Models

Shanglin Zou[1] and Junfan Huang[2]

[1] 98078
[2] 78046

January 2, 2022

1 The Prior

1.1 The theory

1.1.1 Question 1

1. A Gaussian likelihood encodes an assumption that $\boldsymbol{X}$,$\boldsymbol{Y}$ have linear(or non-linear) relationship with noise $\varepsilon \sim N(0, \sigma^2)$

$$p(\varepsilon_i) = p(y_i - f(x_i)) = p(y_i|f, x_i) = N(f(x_i), \sigma^2 I)$$

2. It means that data points are independent and identically distributed.

1.1.2 Question 2

$$\begin{aligned} p(\boldsymbol{Y}|f, \boldsymbol{X}) &= p(y_1, y_2, ..., y_N|f, x_1, x_2, ..., x_N) \\ &= p(y_1|f, \boldsymbol{X})...p(y_N|y_N - 1, ...y_1, f, \boldsymbol{X}) \end{aligned}$$

1.1.3 Question 3

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{W}, x_i)$$

1.1.4 Question 4

The prior and posterior distributions are in the same distribution family, then it called conjugate distribution. The posterior distribution is proportional to likelihood times prior, so we can calculate posterior distribution from this relationship.

1.1.5 Question 5

The figure 1 shows the distributions of w1 and w2. Imagine that the normal distribution curve sits on the plane $w_1 w_2$ with mean $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, with variance τ^2. If we rotate the curve about point(0,0), then we can have many other independent and identically distribution which are normal distribution with mean $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, with variance τ^2. So, we can say that the prior distribution has a spherical co-vairance matrix.
$\boldsymbol{P(W)} = N(\boldsymbol{W_0}, \tau^2 \boldsymbol{I})$, the prior, as an punishment when optimizing the likelihood function, makes sure that the model would be overfitting.

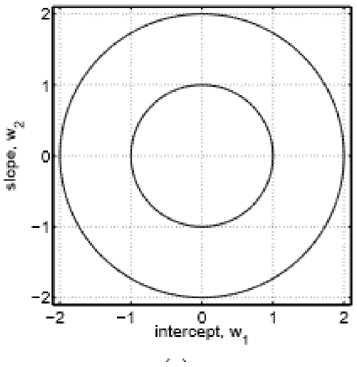


Figure 1: *The spherical Gaussian.*

1.1.6 Question 6

$$p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{x}, \boldsymbol{W}) \cdot p(\boldsymbol{W}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{W}, x_i) \cdot p(\boldsymbol{W})$$

$$\propto \exp^{\sigma^{-2} \sum_{i=1}^{N} (y_i - \boldsymbol{W}^T x_i)^T (y_i - \boldsymbol{W}^T x_i) + \tau^{-2} (\boldsymbol{W} - \boldsymbol{W}_0)^T (\boldsymbol{W} - \boldsymbol{W}_0)}$$

$$= \exp^{\sigma^{-2} (\boldsymbol{Y} - \boldsymbol{AW})^T (\boldsymbol{Y} - \boldsymbol{AW}) + \tau^{-2} (\boldsymbol{W} - \boldsymbol{W}_0)^T (\boldsymbol{W} - \boldsymbol{W}_0)}$$

$$\propto \exp^{\boldsymbol{W}^T (\sigma^{-2} \boldsymbol{A}^T \boldsymbol{A} + \tau^{-2} \boldsymbol{I}) \boldsymbol{W} - 2 \boldsymbol{W}^T (\sigma^{-2} \boldsymbol{A}^T \boldsymbol{y} + \tau^{-2} \boldsymbol{W}_0))}$$

Where $\boldsymbol{A}$ is a design matrix,

$$A = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nn} \end{bmatrix}$$

so we can have

$$\Lambda^{-1} = (\sigma^{-2}\boldsymbol{A}^T\boldsymbol{A} + \tau^{-2}\boldsymbol{I})$$
$$\mu = \Lambda \cdot (\sigma^{-2}\boldsymbol{A}^T\boldsymbol{y} + \tau^{-2}\boldsymbol{W}_0)$$

so posterior distribution follows $N(\mu, \Lambda)$

1.1.7 Question 7

In parametric models, we make an assumption about the distribution of the data, then we use the data from sample to estimate the parametric in this particular distribution. However, in the non-parametric models, we don't make any assumption about the distribution of the data. The non-parametric models focus on the the ranks or scores of the data, which seems to be more robust. In non-parametric models, the complexity of the model grows with the number of training data; in parametric models, we have a fixed number of parameters. For example, we want to generate the cars to the markets, but we have no idea what kinds of type people like. In that case, we should generate the cars, the more cars we producing the more data we getting. But it costs a lot. So it should have a balance between the interpretability of models and the models complexity.

1.1.8 Question 8

It represents a random Gaussian vector with zero-mean and this co-variance matrix, $K(X, X)$. θ is the hyper-parameter that controls the behaviour of the co-variance function. The co-variance function compute how much each location along the input axis co-varies with each other. The closer they are, the stronger relationship they have.

1.1.9 Question 9

It contains all possible functions we have assumed(hypothesis space). Every time given a data point, the interval of the possible functions is shrink.

1.1.10 Question 10

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{X}, f, \theta) &= p(\boldsymbol{Y}|f)p(f|\boldsymbol{X}, \theta)p(\boldsymbol{X}, \boldsymbol{\theta}) \\ &= p(\boldsymbol{Y}|f)p(f|\boldsymbol{X}, \theta)p(\boldsymbol{X})p(\boldsymbol{\theta}) \end{aligned}$$

- f is conditionally depends on $\boldsymbol{X}$ and θ
- $\boldsymbol{Y}$ is conditionally depends on f
- $\boldsymbol{X}$ and θ are independent

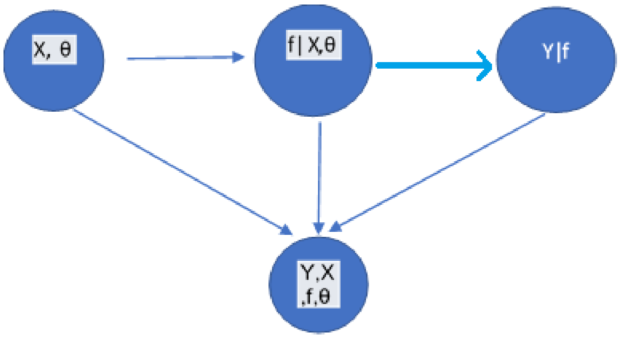


Figure 2: *Graphical model*

1.1.11 Question 11

1. The marginal likelihood is integral of the prior times the likelihood function. The likelihood function is obtained by data.
2.We only care about f being close to y,and not what is the function is, also we do not know what f is and what θ is, so we need to marginalise f out, and determine θ_hat by optimizing $p(\boldsymbol{Y}|\boldsymbol{X}, \theta)$
3.It implies that the hyper-parameter θ effects the relationship between X and Y.

1.1.12 Question 12

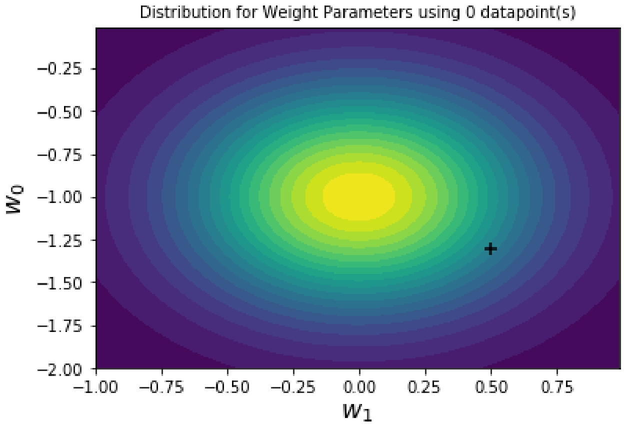
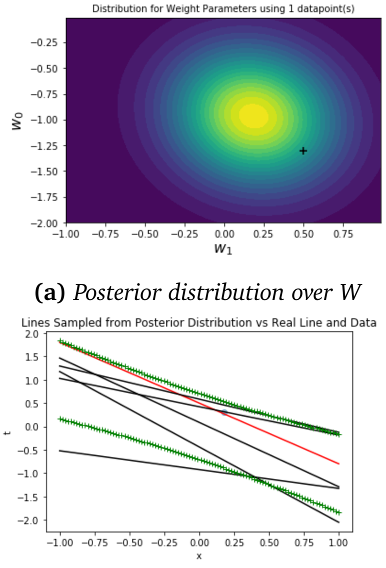


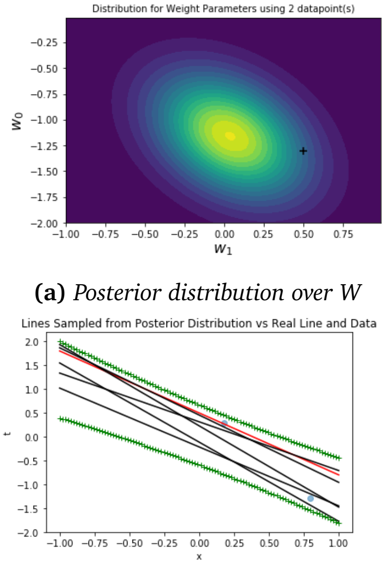
Figure 3: *The prior distribution.*



(a) *Posterior distribution over W*

(b) *Sample*

Figure 4: *After picking a single data-point*



(a) *Posterior distribution over W*

(b) *Sample*

Figure 5: *After picking two data-points*

5. The first graph shows that $W \sim N(\begin{bmatrix}-1\\0\end{bmatrix}, 5I)$. After adding data-points, the mean of the posterior will approach $\begin{bmatrix}-1.3\\0.5\end{bmatrix}$, and the circle will be smaller.
Yes, it is a desirable behavior.
6. Recall from Question 6, by adding many data, the element of co-variance is becoming smaller, Also, μ is becoming closer to $\boldsymbol{W}_0$.

1.1.13 Question 13

Figure 6 shows the samples from GP-prior with a squared exponential co-variance function by using different scales(ie.$l = 0.1, 1, 5$).

when we have smaller length-scale value, like the curve showed in figure when l=0.1, it changes up and down so rapidly, for larger value, it becomes more smoother.

The length-scale encodes the assumption that for squared exponential co-variance function, the smaller length-scale, the smoother prior function would be.

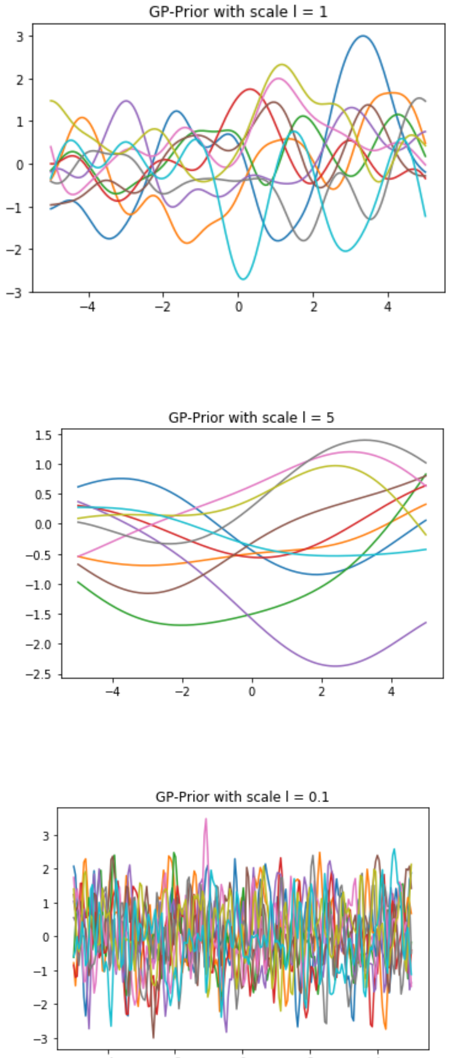


Figure 6: *GP-prior samples*

1.1.14 Question 14

Figure 7 shows the sampling from posterior distribution.
Comparing Figure 6 and Figure 7, it is very clearly that there is less uncertainty. Also, when it is out of range of our training data, the uncertainty is still large, since we know that the prior co-variance is squared exponential, which means if a point is far away from our observed data, then the uncertainty is large.
After adding a diagonal co-variance matrix, it becomes

more uncertain. We can clearly see from two plots from figure 7, the red fill is becoming larger if σ_y becomes larger.σ_y represents the amount of noise in the training data. Higher σ_y values make more coarse approximations which avoids over-fitting to noisy data.

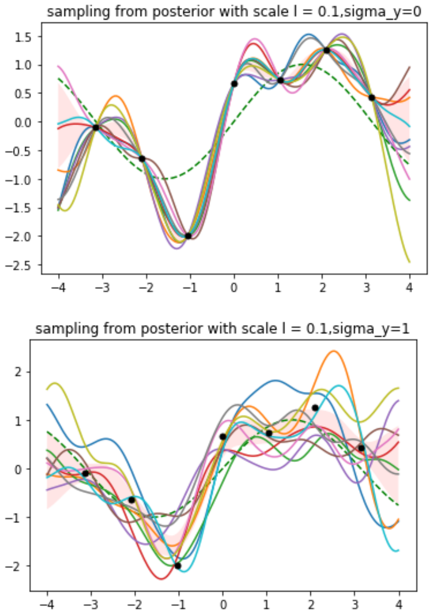


Figure 7: *GP-posterior samples*

2 Posterior

2.1 Theory

2.2 Question 15

In our opinions, building any model has the basis or assumptions. The prior belief, which represents our preference, has been made before some evidence is taken into account. For machine learning, the basic assumption underlying most of machine learning is that the available examples are independent and identically distributed. All the assumptions and belief and preference reflect our intuition. A good assumption at the beginning usually takes much more efficiency for solving problems.

2.3 Question 16

We assume that X is a Gaussian Distribution with zero mean and each elements in X are independently and identically distributed.

2.4 Question 17

$$
\begin{aligned}
p(\boldsymbol{Y}|\boldsymbol{W}) &= \int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W})p(\boldsymbol{X})\,dx \\
&= \int \prod_{i=1}^{N} p(y_i|\boldsymbol{W},x_i)\mathcal{N}(0,I)\,dx_i \\
&= \int \prod_{i=1}^{N} \mathcal{N}(y_i|Wx_i,\sigma^2 I)\mathcal{N}(0,I)\,dx_i \\
&= \mathcal{N}(\mu, WW^T+\sigma^2 I)
\end{aligned}
$$

Since $y = \boldsymbol{W}x+\mu+\epsilon$, then we can have $E(\boldsymbol{Y}|\boldsymbol{W}) = \mu, Cov(\boldsymbol{Y}|\boldsymbol{W}) = WW^T+\sigma^2 I$

2.4.1 Question 18

1. ML is to find parameters that maximize the likelihood function. MAP is to find the parameter that related to prior distribution that maximize posterior distribution.
2.When we are trying to maximize the formula, we use log. So the difference between MAP and ML is that the MAP method has a $log(p(W))$ compared with ML. The more data we observe, the more weight the likelihood function has. In that case, the MAP result is approaching the ML result. Let us take the flipping coins as an example, if my belief on the likelihood of getting head is 0.5 but the true probability is 0.7, as a frequentist I use all the data from the frequency, I will get the likelihood at 0.7. However, if I believe that all the coins have the 0.5 likelihood of getting head(as a prior) I will never get 0.7 just approaching it.
3.Because the formula $\int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W})p(\boldsymbol{W})d\boldsymbol{W}$ has no relationship with $\boldsymbol{W}$. What we want is that find the $\boldsymbol{W}$ which makes the formula maximum, so this two expressions equal.

2.5 Question 19

1. $-\log p(\boldsymbol{Y}|\boldsymbol{W}) = \mathcal{L}(W)$

$$= -\log \Bigg(\prod_{i=1}^{N} \mathcal{N}(y_i|\mu, WW^T+\sigma^2 I)\Bigg)$$

$$= -\sum_{i=1}^{N} \log \big(\mathcal{N}(y_i|\mu, WW^T+\sigma^2 I)\big)$$

$$= -\sum_{i=1}^{N} \log \frac{1}{(2\pi)^{\frac{D}{2}}|WW^T+\sigma^2 I|^{\frac{1}{2}}} + \log \Bigg(\exp \Big(-\frac{1}{2}(y_i - \mu)^T(WW^T+\sigma^2 I)^{-1}(y_i-\mu)\Big)\Bigg)$$

$$= \frac{DN}{2}\log \big((2\pi) + \frac{N}{2}\log|WW^T+\sigma^2 I| + \sum_{i=1}^{N} \frac{1}{2}(y_i - \mu)^T(WW^T+\sigma^2 I)^{-1}(y_i-\mu)$$

$$= \frac{N}{2}\Bigg(D\log 2\pi + \log|WW^T + \sigma^2 I| + \big((WW^T + \sigma^2 I)^{-1}(Y-\mu)(Y-\mu)^T\big)\Bigg)$$

2.we first compute the derivatives of the second item in bracket, we call it as A, then by the appendix,

$$\begin{aligned}\frac{\partial A}{\partial W} &= Tr\Big((WW^T+\sigma^2 I)^{-1}\frac{\partial(WW^T)}{\partial(W_{ij})}\Big)\\ &= Tr\Big((WW^T+\sigma^2 I)^{-1}(WJ_{ij}+J_{ji}W^T)\Big)\end{aligned}$$

Where J_{ij} is a matrix who has all zero entries except for $(J_{ij})_{ij} = 1$.
Then, we compute the derivatives of the third item in bracket, we call it as B,

$$\frac{\partial B}{\partial W} = Tr\Bigg(\frac{\partial\Big(Y(WW^T+\sigma^2 I)^{-1}Y^T\Big)}{\partial W_{ij}}\Bigg)$$

$$= Tr\Bigg((Y-\mu)(Y-\mu)^T\partial\Big((WW^T+\sigma^2 I)^{-1}\Big)$$

$$+ (WW^T+\sigma^2 I)^{-1}\Big(\partial((Y-\mu)^T(Y-\mu))\Big)\Bigg)$$

$$= Tr\Bigg((Y-\mu)(Y-\mu)^T\Big(-(WW^T+\sigma^2 I)^{-1}WJ_{ij}$$

$$+ J_{ji}W^T)(WW^T+\sigma^2 I)\Big)\Bigg)$$

The first item in bracket is a constant, so we just ignore it. Now, we combine these two,

$$\frac{\partial\mathcal{L}}{\partial W} = \frac{\partial A}{\partial W} + \frac{\partial B}{\partial W}$$

2.6 Question 20

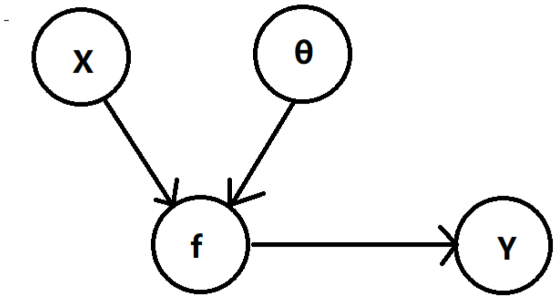


Figure 8: *Graphical model*

2.7 Question 21

Analysis: We use the matrix's pseudo-inverse to solve, then we got the first image.

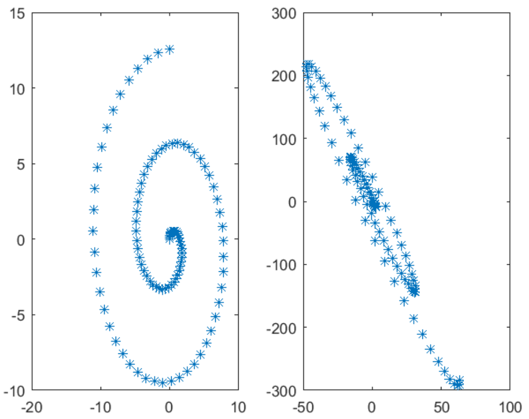


Figure 9: *Comparing the difference*

The rotated and shrunken shape seems to be an shear mapping image.
Then we use linear regression which get a perfect shape.

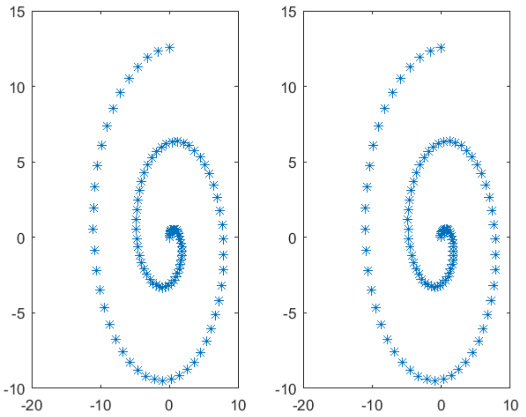


Figure 10: *Comparing the difference*

2.8 Question 22

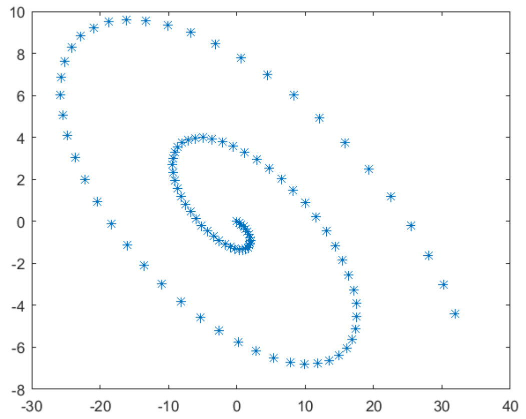


Figure 11: *Random mapping*

The shape is basically the same, but it rotated a little which means that the matrix calculation is a linear transformation, which would change the shape of the data sharply.

3 The Evidence

3.1 Models

3.1.1 Question 23

This assumption takes all the probability mass into the same, no matter which parameter we chose, the data has the equal likelihood. It is the simplest possible model because it does not take any parameters at all as an assumption-free model. But it is the hardest one, because it assigns many different types of behaviors similar probability.

3.1.2 Question 24

For the first model M_0, it gives the uniform model which has the same likelihood for all the data sets. It is more likely to fit those data sets are not well modeled by any sharp linear boundary.
For the second model M_1, it is the simplest logistic function to capture the decision boundaries with just one parameter which is able to fit the simplest boundaries.
For the third one, compared with M_3, it does not have the bias weight to fit some situations like data set has an unequal distribution.
The last one, M_3, is the most complex model but it could fit most linear boundaries. Also, it can transfer to the other models by setting the parameters into zero. But based on Bayesian Occam's razor rule, we should choose the simplest model by analysing our data sets rather than building the full and wide universally applicable models.

3.1.3 Question 25

We have chosen a spherical covariance matrix for the prior which implies that the data points are independent and identically distributed. The mean, $\mu = \mathbf{0}$, which simplifies the model, while the $\sigma^2 = 10^3$ which provides more the choices for the parameters in the data space. Large settings of the weights correspond to a sharp linear boundary in x space. The prior is very vague in parameter space, however, if the priors don't provide a wide space for parameters, the models are restrictive which would be more like the first model H_0.

3.1.4 Question 30

To sum up, in this coursework, we learned a lot. We thought machine learning should be super cool, giving tons of data, calculating automatically. But we are wrong! It should be called as Statistical Learning by using probability to describe what value the random variables value should be choose and by using likelihood to describe how likely the parameter should be this value. Assumption takes an important role in this process. The unsuitable assumption will cost a lot and gain a little. In general, this assignment is made of three parts: Prior, Posterior and Evidence. For each part, it starts at theoretical question, then some calculation and practical question. This path can deepen our understanding those knowledge and help us understand how observed data works for our model. We have learned related knowledge about first two parts in the lecture. However, we need to apply what we learn to solve part three question. This is what we should learn and machine learning scientists learn. Also, we are not really familiar with latex stuff. After finishing this coursework, the latex skills would be improved.