Assignment 1

Feiyang Liu Personal number: 950525-7213
liuf@kth.se

December 4, 2017

1 The prior

1.1 Theory

Question 1 *Why Gaussian form of the likelihood is a sensible choice? What does it mean that we have chosen a spherical covariance matrix for the likelihood?*

Ans: When there is not sufficient knowledge about prior distribution over given samples, Gaussian distribution is a sensible choice. Mainly having following two reasons:
1. According to central limit theorem, when independent random variable are added, their properly normalized sum tends toward a normal distribution;
2. Instead of assigning all possible distributions with same variance, normal distribution gives maximum uncertainty over those all unknown variables.
A spherical covariance matrix means the distribution has spherical (circular) symmetry. To be specific, off-diagonal elements in the matrix equal to zero while diagonal elements are non-zero. This indicates that the distribution of each variable is uncorrelated to others.

Question 2 *If we do **not** assume that the data points are independent how would the likelihood look then? Remember that* $T = [t_1, ..., t_N]$.

Ans: If we don't assume data points are independent, using the product rule, then the likelihood should be like:

$$p(T|f, X) = p(t_1|f, X)p(T_{2:N}|t_1, f, X) = ... = p(t_1|f, X)\prod_{i=2}^{N} p(t_i|t_1, ..., t_{i-1}, f, X)$$

1.1.1 Linear Regression

Question 3 *What is the specific form of the likelihood above, complete the right-hand side of the expression in (6).*

Ans: Assuming outputs are independent, and adding additional noise to the observation, we can get:

$$p(T|X, W) = \prod_{i=1}^{N} p(t_i|W, x_i) = \prod_{i=1}^{N} N(Wx_i, \sigma^2 I)$$

Question 4 *The prior in Eq.8 is a spherical Gaussian. This means that the preference" is encoded in terms of a L2 distance in the space of the parameters. With this view, how would the preference change if the preference was rather encoded using a L1 norm? Compare and discuss the different type of solutions these two priors would encode..*

Ans: Encoding in terms of a L2 distance returns a smoother prior distribution compared with L1. To be specific, let's taking two variable for example. In L2 distance situation, the prior distribution looks like Fig.2 in the following, a bunch of Concentric circles. However for L1 distance, point with same probability is located on a straight line with slope equals to 45 degree(and $-45, 135, -135$ degree), like concentric square. However, in one dimension case, these two cases are almost same.

Question 5 *Derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. However, in order to pass the assignment you only need to outline the calculation and highlight the important steps. You can make derivations for individual samples (x_i, t_i) and then generalize to the dataset or operate on matrices keeping the concept of vectorization in mind.*

- Briefly comment/discuss the form (mean and covariance).
- What is the effect of the constant Z, are we interested in this?

Ans: From what we learned on Exercise 1, we assume that $p(t_i|W, x_i) = N(x_i W, \delta^2 I)$, the brief calculation is:

$$p(W|X,T) = e^{-0.5W^T\Sigma_W^{-1}W} * e^{W^T\Sigma_W^{-1}u_W} * e^{-0.5u_W^T\Sigma_W^{-1}u_W}$$

$$p(t|X,W)p(W) = e^{-0.5\delta^{-2}(t-XW)^T(t-XW)} * e^{-0.5\tau^{-2}(W-W_0)^T(W-W_0)}$$

Expand the equation above, and compare the corresponding quadratic and linear terms, then get the mean and variance for the posterior as following:

$$u_W = (\tau^{-2} + \delta^{-2}X^TX)^{-1}(\delta^{-2}X^Tt + \tau^{-2}W_0)$$

$$\Sigma_W = (\tau^{-2} + \delta^{-2}X^TX)^{-1}$$

We don't care about the constant Z here, since we just need to compare the quadratic and linear terms bewteen both sides of *Eq.8*. Then constant term does not influence the mean and variance, what actually matters.

1.1.2 Non-parameric Regression

Question 6 *Explain what this prior does? Why is it a sensible choice? Use images to show your reasoning. Clue: use the marginal distribution to explain the prior.*

Ans: This prior represents the relation bewteen input data. For similar points x_n and x_m, the corresponding values of $y(x_n)$ and $y(x_m)$ then have a stronger correlation. This correlation can be controlled by the hyperparameter, θ. By using this prior, we can directly formulate the marginal distribution $p(t)$ by intergrate over y.
The reason why choosing Gaussian as the format of prior distribution, is that the multiplication of two Gaussians is also a Gaussian. Moreover, if these two Gaussian are independent, then variance of the output is simply add. Therefore, C in the equation below, equals to the sum of $p(t|y)$ and $p(y)$ variances.

$$\int p(t) = \int p(t|y)p(y)dy = N(t|0, C)$$

Question 7 *Formulate the joint likelihood of the full model that you have defined above,*

$$p(T; X; f;)$$

(Try to draw a very simple graphical model to clearly show the assumptions that you have made.)

Ans: According to the Bayes' theorem, the joint probability can be expanded as following. In this case, x and θ is independent. T is conditional independent of x and θ with f given.

$$p(T, X, f, \theta) = p(T|f)p(f|x, \theta)p(x)p(\theta)$$

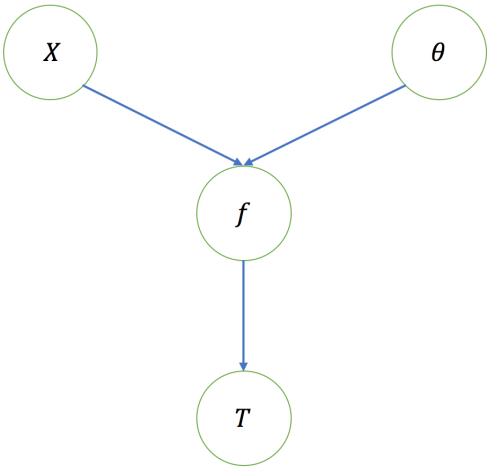

Figure 1: Graphic model

Question 8 *Complete the marginalisation formula in Eq.12 (general form)*

- Explain how this connects the prior and the data?
- How does the uncertainty "filter" through this?
- What does it imply that θ is left on the left-hand side of the expression after marginalisation?

Ans: The Completed formula:

$$p(T|X, \theta) = \int p(T|f)p(f|X, \theta)dy$$

On the right side, the term $p(f|X, \theta)$ is the prior, $p(T|f) = \prod p(t_i|f(x_i))$ is the likelihood over data. Both two distribution on the right-hand side are Gaussian, hence the "uncertainty" is simply added. Intuitively, $variance = k(X, X) + \epsilon^2\delta$. The left θ implies that θ is the model parameter in the integral, therefore it is kept.

1.2 Practical

1.2.1 Linear Regression

Question 9 *Describe the plots and the behavior when adding more data? Is this a desirable behavior? Provide an intuitive explanation.*

Ans: Fig.1 shows the prior distribution over $[w_0, w_1]$. Since we don't have any information on parameters, then set it as normal distribution, center at origin.

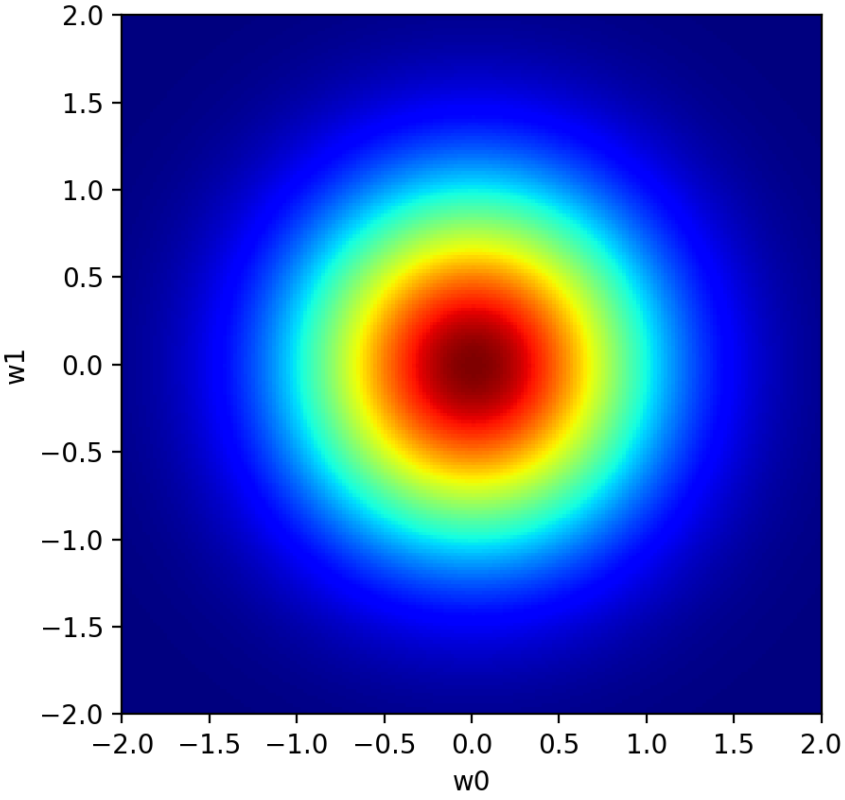

Figure 2: Prior Distribution over W

From Fig.2, it is obvious that parameter distribution over W becomes constraint after one single point observation. In Fig.3, we sampled several points from the posterior distribution and draw the functionality. Because we still have large uncertainty on the distribution, the obtained slopes and intercepts varies from each other.
However, the posterior zone becomes narrower with more observed data coming in. From Fig.4 and Fig.6, we have stronger belief on the distribution over W. Especially in Fig.6, it can almost be inferred from the plot that $[w_0, w_1] = [1.5, 0.8]$, which is exactly the same with the parameter setting. In Fig.7, the lines overlap in a small region, indicating that they have similar slopes and intercepts, which are the w_0 and w_1 respectively.

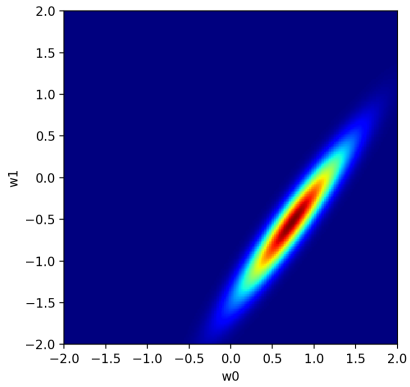

Figure 3: Posterior after 1 observation

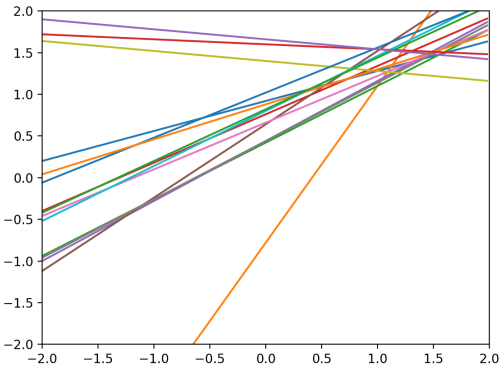

Figure 4: Samples from posterior

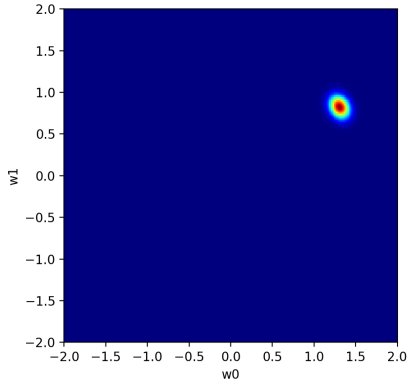

Figure 5: Posterior after 5 observations

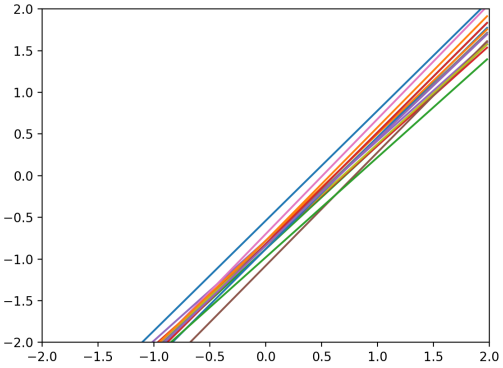

Figure 6: Samples from posterior

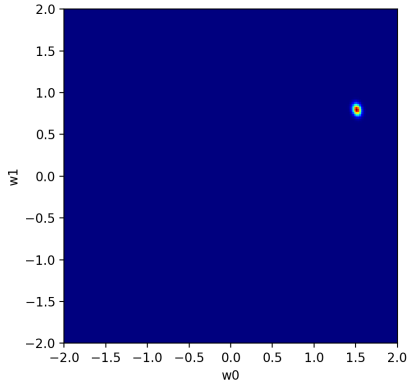

Figure 7: Posterior after 20 observations

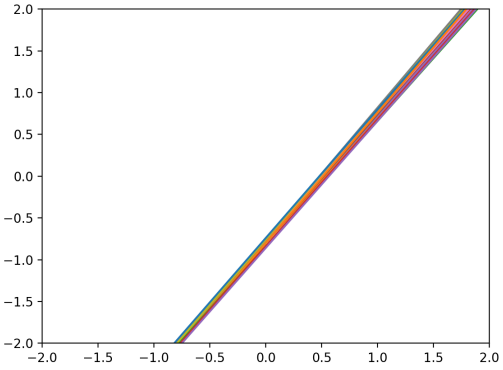

Figure 8: Samples from posterior

1.2.2 Non-parametric Regression

Question 10 *Explain the behavior of altering the length-scale of the covariance function.*

Ans: In Fig.8, the samples from prior distribution are visualized. It is apparent that the higher length-scale, the smoother the sampled lines. Wanna deeper understand the reason, we need to look at the the squared exponential covariance function.

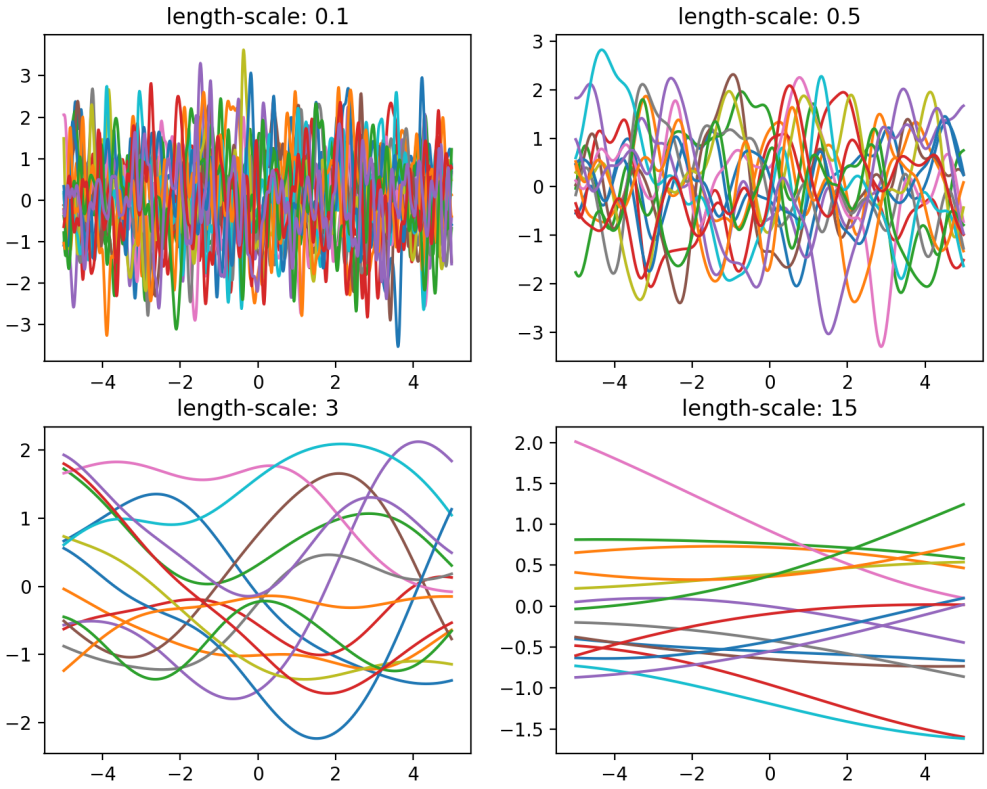

Figure 9: Samples from Prior Distribution of different length scale

$$k(x_i, x_j) = \sigma_f^2 e^{-(x_i - x_j)^T (x_i - x_j) l^{-2}}$$

In the definition of this prior, the kernel is used as covariance function. When length scale is larger, then the output from kernel, covariance is larger. with larger covariance, variables are more tightly correlated, vice versa. Taking the last plot of Fig.8 for example, length-scale is 15, hence samples are strongly related so the lines are smoother (samples are more similar).

Question 11 *Explain the behavior of the samples and compare the samples of the posterior with the ones from the prior. Is this behavior desirable? What would happen if you would add a diagonal covariance matrix to the squared exponential?*

Ans: There are 4 plots in the following. From Fig.9 to Fig.11, they are pure (without adding noises) predictive posterior distribution with length scale 2.8, 3.0, 3.2 respectively. In those three figures, the red line is the cos curve, the other lines are the samples from posterior distribution.
Comparing Fig.9 to Fig.11 with the prior, the posterior is strongly constraint because of those seven observed data. Moreover, the uncertainty between $[0, 2\pi]$ is much smaller than the outside regions (samples are totally overlapped in the observation region, getting diverge in the outside zone).
As for noise, comparing Fig.10 and Fig.12. Since noise is added to the the squared exponential, the obtained blue curve can not exactly pass through observed data. Diagonal elements are added with noise, it means that we do not have the high certainty on those observed data anymore. Here only the mean curve is plotted, because the sampled curve is a mess with noise added.

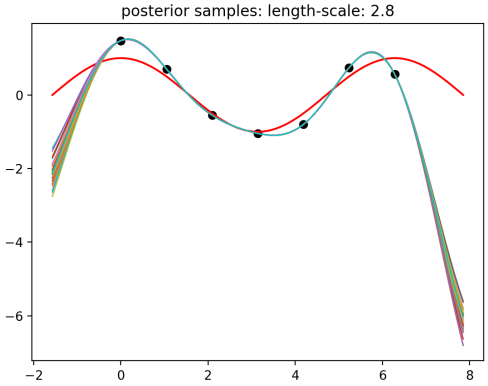

Figure 10: **pure** predictive posterior distribution with length scale: 2.8

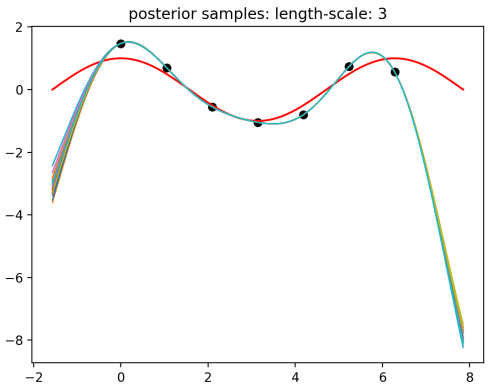

Figure 11: **pure** predictive posterior distribution with length scale: 3

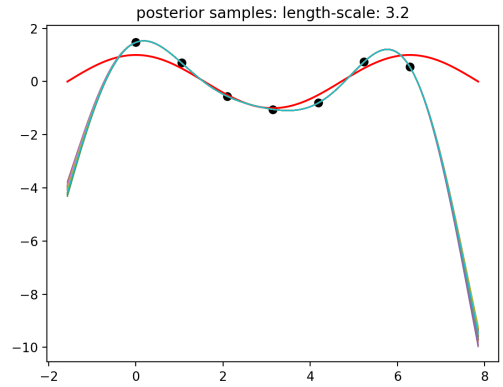

Figure 12: **pure** predictive posterior distribution with length scale: 3.2

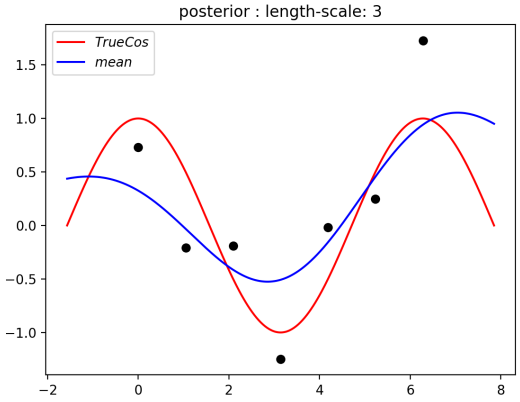

Figure 13: **noise** predictive posterior distribution with length scale: 3

2 The Posterior $p(X|Y)$

1.3 Theory

Question 12 *What type of preference does this prior encode?*

Ans: This prior indicates the probability is normally distributed in the space, variables are uncorrelated in each dimension. Hence, this prior does not have specified preference over X. However, we can determine the preference by tuning the parameter of the prior, this preference can be passed through by marginalization.

Question 13 *Perform the marginalisation in Eq.23 and write down the expression. As previously, it is recommended that you do this by hand even though you only need to outline the calculations and show the approach that you would take to pass the assignment.*

Ans: The *Eq.*23 can be expanded by Integrating over X

$$p(Y|W) = \int p(Y|X,W)p(X)dX$$

Let's take a single point for example, the latent variable is from normal distribution:

$$p(x_i) = N(0, I)$$

The conditional probability follows by:

$$p(y_i|x_i, W) = N(Wx_i + \mu, \Psi)$$

where $\Psi = \sigma^2 I$ is usually diagonal, assuming same noise is added to each variable. Therefore, since these two distributions are Gaussian, we can easily obtain:

$$p(y_i|W) = \int p(y_i|x_i, W)p(x_i)dX = N(\mu, \Psi + WW^T)$$

Then it is quite intuitive to compute the probability of the whole sequences:

$$p(Y|W) = \prod p(y_i|W)$$

1.3.1 Learning

Question 14 *Compare these three estimation procedures above in log-space.*

- How are they different?
- How are MAP and ML different when we observe more data?
- Why are the two last expressions of $Eq.25$ equal?

Ans:

- Compared with ML, MAP introduce the prior over M. Hence, ML can be considered as a special case of MAP where prior over M is uniform. Type-II ML is to maximize the marginal distribution, not related to latent variables X.
- Since the difference between ML and MAP is the prior term. When there are more data coming in, in MAP, term $p(Y|X, W)$ varies while $p(W)$ is constant. In this case, it is the $p(Y|X, W)$ that makes the decision of maximization, which is quite similar with ML algorithm.
- It is quite intuitive that for all W, the denominator of the following term is a constant. Hence, We want to find the W which maximize this term, it is equivalent to find the W maximizing the numerator. Consequently, the two last expressions of $Eq.25$ equal.

$$\frac{p(Y|X, W)p(W)}{\int p(Y|X, W)p(W)dW}$$

Question 15 *1. Write down the objective function $-log(p(Y|W)) = L(W)$ for the marginal distribution in Eq.23*
2. Write down the gradients of the objective with respect to the parameters $\frac{\sigma L}{\sigma W}$

Ans:

$$-\log P(Y|W) = -\log \prod p(y_i|W) = -\sum \log P(y_i|W)$$

Here, $\log P(y_i|W)$ is a Gaussian distribution, from above we have derivate its mean and variance already, hence:

$$\begin{aligned}
\log P(y_i|W) &= \log \big[\frac{1}{(2\pi)^{\frac{D}{2}}(\sigma^2 I + WW^T)^{\frac{1}{2}}} e^{-\frac{1}{2}(y_i-\mu)^T(\sigma^2 I+WW^T)^{-1}(y_i-\mu)}\big] \\
&= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2 I + WW^T) - \frac{1}{2}(y_i-\mu)^T(\sigma^2 I + WW^T)^{-1}(y_i-\mu) \\
&= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2 I + WW^T) - \frac{1}{2}\big[y_i^T(\sigma^2 I + WW^T)^{-1}y_i - 2\mu^T(\sigma^2 I + WW^T)^{-1}y_i \\
&+ \mu^T(\sigma^2 I + WW^T)^{-1}\mu\big]
\end{aligned}$$

Then we sum the sequences:

$$\begin{aligned}
-\log P(Y|W) &= -\sum \log \big[\frac{1}{(2\pi)^{\frac{D}{2}}(|\sigma^2 I + WW^T|)^{\frac{1}{2}}} e^{-\frac{1}{2}(y_i-\mu)^T(\sigma^2 I+WW^T)^{-1}(y_i-\mu)}\big] \\
&= -\sum \big\{ -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\sigma^2 I + WW^T|) - \frac{1}{2}(y_i-\mu)^T(\sigma^2 I + WW^T)^{-1}(y_i-\mu)\big\} \\
&= \frac{ND}{2}\log(2\pi) + \frac{N}{2}\log(|\sigma^2 I + WW^T|) + \frac{1}{2}\sum (y_i-\mu)^T(\sigma^2 I + WW^T)^{-1}(y_i-\mu) \\
&= \frac{ND}{2}\log(2\pi) + \frac{N}{2}\log(|\sigma^2 I + WW^T|) + \frac{N}{2}Tr(C^{-1}S) \\
&= \log L(W)
\end{aligned}$$

where S is the data covariance matrix, $S = \frac{1}{2}\sum (y_i-\mu)^T(y_i-\mu)$, C is the covariance matrix of predictive distribution, $C = (\sigma^2 I + WW^T)$
Easy to know, the first term is constant over different W. Before we take the gradient, we first do some preparatory tasks:

$$C^{-1} = \sigma^{-1}I - \sigma^{-2}WM^{-1}W^T$$

$$M = \sigma^2 I + W^T W$$

$$\frac{\delta C(W)}{\delta W} = \frac{\delta WW^T}{\delta W} = \Delta(W)$$

Thus, we have less computation cost by inversing M instead of C. Finally, take the gradient:

$$\frac{\delta L(W)}{\delta W} = \frac{N}{2}Tr(C^{-1}\Delta(W)) + \frac{N}{2}Tr(-C^{-1}\Delta(W)C^{-1}S)$$

1.4 Practical

1.4.1 Linear Representation Learning

Question 16 *Plot the representation that you have learned (hint: plot X as a two-dimensional representation). Explain the outcome and discuss key features, elaborate on any invariance you observe. Did you expect this result?*

Ans: Fig.14 to Fig.17 are actual X locations and recovered X with different σ settings as two-dimensional representation. Here we don't add noise to the observed data Y. From the following plots, there are almost no variation between different σ. However, after we add noise Y, we get two plots, Fig.18 and Fig.19. In this case, it seems impossible for us to recover the latent data when noise is large.
As for invariance, I find out that W is invariant to rotations. When we look back to Question.15, if we replace W with $W' = WR$, where is the orthogonal rotation matrix. Then it is obvious that $W'W'^T = WRR^TW^T = WW^T$, making no difference to the optimization result. On the text book $P575$, this problem is also well explained.

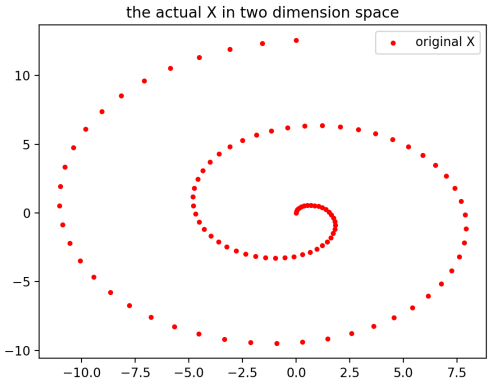

Figure 14: Actual X locations in 2-D

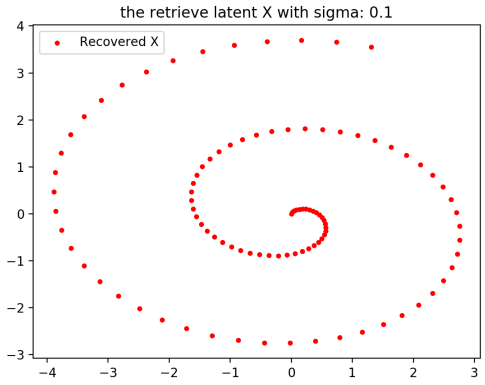

Figure 15: Recovered X in 2-D, $\sigma = 0.1$

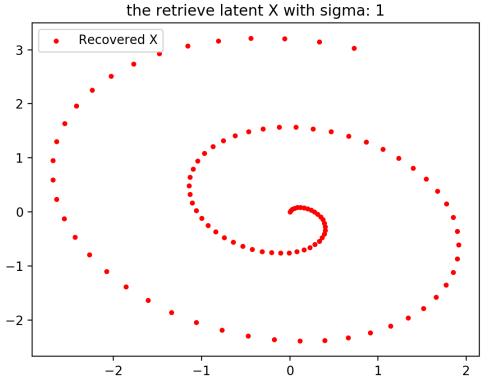

Figure 16: Recovered X in 2-D, $\sigma = 1$

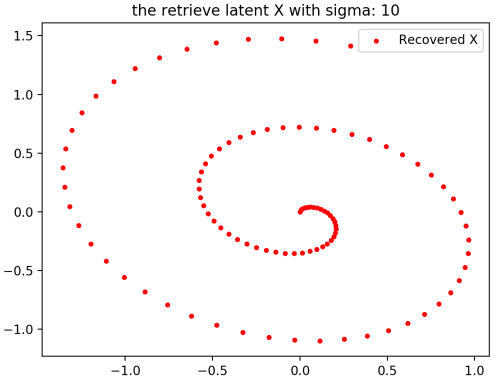

Figure 17: Recovered X in 2-D, $\sigma = 10$

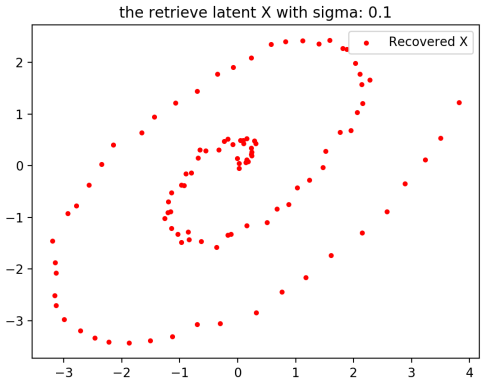

Figure 18: Recovered noisy X, $\sigma = 1$

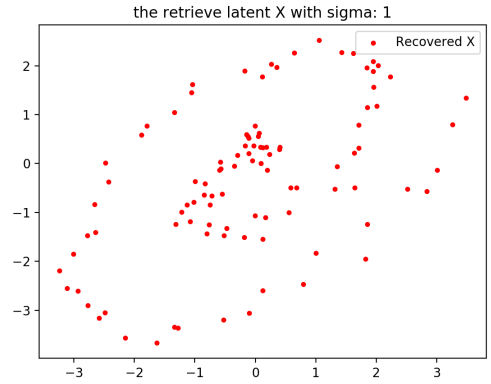

Figure 19: Recovered noisy X, $\sigma = 10$

3 The Evidence $p(D)$

1.5 Theory

Question 17 *Why is this the simplest model, and what does it actually imply? Discuss its implications, why is this a bad model and why is it a good model?*

Ans: Obviously, it is the simplest model since it assigns all possible data with same probability. It is also a bad model because it indicates there is no useful information over the data space. By allocating possibilities over a too wide range, this model leads to large computational cost. Hence, it usually loses when competing with other models, shown in the literature of Murray 2005.

Question 18 *Explain how each separate model works. In what way is this model more or less flexible compared to M_0? How does this model spread its probability mass over D?*

Ans: Let's take a look at this model, instead of using uniform probability, M_1 introduces x_1 to compute the possibility. Hence, if the decision boundary is strongly related to x_1, M_1 can perfectly deal with it. To be specific, we quote a picture from the literature. For data set(b) in Fig.20, M_1 model can capture the vertical decision boundary, using x_1 to classify cross point from circle point.

$$p(D|M_1, \theta_1) = \sum_{n=1}^{9} \frac{1}{1 + e^{-t^n \theta_1^1 x_1^n}}$$

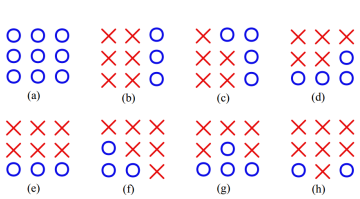

Figure 20: Labellings for the grid of inputs corresponding to data sets,from Murray 2005[1]

Question 19 *How have the choices we made above restricted the distribution of the model? What datasets are each model suited to model? What does this actually imply in terms of uncertainty? In what way are the different models more flexible and in what way are they more restrictive? Discuss and compare the models to each other?*

Ans: M_2 introduces x_2 and M_3 has an additional offset term. Here M_2 can deal with rotated decision boundary which M_2 is not able to, like data set(d) in Fig.20. The full logistic regression model, M_3, is a more general but also complicated model. General means that it can realize M_1 and M_2 by setting some parameters as zero. For some extreme data, like set(a) in the literature, model M_3 can cope with it by adding the offset. Besides, for some sharp linear boundaries like set(f), this model can give a better performance than other three models. M_3 is more flexible since it can spread its importance to regions by tuning the parameters, while this flexibility comes at the expense of sometimes losing out to a simpler model. For some "simple" data sets, M_1 and M_2 are more intuitive and typical, but this also leads to restriction over the whole data sets.

$$p(D|M_2, \theta_2) = \sum_{n=1}^{9} \frac{1}{1 + e^{-t^n(\theta_2^1 x_1^n + \theta_2^2 x_2^n)}}$$

$$p(D|M_3, \theta_3) = \sum_{n=1}^{9} \frac{1}{1 + e^{-t^n(\theta_3^1 x_1^n + \theta_3^2 x_2^n + \theta_3^3)}}$$

Question 20 *Explain the process of marginalisation and briefly discuss its implications.*

Ans: This marginalization process integrates $p(D|M, \theta)$ over all possible model parameter θ, implies where this model places its probability mass. In some way, the evidence of this model is equivalent to the weighted sum of $p(D|M, \theta)$, corresponding weight comes from the prior distribution.

Question 21 *What does this choice of prior imply? How does the choice of the parameters of the prior μ and Σ effect the model?*

Ans: This prior almost implies we have no idea about the parameters.

$$\mu = 0$$

$$\Sigma = \sigma^2 I$$

$$\sigma^2 = 1000$$

This mean implies the prior is centering at the origin. This variance implies two aspects. First, variables in each dimension are uncorrelated; Second, we have large uncertainty on this prior. Take a two-dimensional distribution for example, shown in Fig.21. The probability is located in a wide range over the 2-d space, hence we have more choices for different model parameter settings. This will make capable of having sharp decision boundary over more kinds of data sets.

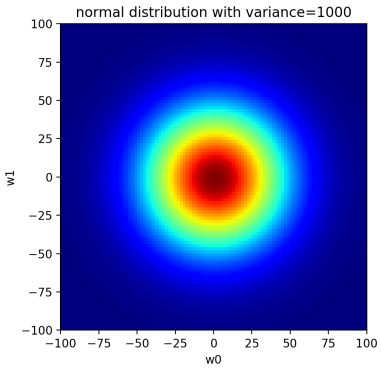

Figure 21: Gaussian distribution with mean=0, variance=1000

1.6 Practical

Question 22 *Plot the evidence over the whole dataset for each model (and sum the evidence for the whole of D, explain the numbers you get). The* ***x-axis*** *index the different instances in D and each models evidence is on the* ***y-axis****. How do you interpret this? Relate this to the parametrisation of each model.*

Ans: The sum of evidence for the whole of dataset is $evidence = [1, 1.00226612, 1.00645773, 0.88852154]$. For M_0, because the distribution over each data set is uniform, hence the sum is 1. As for the other three models, we use using a naive Monte Carlo approach to approximate, which is lack of accuracy. The second and third sum is larger than one, this is because when sampling from prior, some giving high probability to $p(D|M_i, \theta^s)$ might be sampled several times. Hence the outcome evidence is higher than the actual value. For M_4, the sum is smaller than 1 is because its prior has higher dimension than others, hence it have lower percentage of samples when we having same number of samples for all 4 models. However, if we keep increasing the number of samples (here I have 10000 samples, much less than the literature), the sum of its evidence can also be larger than one.

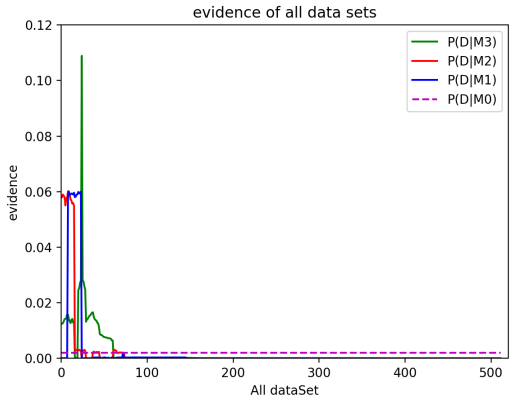

Figure 22: Evidence of all data sets

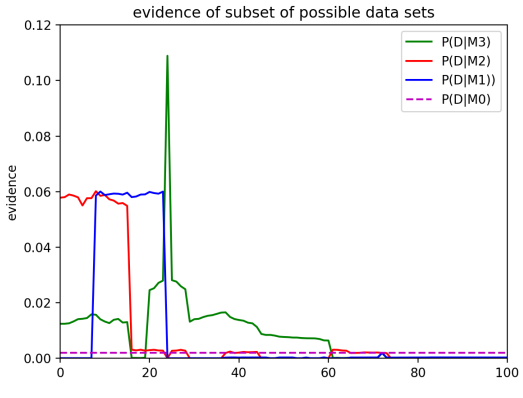

Figure 23: Evidence of subsets of possible data sets

Fig.22 and Fig.23 shows the evidence probability over each model. From the plot, it obvious that M_0 has the largest evidence range over th whole data set. Except M_0, the other three models are located closely. This is within our expectation, because M_2 is the same as M_1 when θ_2 is set to be zero. For some data set, M_3 has a extreme high evidence probability, this case might be resulted from several reasons. One of reasons is that this data set has large offset, like all positive or negative. And M_3 is the only model among four has the bias term, θ^3, which can compensate for this offset.

Question 23 *Find using* ***np.argmax*** *and* ***np.argmin*** *which part of the D that is given most and least probability mass by each model. Plot the data-sets which are given the highest and lowest evidence for each model. Discuss these results, does it make sense?*

Ans: Here I just directly screenshot from the terminal window, where cross represents one while circle means negative one, then obtain Fig.24 to Fig.26:
M_1. The decision boundary of most probability mass is quite like a vertical line, which is strongly related to x_1. As for the least probability mass, obviously, these nine data points can not be linearly separated.
M_2. The decision boundary of most probability mass goes through the origin, no offset needed. Hence, M_2 can perfectly handle it. As for the least probability mass, the reason is the same with M_1, the cross are located in both sides of circle points.
M_3. The decision boundary of most probability mass goes away from the origin, where bias term is required. Hence, M_3 has the most probability mass of this dataset rather than the other two. As for the least probability mass, the circles are located in both sides of cross points, again not linearly separated.
Consequently, all the results shown below are reasonable.

```
most probability mass of model 1
oxx
oxx
oox

least probability mass of model 1
ooo
oxx
xxo
```

Figure 24: M_1

```
most probability mass of model 2
oxx
oox
oox

least probability mass of model 2
xxo
ooo
oxx
```

Figure 25: M_2

```
most probability mass of model 3
ooo
ooo
ooo

least probability mass of model 3
xxo
oxo
oxx
```

Figure 26: M_3

Question 24 *What is the effect of the prior $p(\theta)$?*

- What happens if we change its parameters?
- What happens if we use a non-diagonal covariance matrix for the prior?
- Alter the prior to have a non-zero mean, such that $\mu = [5; 5]^T$?
- Redo evidence plot for these and explain the changes compared to using zero-mean.

Ans: In Fig.27 and Fig.28, $\mu = [5; 5]^T$. In Fig.29 and Fig.30, the covariance matrix is randomly generated from normal distribution times σ. μ. From the plot, there are not too much difference with zero mean, but the peaks of three models have somehow changed.

Taking first model for example, now M_1 has higher evidence in first 15 datasets instead of 10 to 25. From my perspective, because now θ_1 has large probability being positive, this parameter setting is more capable to draw decision boundaries to first 15 data sets. Σ. Now the variance matrix is non-diagonal, hence parameters are correlated among each dimension. Since the way I generate data set is convert number into binary code, hence the lower right of $3 * 3$ matrix in first is constant negative one, only variations in the upper left corner(top of the off-diagonal line). So when parameter is strongly dependent on each other, M_2 has a relatively higher evidence than previous cases. There are not obvious change in M_1 because only one parameter is used in this model.

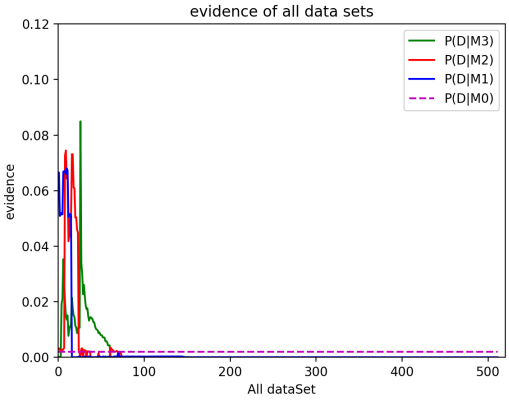

Figure 27: Evidence with $\mu = [5; 5]^T$

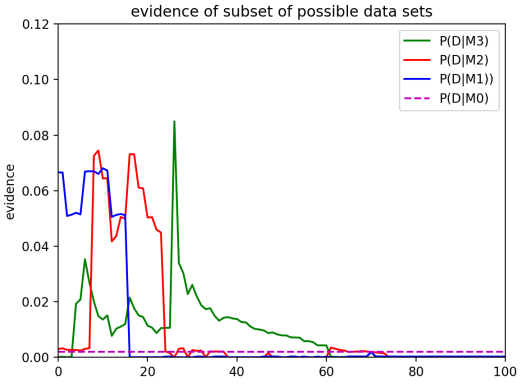

Figure 28: Evidence of subsets

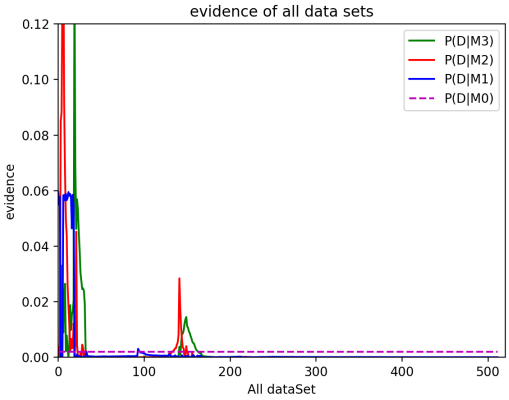

Figure 29: Evidence with no-diagonal Σ

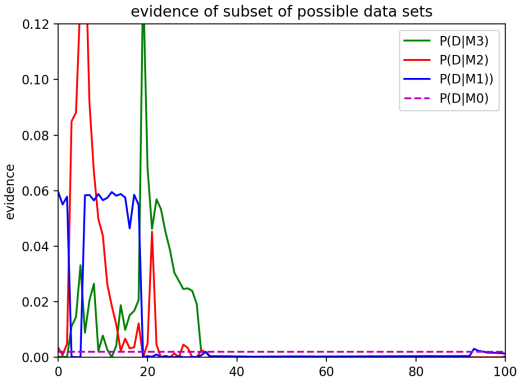

Figure 30: Evidence of subsets

References

[1] I. Murray and Z. Ghahramani. (2005) A note on evidence and Bayesian Occams razor. Technical report.