{\sigma^4}y_{n+1}^T\mathbf{x_{n+1}^T}\Sigma_1\mathbf{x_{n+1}}y_{n+1} - \frac{2}{\sigma^2}\mu_{\mathbf{w}}^{\mathbf{T}}\Sigma_w^{-1}\Sigma_1\mathbf{x_{n+1}}y_{n+1})]$ (2)
$\propto exp[-\frac{1}{2}(y_{n+1} - \mu_y)^T\Sigma_y^{-1}(y_{n+1} - \mu_y)]$
Here $\mu_y = \mu_{\mathbf{w}}^{\mathbf{T}}\mathbf{x_{n+1}}, \Sigma_y = \sigma^2 + \mathbf{x_{n+1}^T}\Sigma_w\mathbf{x_{n+1}}$. Now we prove it.
Since $(y_{n+1} - \mu_y)^T\Sigma_y^{-1}(y_{n+1} - \mu_y) = y_{n+1}^T\Sigma_y^{-1}y_{n+1} - 2\mu_y^T\Sigma_y^{-1}y_{n+1}$ (3), from (2) and (3), we have
$\Sigma_y^{-1} = \frac{1}{\sigma^2} - \frac{1}{\sigma^4}\mathbf{x_{n+1}^T}\Sigma_1\mathbf{x_{n+1}}$ (4), $\mu_y^T\Sigma_y^{-1} = \frac{1}{\sigma^2}\mu_{\mathbf{w}}^{\mathbf{T}}\Sigma_w^{-1}\Sigma_1\mathbf{x_{n+1}}$.
So $\mu_y = \Sigma_y(\frac{1}{\sigma^2}\mathbf{x_{n+1}^T}\Sigma_1\Sigma_w^{-1}\mu_{\mathbf{w}}) = \frac{1}{\sigma^2}\mu_{\mathbf{w}}^{\mathbf{T}}\Sigma_w^{-1}\Sigma_1\mathbf{x_{n+1}}\Sigma_y$ (5).
We have computed Σ_1^{-1} above. $\Sigma_1^{-1} = \frac{1}{\sigma^2}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T} + \Sigma_w^{-1}$. Then we can compute
$\Sigma_y = (\frac{1}{\sigma^2} - \frac{1}{\sigma^4}\mathbf{x_{n+1}^T}\Sigma_1\mathbf{x_{n+1}})^{-1} = \sigma^4(\sigma^2 - \mathbf{x_{n+1}^T}\Sigma_1\mathbf{x_{n+1}})^{-1}$
$= \sigma^4(\frac{1}{\sigma^2} + \frac{1}{\sigma^2}\mathbf{x_{n+1}^T}(\Sigma_1^{-1} - \mathbf{x_{n+1}}\frac{\mathbf{1}}{\boldsymbol{\sigma^2}}\mathbf{x_{n+1}^T})^{-\mathbf{1}}\mathbf{x_{n+1}}\frac{\mathbf{1}}{\boldsymbol{\sigma^2}})$
$= \sigma^2 + \mathbf{x_{n+1}^T}(\Sigma_1^{-1} - \mathbf{x_{n+1}}\frac{\mathbf{1}}{\boldsymbol{\sigma^2}}\mathbf{x_{n+1}^T})^{-1}\mathbf{x_{n+1}}$
$= \sigma^2 + \mathbf{x_{n+1}^T}(\frac{1}{\sigma^2}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T} + \Sigma_w^{-1} - \frac{1}{\sigma^2}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T})^{-1}\mathbf{x_{n+1}}$
$= \sigma^2 + \mathbf{x_{n+1}^T}\Sigma_w\mathbf{x_{n+1}}$. So we have proved $\Sigma_y = \sigma^2 + \mathbf{x_{n+1}^T}\Sigma_w\mathbf{x_{n+1}}$.
To get $\mu_y = \frac{1}{\sigma^2}\mu_{\mathbf{w}}^{\mathbf{T}}\Sigma_w^{-1}\Sigma_1\mathbf{x_{n+1}}\Sigma_y = \mu_{\mathbf{w}}^{\mathbf{T}}\mathbf{x_{n+1}}$, that is to prove
$\mathbf{x_{n+1}} = \frac{1}{\sigma^2}\Sigma_w^{-1}\Sigma_1\mathbf{x_{n+1}}\Sigma_y$. That is to prove
$\Sigma_1^{-1}\Sigma_w\mathbf{x_{n+1}} = \frac{1}{\sigma^2}\mathbf{x_{n+1}}\Sigma_y$. Since
$\Sigma_1^{-1}\Sigma_w\mathbf{x_{n+1}} = (\frac{1}{\sigma^2}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T} + \Sigma_w^{-1})\Sigma_w\mathbf{x_{n+1}} = \frac{1}{\sigma^2}\mathbf{x_{n+1}}\mathbf{x_{n+1}^T}\Sigma_w\mathbf{x_{n+1}} + \mathbf{x_{n+1}}$
$= \frac{1}{\sigma^2}\mathbf{x_{n+1}}(\mathbf{x_{n+1}^T}\Sigma_w\mathbf{x_{n+1}} + \sigma^2)$
$= \frac{1}{\sigma^2}\mathbf{x_{n+1}}\Sigma_y$.
So we have proved $\mu_y = \mu_{\mathbf{w}}^{\mathbf{T}}\mathbf{x_{n+1}}$.
Now we can get that $P(y_{n+1}|x_{n+1}, X, \sigma^2) = N(y_{n+1}|\mu_y, \Sigma_y)$, where
$\mu_y = \mu_{\mathbf{w}}^{\mathbf{T}}\mathbf{x_{n+1}}, \Sigma_y = \sigma^2 + \mathbf{x_{n+1}^T}\Sigma_w\mathbf{x_{n+1}}$.

(c) Answer:
We could use point estimate on hyperparameters. $\hat{\eta} = argmax_\eta\ p(\eta|D)$, where η is hyperparameter, D is data.

(d) Answer:
Since the posterior predictive distribution contains integral, it takes a lot of time to compute. We can use plugin approximation to reduce time cost. We should use this approximation under the condition that when the dataset is not good, or the dataset's distribution is too broad. Because in this case, the dataset provides little information, and MAP will mostly depend on the prior.

(e) Answer:

For model $\mathcal{M}_0$, the covariance matrix of the prior distribution is $\Sigma_0 = \begin{bmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & 0 \\ 0 & 0 & 0 & \sigma_0^2 \end{bmatrix}$

For model $\mathcal{M}_{AB}$, the covariance matrix of the prior distribution is $\Sigma_{AB} = \begin{bmatrix} \sigma_0^2 & \gamma_0^2 & 0 & 0 \\ \gamma_0^2 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & 0 \\ 0 & 0 & 0 & \sigma_0^2 \end{bmatrix}$

For model $\mathcal{M}_{CD}$, the covariance matrix of the prior distribution is $\Sigma_{CD} = \begin{bmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & \gamma_0^2 \\ 0 & 0 & \gamma_0^2 & \sigma_0^2 \end{bmatrix}$

(f) Answer:
Since $P(y_{1:n}|x_{1:n}, M_i) = P(\mathbf{y}|X, M_i) = \int N(\mathbf{y}|X\mathbf{w}, \sigma^2 I_n)N(\mathbf{w}|\mu_{\mathbf{0}}, \Sigma_i)d\mathbf{w}$, this form is almost the same as that in (b). So with the same derivation, the result is $N(\mathbf{y}|\mu_{\mathbf{y}}, \mathbf{\Sigma_y})$, where $\mu_{\mathbf{y}} = X\mu_0, \mathbf{\Sigma_y} = \sigma^2 I + X\Sigma_i X^T$.

(g) Answer:
From (f), we know that

$$P(\mathbf{y}|\mathbf{X}, \mathcal{M}_i) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mu_0, \sigma^2 I_n + \mathbf{X}\Sigma_i\mathbf{X}^T)$$
$$= (2\pi)^{-\frac{n}{2}}|\Gamma_i|^{-1}\exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mu_0)^T\Gamma_i^{-1}(\mathbf{y} - \mathbf{X}\mu_0)]$$

where $\Gamma_i = \sigma^2 I_n + \mathbf{X}\Sigma_i\mathbf{X}^T$

When we directly computed the posterior, we encountered an overflow problem and it also took a long time before we got the overflow problem. The reason is that Γ is a 10000x10000 matrix in this problem. In order to simply the computations, $|\Gamma|$ and Γ^{-1}, of Γ matrix, we adopted Corollary 4.3.1 Matrix Inversion Lemma in the Murphy's machine learning textbook.

The equation (4.106) shows that

$$(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1}$$

Let $E = \sigma^2 I_n$, $F = \mathbf{X}$, $H = -\Sigma_i^{-1}$, $G = X^T$, then we have

$$\Gamma_i^{-1} = (\sigma^2 I_n + \mathbf{X}\Sigma_i\mathbf{X}^T)^{-1}$$
$$= \sigma^{-2}I_n + \sigma^{-4}\mathbf{X}(-\Sigma_i^{-1} - \mathbf{X}^T\sigma^{-2}I_n\mathbf{X})^{-1}\mathbf{X}^T$$

The equation (4.108) shows that

$$|E - FH^{-1}G| = |H - GE^{-1}F||H^{-1}||E|$$

Then we have

$$|\Gamma_i| = |\sigma^2 I_n + \mathbf{X}\Sigma_i\mathbf{X}^T|$$
$$= |-\Sigma_i^{-1} - \mathbf{X}^T\sigma^{-2}I_n\mathbf{X}||-\Sigma_i||\sigma^2 I_n|$$

Since the posterior is proportional to the likelihood times the prior and we assume a uniform prior over the three models, we only have to compare their likelihood. In addition, to avoid the numerical problem that the high power of a number will cause the likelihood become extremely small, we instead calculated their log-likelihood as shown below:

$$\log P(\mathbf{y}|\mathbf{X}, \mathcal{M}_i) = -\frac{n}{2}\log 2\pi - \log|\Gamma_i| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\mu_0)^T\Gamma^{-1}(\mathbf{y} - \mathbf{X}\mu_0)$$
$$= -\frac{n}{2}\log 2\pi - n\log\sigma^2 + \log|-\Sigma_i^{-1} - \mathbf{X}^T\sigma^{-1}I_n\mathbf{X}| + \log|-\Sigma_i| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\mu_0)^T\Gamma^{-1}(\mathbf{y} - \mathbf{X}\mu_0)$$

Table 1: log-likelihood

model	log-likelihood
$\mathcal{M}_0$	-113939.236643
$\mathcal{M}_{AB}$	-113939.248511
$\mathcal{M}_{CD}$	-113939.248698

The result of log-likelihood for the three models are shown in Table 1. For choosing a model, we should pick the model with the largest log-likelihood, i.e. $\mathcal{M}_0$ in this problem.

(h) Answer:

$\epsilon \sim \theta\mathcal{N}(0, \sigma^2) + (1-\theta)\mathcal{N}(0, 50) = \mathcal{N}(0, \theta^2\sigma^2 + 50(1-\theta)^2)$

Let $\delta^2 = \theta^2\sigma^2 + 50(1-\theta)^2$

From (a), we know that

$P(\mathbf{w}|\mathbf{X}, y, \sigma^2) \propto P(y|\mathbf{X}, \mathbf{w}, \sigma^2)P(\mathbf{w})$

$= \mathcal{N}(y|\mathbf{X}\mathbf{w}, \sigma^2 I_n)\mathcal{N}(\mathbf{w}|\mu_0, \Sigma_0)$

$= \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}})$, where $\mu_{\mathbf{w}} = \Sigma_{\mathbf{w}}(\frac{1}{\sigma^2}\mathbf{X}^T y + \Sigma_0^{-1}\mu_0)$ and $\Sigma_{\mathbf{w}} = \sigma^2(\sigma^2\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}$.

But in this problem part, $y \sim \mathcal{N}(y|\mathbf{X}\mathbf{w}, \delta^2 I_N)$

So the new posterior distribution is

$P(\mathbf{w}|\mathbf{X}, y, \sigma^2) \propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \Sigma_0)\mathcal{N}(y|\mathbf{X}\mathbf{w}, \delta^2 I_N) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}^{'}, \Sigma_{\mathbf{w}}^{'})$

where $\mu_{\mathbf{w}}^{'} = \mu_{\mathbf{w}} = \Sigma_{\mathbf{w}}(\frac{1}{\sigma^2}\mathbf{X}^T y + \Sigma_0^{-1}\mu_0)$ and $\Sigma_{\mathbf{w}}^{'} = \delta^2(\delta^2\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}$, $\delta^2 = \theta^2\sigma^2 + 50(1-\theta)^2$.