DD2434 Machine Learning, Advanced Course
Assignment 1

Qiao Ren

Kth mail: qiaor@kth.se

Dec 2019

Question1. Explain why Gaussian form of the likelihood is a sensible choice and what assumptions we make by this choice. What assumptions do we make about the data by choosing a spherical covariance matrix for the likelihood?

Answer:
Gaussian form of the likelihood is a sensible choice. Because: the central limit theorem tells us: for large number of data points, the distribution tends to be a gaussian distribution. Likelihood is defined as: the joint probability of sampled data, given a set of model parameter values. Likelihood is written as: p(D|w). Therefore, gaussian distribution is a highly probable type of distribution, when we estimate how the data are distributed, with given model parameters. After we make this assumption, the things that we do not know is: what is the value of mean and what is the value of variance or covariance.

The assumptions that we make about the data by choosing a spherical covariance matrix for the likelihood: we assume that the dimensions in data are independent to each other. We use a spherical covariance matrix to express this relation. Knowing the information about one variable does not help us getting the information about another variable. Spherical covariance matrix C is propotional to identity matrix I. All the off-diagonal values are zero, as shown in the following equation.

$$C=\lambda I$$

Question2. If we do not assume that the data points are independent how would the likelihood look then? Remember that $T = [t_1,...,t_N]$

Answer:

The likelihood becomes the following equation. Because each data point t_i depends on all the observations in the previous time, which are $t_1,...,t_{i-1}$.

$$p(T \mid f, X) = \prod_{i=1}^{N} \mathrm{p}(\mathrm{t_i} \mid \mathrm{t_1},...,\mathrm{t_{i-1}}, \mathrm{f}, \mathrm{X})$$

Question 3. Please, complete the right-hand side of the expression in equation (6).
Answer:

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

Because t_i is computed by a linear equation Wx_i plus the noise. Noise follows gaussian distribution N(0, σ^2I)

Question 4. The prior over each row of W in Eq.8 is a spherical Gaussian: p(w) = N(w0; τ²I).This means that the preferred model is encoded in terms of L2 distance in the parameter space.

• What would be the effect of encoding the preferred model with L1 norm (for model parameters)?

• Discuss how these two types of priors affect the posterior from the regularization perspective. Write down the penalization term, i.e. the negative log-prior, and illustrate for a two-dimensional problem (in the two-dimensional parameter space)

Answer:
The goal or regularization is to avoid our model being too complicated. We want to avoid overfitting. L1 norm is also called Lasso regression. The penalty is expressed as the sum of all the absolute values of weights: $\sum |w_j|$. In terms of equation, the goal is to minimize:

$$ErrorFunction + L1PenaltyTerm$$
$$= \frac{1}{2}\sum_{i=1}^{n}\{t_n - w^T\psi(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|$$

The effect is: the optimal weight w^* is the one which is the closest to the minimum value of the error function, with the constraint: $\sum |w_j| \leq$ a constant. Please see Image 1. λ controls the strength of the L1 penalty norm. λ controls the amount of penalty. When λ=0, no parameters in the model are eliminated. When λ increases, more and more parameters in the model are set to zero. This means that those parameters are eliminated. A complicated model is simplified to the simple model.

L2 norm is: $\sum |w_j|^2$. When L2 norm ≤ a constant, the optimal weight is shown in image 2. The blue ellipses are the 3-dimensional error function projected to a 2D plane. Each ellipse represents a contour. The goal is to minimize:

$$ErrorFunction + L2PenaltyTerm$$
$$= \frac{1}{2}\sum_{i=1}^{n}\{t_n - w^T\psi(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^2$$

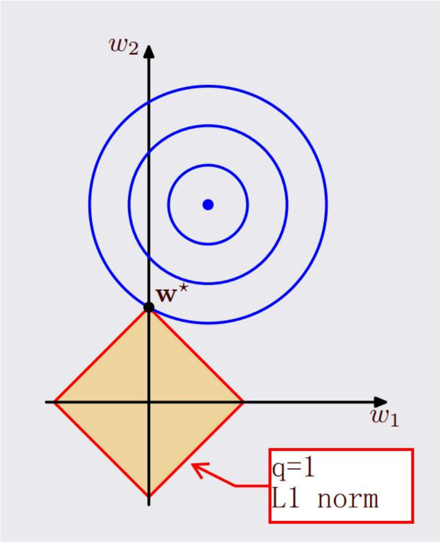


Image 1 L1 norm

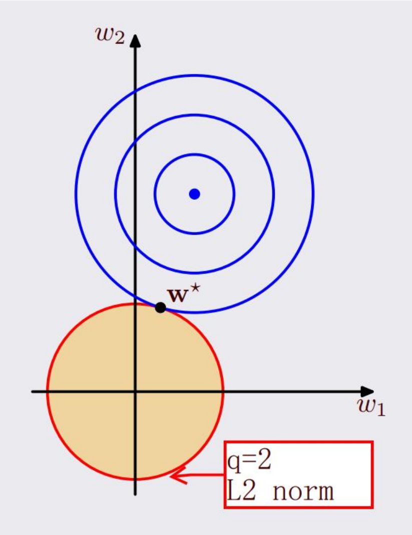


Image 2 L2 norm

Q5 Assuming conditional independence of the target variables in t, derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. To pass the assignment you only need to outline the calculation and highlight the important steps. In summary, please complete the following tasks

• Derive the posterior over the parameters and explain the final form in terms of the mean and covariance.

• How does the posterior form relate to the least square estimator of W (equivalent to the maximum likelihood approach) for this linear regression problem?

• How does the constant Z (Eq.7) affect the solution? Are we interested in it?

Answer:

• Derive the posterior over the parameters and explain the final form in terms of the mean and covariance.

Based on: $p(W) = MN(W_0, I, \tau^2 I)$, we know that each row in the matrix W follows the normal distribution with diagonal covariances. It means that each element w_{ij} in the matrix W are independent to each other. So we can convert W from a matrix to a vector, which means:

Matrix $W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ w_{m1} & \cdots & \cdots & w_{mn} \end{bmatrix}$ is converted to

Vector $W = \begin{bmatrix} w_{11} \\ w_{12} \\ \cdots \\ w_{mn} \end{bmatrix}$

We derive the posterior p(W|X,T) in two different ways.

The first equation of posterior is:

$$\begin{aligned}
&p(W \mid X,T) \\
&= \frac{1}{Z} * likelihood * prior \\
&= \frac{1}{Z} p(T \mid X,W) p(w) \\
&= \frac{1}{Z} * N(XW, \sigma^2 I) * N(0,\Sigma) \\
&= \frac{1}{Z}[-\frac{1}{2\sigma^2}(T-XW)^T(T-XW)-(-\frac{1}{2}W^T\Sigma^{-1}W)] \\
&= \frac{1}{Z}\{[-\frac{1}{2\sigma^2}T^T T]+[\frac{1}{\sigma^2}T^T XW]+[-\frac{1}{2\sigma^2}(XW)^T XW-\frac{1}{2}W^T\Sigma^{-1}W]\} \\
&= const * \{const + linear + quadratic\} \\
&= equation(1)
\end{aligned}$$

The second equation of posterior is:

$$\begin{aligned}
&p(W \mid X,T) \\
&= N(W \mid \mu_W, \Sigma_W) \\
&= \frac{1}{(2\sqrt{1})^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(W-\mu_W)^T\Sigma_W^{-1}(W-\mu_W)} \\
&= c * e^{-\frac{1}{2}W^T\Sigma_W^{-1}W} * e^{W^T\Sigma_W^{-1}W} * e^{-\frac{1}{2}\mu_W^T\Sigma_W^{-1}\mu_W} \\
&= c * e^{quadratic} * e^{linear} * e^{const} \\
&= equation(2)
\end{aligned}$$

quadratic term in equation(1) = quadratic term in equation(2)

$$-\frac{1}{2}W^T\Sigma_W^{-1}W = -\frac{1}{2}W^T(\frac{1}{\sigma^2}X^T X+\Sigma^{-1})W$$

$$\Rightarrow \Sigma_W = \frac{1}{\frac{1}{\sigma^2}X^T X+\Sigma^{-1}}$$

So we get the covariance matrix Σ_W.

linear term in equation(1) = linear term in equation(2)

$$W^T\Sigma_W^{-1}\mu_W = \frac{1}{\sigma^2}T^T XW$$

$$\Rightarrow \mu_W = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}X^T X+\Sigma^{-1})^{-1}X^T T$$

So we get the mean μ_W.

• How does the posterior form relate to the least square estimator of W (equivalent to the maximum likelihood approach) for this linear regression problem?

Maximum likelihood is equivalent to minimize the error function with the addition of a regularization term.

Least square estimator is to minimize: $\frac{1}{2}\sum_{n=1}^{N}\{t_n - W^T\phi(x_n)\}^2 + \mathrm{Re}\,gularzationTerm$

Maximum posterior can be written in the following way:

Maximize posterior
→Maximize (likelihood*prior)
→Maximize log (likelihood*prior)
→Maximize log (likelihood) + maximize log(prior)
→Maximize $-\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - W^T\phi(x_n)\}^2$ + maximize log ($ce^{\frac{1}{2}(w-\mu)^T\frac{1}{\alpha}(w-\mu)}$)
→Minimize: $\frac{1}{2}\sum_{n=1}^{N}\{t_n - W^T\phi(x_n)\}^2 + \mathrm{Re}\,gularzationTerm + const$

Therefore, maximum posterior is equivalent to least square estimator.

• How does the constant Z (Eq.7) affect the solution? Are we interested in it?

Constant Z is a normalization constant. By using Z, the posterior represents a true probability distribution. P(W|X,T)< or = 1. Z can be seen as the evidence. Constant Z does not influence mean and variance.

Question 6 Explain what this prior does. Motivate the choice of this prior and use images to show your reasoning. Clue: use the marginal distribution to explain the prior.

Answer:
The mean in the prior is zero. Because we assume that the data have been translated to zero-mean. So the mean is not a function, but zero.

The covariance in the prior is a Gram matrix. The Gram matrix is determined by a kernel function. The kernel function describes the similarity between two points x and x’. If two points x_n and x_m are similar, then their corresponding values $y(x_n)$ and $y(x_m)$ are strongly correlated, compared with other dissimilar points.

The marginal distribution of **t** is:

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

$$t_i = f(x_i) + \varepsilon = f_i + \varepsilon$$

The reason that we choose Gaussion distribution as the prior distribution is: p(t|y) follows gaussian distribution and p(y) also follows gaussian distribution. The multiplication of two gaussian distribution is also a gaussian distribution. More than that, the two Gaussian sources of randomness are independent and so their covariances simply add. The two Gaussian sources of randomness are the randomness associated with $f(x_i)$ and the randomness associated with ε .

Image3 below gives an example for the prior distribution. The red straight line represents that mean= 0. The curved functions (in blue, purple, cyan, red and green) are the samples taken from the prior distribution.

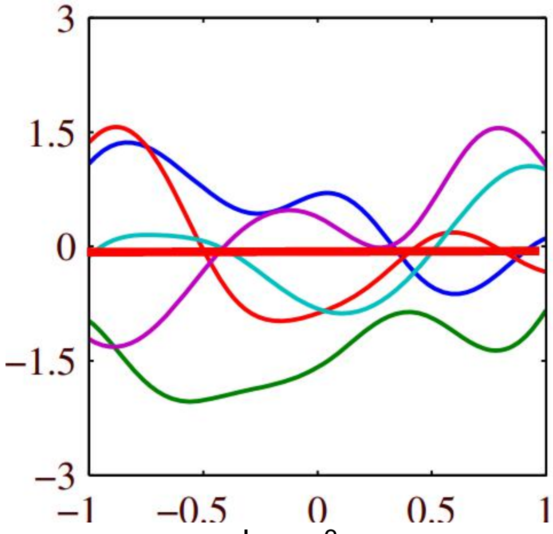


Image 3

Question 7 Formulate the joint likelihood of the full model defined above, p(T; X; f; θ) and draw a simple graphical model reflecting the assumptions that you have made.

Answer:

$$p(T, X, f, \theta) = p(T \mid f) p(f \mid X, \theta) p(X) p(\theta)$$

T depends on f. f depends on data X and parameters θ. X and θ are independent. Image 4 represents this relation.

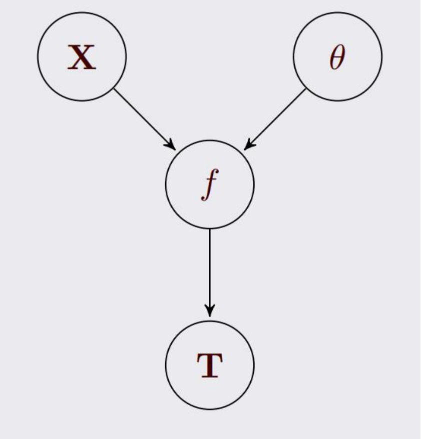


Image 4

*Question 8 Complete the marginalisation formula in Eq.12 p(**T**|**X**, **θ**) =?. and discuss the following:*

- *Explain how it connects the prior and the data.*
- *How does the uncertainty filter through the marginalisation?*
- *Why do we still condition on θ after the marginalisation?*

Answer:

$$p(T \mid X,\theta) = \int p(T \mid f)p(f \mid X,\theta)df$$

p(f|X, θ) is the prior. p(T | f) is the likelihood, in which $p(T \mid f) = \prod p(t_i \mid f(x_i))$. Because we are interested in the relation between X and θ. We are not interested in f. So we marginalize out f. The uncertainty has been added into each target value $t_{i.}$ Θ is the parameters in the model. After we marginalize out f, θ remains in the condition.

Question 9

1. Set the prior distribution over W and visualise it.

2. Pick a single data point (x;t) and visualise the posterior distribution over W.

3. Draw 5 samples from the posterior and plot the resulting functions.

4. Repeat 2 − 3 by adding additional data points up to 7.

5. Given the plots explain the effect of adding more data on the posterior as well as the functions. How would you interpret this effect?

6. Finally, test the exercise for different values of σ, e.g. 0.1, 0.4 and 0.8. How does your model account for data with varying noise levels? What is the effect on the posterior?

Answer:

The prior over W is that the mean of weight w_0 =0, and the mean of weight w_1 =0. the distribution of weight W follows a multivariate gaussian distribution. Therefore, as shown in image 5, the distribution is shown as a circle.

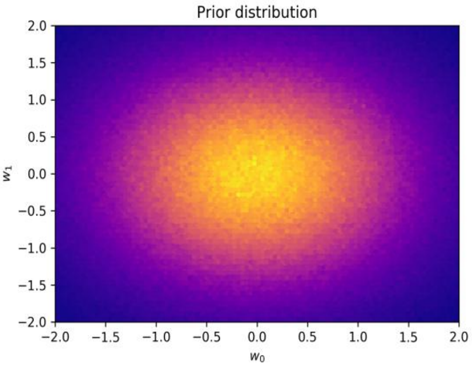


Image 5

My dataset has input x which are in range [-1,1], with step 0.01. I applied the linear model with the noise so I generated 200 x_i and their corresponding target values. I picked one data point, which is the 100th data point, from the data set. The posterior distribution is shown in image 6. Image 7 shows the five samples selected from this posterior distribution.

In terms of the color range, blue part means the probability is low. Yellow part means the probability is large.

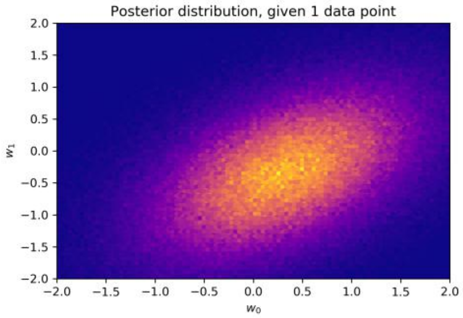


Image 6. posterior given 1 data point

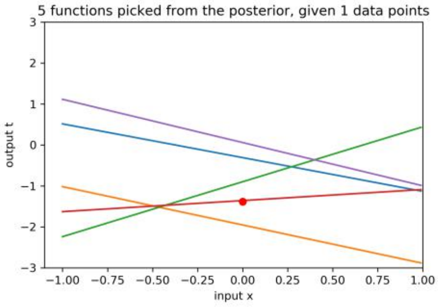


Image 7. five functions randomly drawn from posterior distribution, given 1 data point

I picked 5, 6, 7 data points. The posterior distribution is shown in image 8, image 10, image 12. Image 9 shows the five linear functions randomly drawn from the posterior distribution in image 8. Same goes images 10 and 11. Same goes images 12 and 13.

The result shows that, with the increasing amount of given data points, the posterior distribution become more precise. The mean of posterior distribution is closer to [0.5, -1.5]. So in the images, the center of the ellipse gets closer to the point [0.5,-1.5]. The variance of posterior distribution become smaller. So in the image 6,8,10,12, the size of the yellow ellipse becomes smaller and smaller. The center of the ellipse becomes brighter and brighter. It means that it has more confidence that the mean of weight is [0.5, -1.5]. When I provide 200 data points (image 14), the ellipse shrinks to the point [0.5,-1.5].

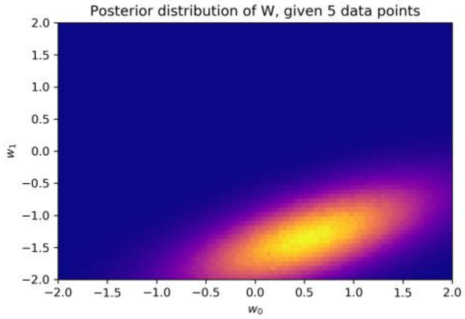


Image 8. posterior given 5 data point

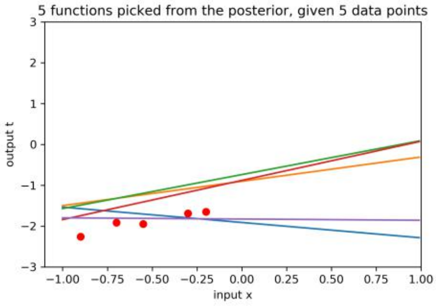


Image 9. five functions randomly drawn from posterior distribution, given 5 data point

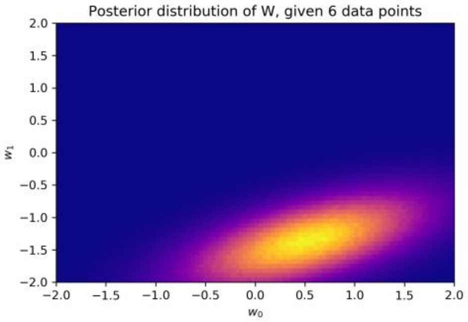


Image 10. posterior given 6 data point

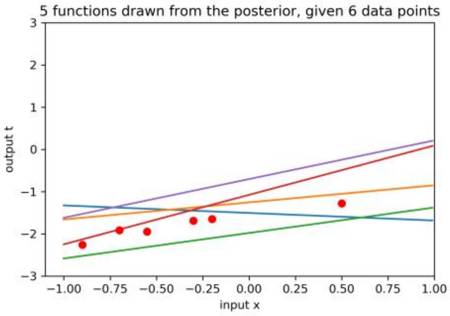


Image 11. five functions randomly drawn from posterior distribution, given 6 data point

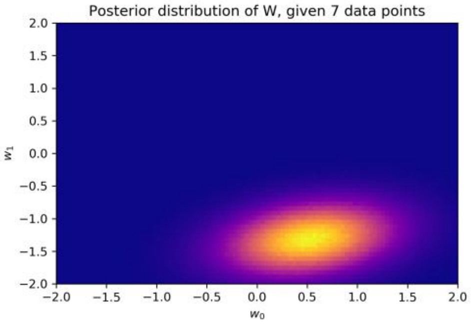


Image 12. posterior given 7 data point

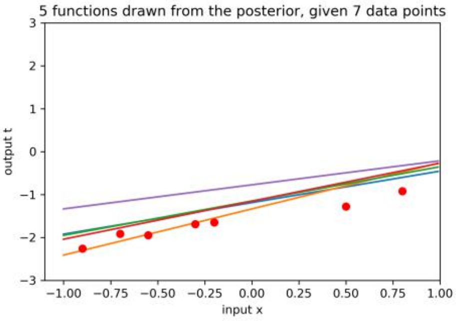


Image 13. five functions randomly drawn from posterior distribution, given 7 data point

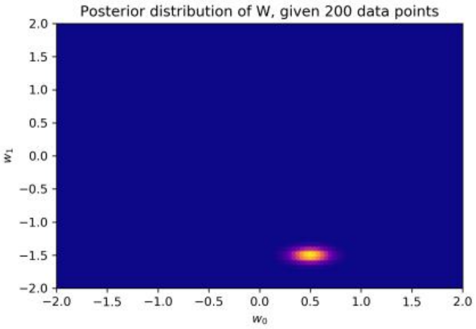


Image 14. posterior given 200 data point

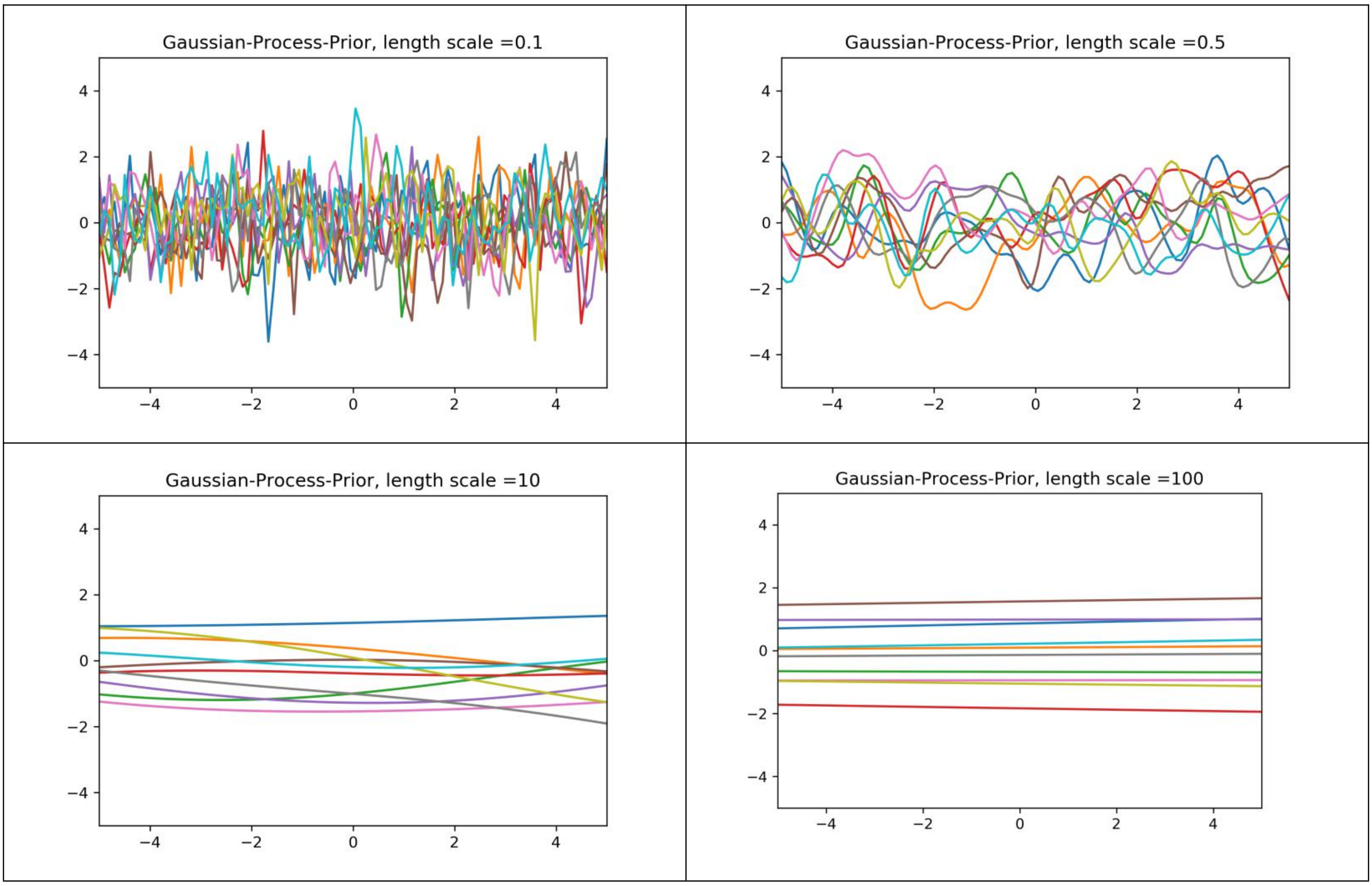
Values of σ controls the noise level in the dataset. Large values of σ means that the target value of the data has large uncertainty.Therefore, the larger the values of σ is, the more uncertainty the posterior has. The posterior with σ=0.8 has larger uncertainty than the posterior with σ=0.1.

Question 10
1. Create a GP-prior with a squared exponential covariance function.
2. For each of 4 different length scales, please draw 10 samples from this prior and visualise them. Explain the observed consequences of altering the length-scale of the covariance function.

Answer:
The squared exponential covariance function is:

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

The following 4 images shows the GP-prior with different length scale. On each image, 10 samples have been selected from the prior.



Length-scale l controls the smoothness of the prior. When l is 100 or goes to infinity, $-\{\frac{(x_i - x_j)^2}{l^2}\}$ approaches to 0. $e^{-\{\frac{(x_i - x_j)^2}{l^2}\}}$ becomes close to 1. So kernel K(x_i,x_j) reaches its maximum. This means that the dimensions in the covariance matrix reaches their largest correlation. They are highly correlated. Therefore, the value on one dimension is stable and it is similar to the value on the other dimensions.

When l is 0.1 or close to zero, $e^{-\{\frac{(x_i - x_j)^2}{l^2}\}}$ becomes close to 0. So kernel K(x_i,x_j) approaches to zero. This means that the dimensions in the covariance matrix are independent from each other. Their correlation is very low. Therefore, the value on each dimension varies a lot.

Question 11
1. What is the posterior before we observe any data?
2. Compute the predictive posterior distribution of the model.
3. Sample from this posterior with points both close to and far away from the observed data. Explain the observed effects.
4. Plot the data, the predictive mean and the predictive variance of the posterior from the data.
5. Compare the samples of the posterior with the ones from the prior. Is the observed behavior desirable?
6. *What would happen if you added a diagonal covariance matrix to the squared exponential?*

Answer:
1. What is the posterior before we observe any data?
Before we observe the data, the posterior is prior. Because we know nothing about the observations. The posterior is:

$$p(\mathrm{f}|\mathbf{x}, \theta) = N\,(\mathbf{0}; \mathrm{k}(\mathbf{x_i},\mathbf{x_j}))$$

2. Compute the predictive posterior distribution of the model.
Our goal is to make a prediction on the new input data. It means that we want to predict the target t_{N+1} for the new input x_{N+1}. Predictive posterior distribution is $p(t_{N+1}|\mathbf{t}_N)$. $\mathbf{t}_N$ is a vector $[t_1,t_2,\ldots,t_n]$. In order to find out $p(t_{N+1}|\mathbf{t}_N)$, we start with the joint distribution $p(\mathbf{t}_{N+1})$:

$$p(\mathbf{t}_{N+1}) = N(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

in which, $$\mathbf{C}_{N+1} = \begin{bmatrix} C_N & k \\ k^T & c \end{bmatrix}$$

Then we obtain predictive posterior distribution:

$$p(t_{N+1}|\mathbf{t}) = N\,(t_{N+1}\,|m(\mathbf{x}_{N+1}), \sigma^2(\mathbf{x}_{N+1}))$$

in which, mean: $m(\mathbf{x}_{N+1}) = k^T C_N^{-1} t$

covariance: $\sigma^2(\mathbf{x_{N+1}}) = c - k^T C_N^{-1} k$

$$\mathbf{C}_{N+1} = \begin{bmatrix} C_N & k \\ k^T & c \end{bmatrix}$$

the vector k has elements $k(x_n, x_{N+1})$ for n = 1, . . . , N . $c = k(x_{N+1}, x_{N+1}) + \beta^{-1}$. β^{-1} is the precision of the noise.

Image 15 shows how I generated the data points from the function and the noise. The function in light blue is the function: $(2+(0.5x_i-1)^2)*\sin(3x_i)$. After I computed the posterior distribution, I randomly picked 10 samples from the posterior. Image 16 shows the 10 samples. When the posterior is close to the data points, 10 functions approach to the observation (green dots). When the posterior is far away from the data points, the 10 functions vary a lot, as seen on the left side and right side of the green dots in image 16. Because no information is available.

Image 17 shows the data, the predictive mean and the predictive variance of the posterior.

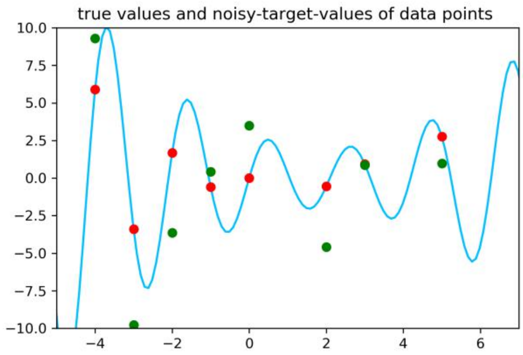


Image 15 The function in light blue is the function: $(2+(0.5x_i-1)^2)*\sin(3x_i)$. I used it to generate data points. Red dots denote the true values which are generated from the light blue function. Green dots denote: x_i with corresponding target values t_i. Target value t_i is the sum of true values and the noise.

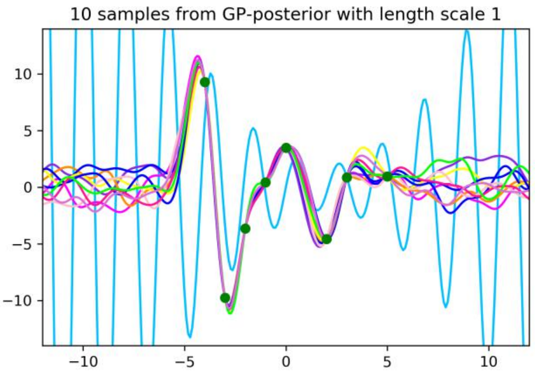


Image 16. 10 samples randomly picked from the gaussian process posterior distribution. The length scale is 1. The function in light blue is the function that I used to generate data points. Green dots denote the observations: x_i with corresponding target values t_i.

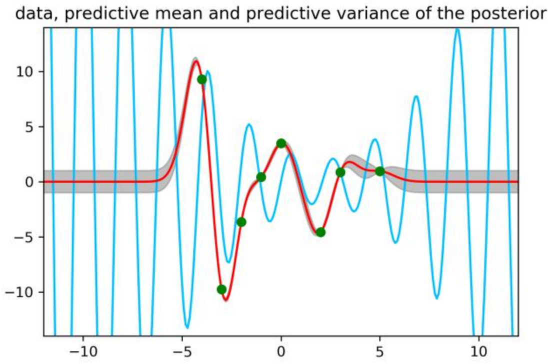


Image 17. Red line represents the predictive mean of the posterior, on all the infinite amount of dimensions. The grey area represents the predictive variance of the posterior, on all the infinite amount of dimensions. Green dots denote the target values.

5. Compare the samples of the posterior with the ones from the prior. Is the observed behavior desirable?
The samples in posterior are more similar to the function in light blue, when the data points are available. The samples in prior are too random.This behaviour is desirable. Because posterior learns knowledge from the given data points. But prior knows nothing about the data points.

Question 12 What type of "preference" for the latent variable X does this prior encode?
Answer:

We want to build up a simple linear relationship between X and weight W. So we specify the prior over the latent variables as spherical gaussian: p(**X**) = *N* (**0, I**) . It means that the dimensions of X are uncorrelated. The dimensions of X are independent from each other. The mean is zero.

Question 13 Perform the marginalisation in Eq. 23 and write down the expression. As previously, it is recommended that you do this by hand. In the answer outline the calculations and highlight the important steps.
Hint: The marginal can be computed by integrating out X with the use of Gaussian algebra we exploited in the exercise derivations and, in particular, by completing the square. However, it is much easier to derive the mean and covariance, knowing that the marginal is Gaussian, from the linear equation of Y(X).

Answer:
According to the linear equation Y(X), we know that each single y_i satisfies:
$y_i = Wx_i + \varepsilon$ and $\varepsilon \sim N(0, \sigma^2 I)$
So the likelihood function becomes:

$$p(y_i \mid x_i, W) = N(y_i \mid Wx_i, \sigma^2 I)$$

Then we convert the marginal distribution $p(Y \mid W) = \int p(Y \mid X, W) p(X) dX$ to each single data point:

$$p(y_i \mid W) = \int p(y_i \mid x_i, W) p(x_i) dx$$

For each single data point, we want to find out its mean and covariance matrix:

$$\begin{aligned} & E[y_i \mid W] \\ & = E[Wx_i + \varepsilon] \\ & = WE[x_i] + E[\varepsilon] \\ & = W\mu_{x_i} + \mu_\varepsilon \\ & = 0 \end{aligned}$$

$$\begin{aligned} & \operatorname{cov}[y_i \mid W] \\ & = E[(y_i - E[y_i])(y_i - E[y_i])^T] \\ & = E[(Wx_i + \varepsilon)(Wx_i + \varepsilon)^T] \\ & = E[Wx_i x_i^T W^T] + E[Wx_i \varepsilon^T] + E[W^T x_i \varepsilon] + E[\varepsilon\varepsilon^T] \\ & = E[Wx_i x_i^T W^T] + E[\varepsilon\varepsilon^T] \\ & = WW^T + \sigma^2 I \end{aligned}$$

We plug the mean and cov into the marginal distribution for each data point:

$$p(y_i \mid W) = N(y_i \mid 0, WW^T + \sigma^2 I)$$

Finally, we get the marginal distribution for the whole data set:

$$p(Y \mid W) = \prod_{i=1}^{N} N(y_i \mid 0, WW^T + \sigma^2 I)$$

Question 14 Compare the three different estimation procedures above in log-space.
1. What are their distinctive features and how are they different when we observe more data?
2. Why are the two last expressions of Eq. 25 equal?
3. Explain why Type-II Maximum-Likelihood is a sensible approach to learn the model.

Answer:

1. Maximum the log likelihood is shown in eq(3). Maximum the log likelihood function is equivalent to minimizing the sum-of-squares error function.

$$-\ln p(Y \mid X, W) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_i - W^T x_i)^2 + const \qquad \text{eq(3)}$$

Maximizing a posterior is shown in eq (4). Maximizing a posterior is equivalent to minimizing the sum-of-square error function with the regularization term.

$$-\ln p(W \mid X, Y) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_i - W^T x_i)^2 + \frac{1}{2}\sum_{n=1}^{N} w_i^2 + const \qquad \text{eq(4)}$$

Type-II Maximum-Likelihood is shown in eq(5). We maximize the marginal likelihood function obtained by integrating out the weight parameters.

$$-\ln p(Y \mid W) = \frac{N}{2}\ln(|WW^T - \sigma^2 I|) + \frac{1}{2}\sum_{i=1}^{N}(y_i^T (WW^T + \sigma^2 I) y_i) + const \qquad \text{eq(5)}$$

MAP has larger cost for computation than ML.

When we observe more data, MAP will approach to ML. Because p(W) is a constant. The MAP, which is p(Y|X,W), changes, when more data are fed in.

2.

$$W = \arg\max_W \frac{p(Y \mid X, W)p(W)}{\int p(Y \mid X, W)p(W)dW}$$

We know the Bayes theorem:

$$p(W \mid Y, X) = \frac{p(Y \mid X, W)p(W)}{p(Y \mid X)}$$

Because the denominator $\int p(Y \mid X, W)p(W)dW$ is independent from W. It can be seen as a constant. So maximizing is the whole term is same as maximizing the numerator.

3. Type-II Maximum Likelihood is a sensible approach because it is an empirical Bayes.

Question 15
1. Compute the objective function −log(p(YjW)) = L(W) for the marginal distribution in

Eq. 23.
2. Compute the gradients of the objective with respect to the parameters δδWL

Answer:
1.Compute the objective function −log(p(YjW)) = L(W) for the marginal distribution in Eq. 23.
We already know:

$$p(Y|W) = \prod_{i=1}^{N} N(y_i | 0, WW^T + \sigma^2 I)$$

Then we compute the log of p(Y|W):

$$\begin{aligned}
&\log(p(Y|W)) \\
&= \log[\prod_{i=1}^{N} N(y_i | 0, WW^T + \sigma^2 I)] \\
&= \sum_{i=1}^{N} \log[(y_i | 0, WW^T + \sigma^2 I)] \\
&= \sum_{i=1}^{N} \log[\frac{1}{(2\pi)^{\frac{D}{2}} (|WW^T + \sigma^2 I)^{\frac{1}{2}}} e^{-\frac{1}{2}(y_i-\mu)^T (WW^T+\sigma^2 I)^{-1} (y_i-\mu)}] \\
&= \sum_{i=1}^{N} [-\frac{D}{2}\log(2\pi)] - \sum_{i=1}^{N} [\frac{1}{2}\log(|WW^T + \sigma^2 I|)] - \frac{1}{2}\sum_{i=1}^{N} (y_i - \mu)^T (WW^T + \sigma^2 I)^{-1} (y_i - \mu) \\
&= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log(|WW^T + \sigma^2 I|) - \frac{N}{2} Tr((WW^T + \sigma^2 I)^{-1} YY^T)
\end{aligned}$$

We compute the negative log of p(Y|W):

$$\begin{aligned}
&L(W) \\
&= -\log(p(Y|W)) \\
&= \frac{ND}{2}\log(2\pi) + \frac{N}{2}\log(|WW^T + \sigma^2 I|) + \frac{1}{2} Tr((WW^T + \sigma^2 I)^{-1} YY^T)
\end{aligned}$$

in which, Tr (A) is the trace of a matrix A: $Tr(A) = \sum_{i=1}^{n} a_{ii}$

2.Compute the gradients of the objective with respect to the parameters $\frac{\partial L}{\partial W}$

$$\begin{aligned}
&\partial tr(X) = Tr(\partial X) \\
&\partial(\log[\det(X)]) = Tr(X^{-1}\partial X) \\
&\partial X^{-1} = -X^{-1}(\partial X)X^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial L(W)}{\partial W}\\
&=\frac{N}{2}tr[(WW^{T}+\sigma^{2}I)^{-1}(J_{ij}W^{T}+WJ_{ij}^{T})]+\\
&\frac{1}{2}tr\{YY^{T}[-(WW^{T}+\sigma^{2}I)^{-1}(J_{ij}W^{T}+WJ_{ij}^{T})(WW^{T}+\sigma^{2}I)^{-1}]\}
\end{aligned}
$$

Question 16

1. Plot the representation that you have learned (hint: plot X as a two-dimensional representation).

2. Explain the outcome and discuss key features, elaborate on any invariance you observe. Did you expect this result?

3. How is the effect of representation learning dependent on the number of available samples? Please test lower values of N and discuss the observed implications.

Answer:

W is initialized as a 10*2 matrix. Each element in W is randomly selected from gaussian distribution N(0,1).

In all my images in this question:

Red dots represent **X.**

$$\mathbf{X} = f_{\text{non-linear}}(x_{\text{true}})$$
$$= [\sin(x_i)-x_i*\cos(x_i), \cos(x_i)+x_i*\sin(x_i)],$$
in which $x_{\text{true}} = [0, 4\pi]$ with 200 elements.

Blue dots represent “the learned X” or “the recovered X”. It is written as **X’. X’** is calculated in the following way:

$$Y = \mathbf{X'}W^T$$
$$\rightarrow YA = \mathbf{X'}W^TW$$
$$\rightarrow \mathbf{X'} = YW(W^TW)^{-1}$$
in which, $Y = f_{\text{linear}}(f_{\text{non-linear}}(x_{\text{true}})) + \text{noise}$

The first step is to find the optimal weight W such that f=-log (p(Y|W)) reaches its minimum. The second step is to use this optimal W to calculate **X’**. Therefore, even though we could not observe x, we still are able to estimate the best weight W and are able to recover x.

Image 19, 20, and 21 are my results. In those 3 images, the noise were not added. The red dots represent X. Blue dots represent learned X (or **X’**). The rotation matrix is removed in the likelihood function. Because $W'W^T = WRR^TW^T = WW^T$. So we could not find a single unique solution for W. The weight in image 19, 20 and 21 are all optimal results. The recovered X (blue dots) are shown from different angles.

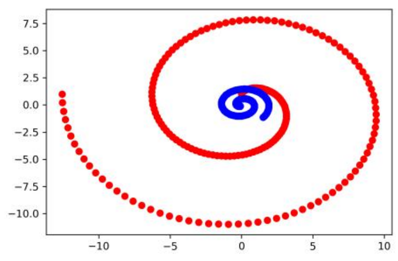


Image 19

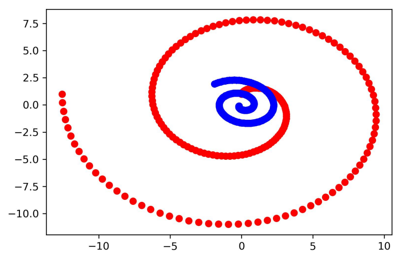


Image 20

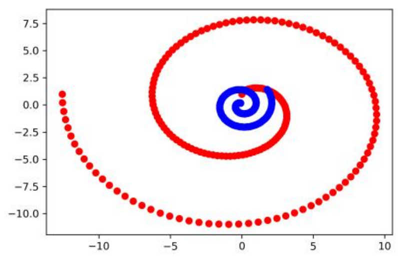


Image 21

After I added noise to Y, the results are shown in image 22 (noise variance=0.1) and 24 (noise variance=0.9). When noise variance is large, the oscillation in the recovered X is large. This can be seen by comparing image 23 and 25.

Y has noise with variance=0.1:

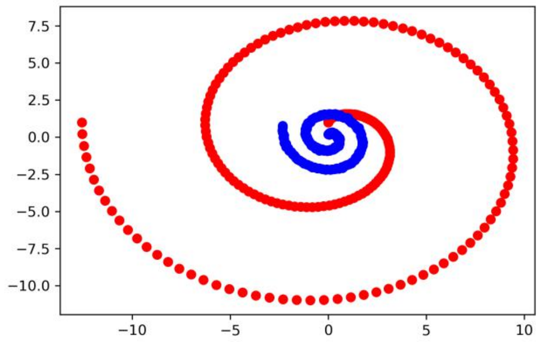


Image 22 X and learned X, noise variance =0.1

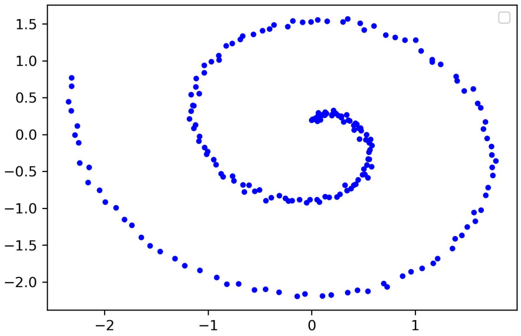


Image 23 zoomed in blue dots from image 22

Y has noise with variance=0.9

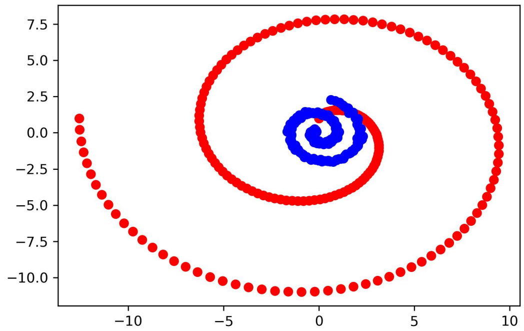


Image 24 X and learned X, noise variance =0.9

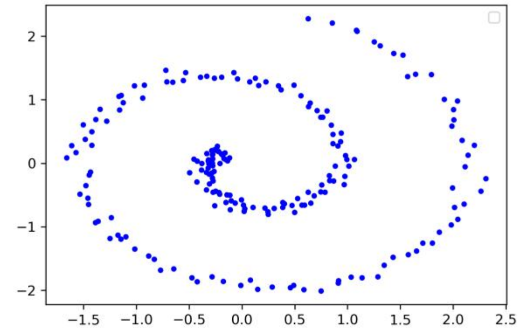
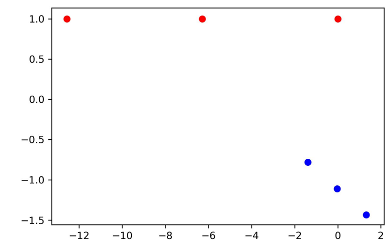


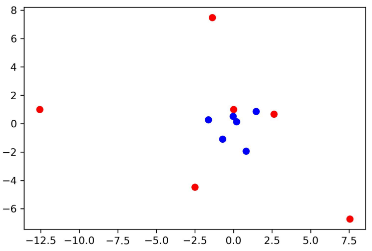
Image 25 zoomed in blue dots from image 24

The likelihood is invariant to rotations. W is invariant to rotations.

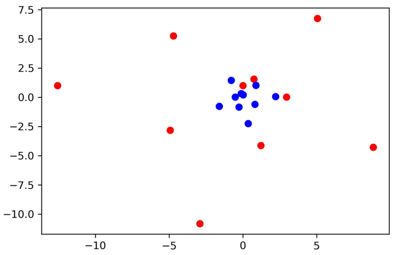
The number of available samples has an influence on the representation learning. The following images show the cases when N= 3,6,10,20,40. When N is too small (N=3), the learned X (blue dots) are very bad. The learned X are far away from the true X (red dots). When N increases (N=20, N=40), the learned X become similar to the true X. So we conclude that N should not be too small. If N is too small, the learned result is poor.



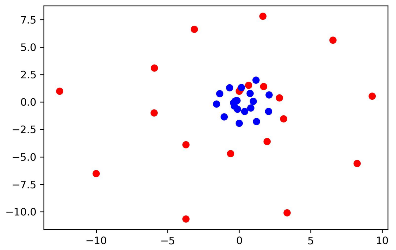
N=3



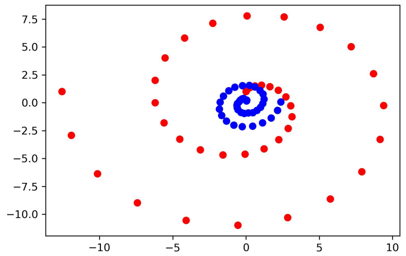
N=6



N=10



N=20



N=40

Question 17: Why is this the simplest model, and what does it actually imply? What makes it a bad model on the one hand, and a good model on the other hand?

Answer:

M_0 is the simplest model because M_0 treats all data sets equally. It assigns each data set the same probability which is $\frac{1}{512}$. M_0 does not have any free parameters.

The good side of M_0 is that it has the largest evidence over the whole range of all the possible data sets. The bad side of M_0 is that it is unable to assign much probability mass to simple data set. Because it assigns different types of behaviors the same probability. It does not use any information in the data sets.

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?

Answer:

M_3 is standard logistic regression. It is the most complex model among the four models (M_0, M_1,M_2,M_3). It has the most parameters and can realize the other models by setting some of its parameters to zero. Its flexibility: We expect it to spread its large unit probability mass over a wider range of data sets, compared with the other models. The drawback is that, for simple observations, H_3 is not suitable because it is over-complicated.

M_2 is the same as H_3 but without the bias weight w_0. M_2 gives a higher probability to data sets with decision boundaries crossing the origin. But M_3 with the bias term allows decision boundaries to be offset from the origin. M_2 includes both x_1 and x_2. So it is able to model decision boundaries for data sets that are due to rotation invariance. However this is not possible for M1

M_1 is the same as H_2 but it only includes the first dimension of x. M_1 is suitable to model the data set whose decision boundary is a function of x_1, but not x_2. M_1 has higher probabilities for data sets where the decision boundaries crosses the origin. M1 is not able to model decision boundaries with the rotation invariance.

Question 19: Discuss and compare the models. In particular, please address the following questions in your discussion

- *How have the choices we made above restricted the distribution of the model?*
- *What dataset is 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?*

Answer:

M1 is suitable for data set which decision boundary is a function of x_1, but not x_2. It has higher probabilities for data sets where the decision boundaries crosses the origin. M1 is not able to model the data set which has decision boundary with rotation invariance.

M2 is suitable for data set which decision boundary is a function of both x_1 and x_2. M2 stretches out more than M1. M2 is able to model the data set which has decision boundary with rotation invariance.

M3 is suitable for complicate data set. M3 stretches out more than M2 and more than M1. Because it has the bias term θ_3^3. The bias term allows decision boundary to be offset from the origin. M3 might have the risk of over-fitting, because of its high complexity.

M1 and M3 covers a large amount of data sets. So they have a large flexibility. The other points have been illustrated in question 18.

Question 20: Explain the process of marginalisation and briefly discuss its implications in the given context of the model evidence.

Answer:
The marginalization process means that we find out the evidence of a model p(D|M$_i$) by computing a weighted sum over p(D|M,θ). The weight is p(θ). p(θ) is our belief on the parameters in the model. After taking the integration, θ is marginalized out.

$$p(D \mid M_i) = \int_{\forall \theta} p(D \mid M_i, \theta) p(\theta) d\theta$$

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

Answer:
This choice of prior implies:
1) The dimensions of the parameters θ are independent. Because the off-diagonal elements in the covariance matrix Σ are zero.
2) The distribution of the parameters θ is sharply peaked around the most probable θ. Because the variance is very large: $\sigma^2 = 10^3$. We assume this is a sharp linear boundary.