MLAI Week 9 Exercise: Generative Models

Note: An indicative mark is in front of each question. The total mark is 7. You may mark your own work when we release the solutions.

$\boxed{1}$ 1. Slide 19: if the observed data point is $(x = -0.9, y = -0.1)$ instead, sketch what the likelihood will look like.

Solution: *Note: The solution provided here is made to be comprehensive to give you more insights. However, your answer will be considered as correct as long as you can sketch correctly to show what the likelihood looks like typically.*

Likelihood function (`https://en.wikipedia.org/wiki/Likelihood_function`) tells how likely the observed data can be generated by a given model. Given observed data $\mathcal{D}$, the likelihood is expressed as $p(\mathcal{D} \mid \mathbf{w})$ or more specificly $p(\mathbf{y} \mid \mathbf{w}, \mathbf{x})$ with $\mathbf{y} = f(\mathbf{x}, \mathbf{w})$. With the linear model defined by:

$$y = w_0 + w_1 x \tag{1}$$

To make the model generate the observed data point $(x = -0.9, y = -0.1)$, w_0 and w_1 need to satisfy:

$$-0.9 = w_0 - 0.1 w_1 \tag{2}$$

Eqn. (2) tells a model with any pair of w_0 and w_1 that satisfies it can confidently generate a data point like $(x = -0.9, y = -0.1)$, therefore the the likelihood w.r.t such pairs of w_0 and w_1 is 1. We can draw a line in the space formed by w_0 and w_1, showing the models with the parameters on the line having a high likelihood.

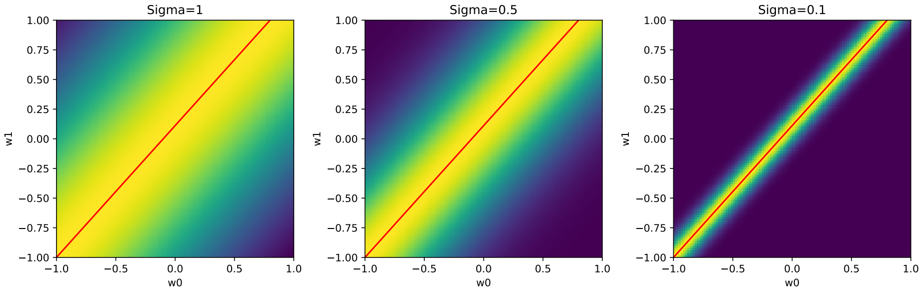
However, our belief is the output of the model is disturbed by noise which is independent from the input, thus the linear model becomes:

$$y = w_0 + w_1 x + \epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. With the linear model expressed by Eqn. (3), the likelihood function can be written as:

$$p(y \mid w_0, w_1, x) = \mathcal{N}(w_0 + w_1 x, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y - w_0 - w_1 x)^2} \tag{4}$$

With $(x = -0.9, y = -0.1)$ given, we can calculate Eqn. (4) in the space formed by w_0 and w_1. We show the sketch of the likelihood in the following figure under different σ value. Eqn. (2) is shown as the red straight line in the figure to illustrate the likelihood when Eqn. (1) is used as the model. Note that σ reflects the uncertainty of the output generated by Eqn. (3). As such, the smaller σ is, the more confident we are that the observed data is generated by the models that lies on the line expressed by Eqn. (2).

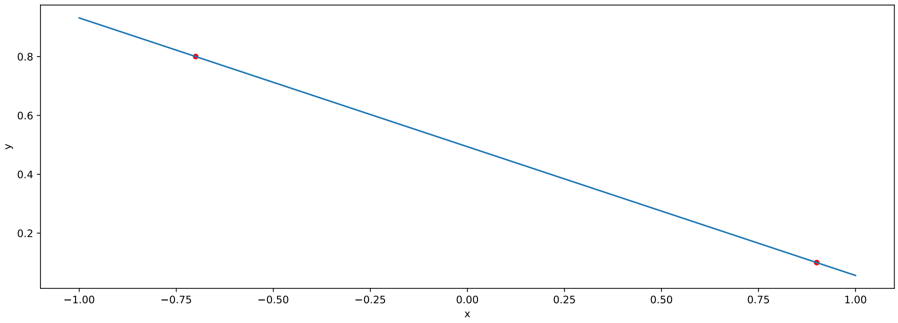


$\boxed{2}$ 2. Slide 20: if the second observed data point is $(x = -0.7, y = 0.8)$ instead, sketch what the posterior will look like on this slide, assuming the first observed data point is still as it is $(x = 0.9, y = 0.1)$.

Solution: *Note: The solution provided here is made to be comprehensive to give you more insights. However, your answer will be considered as correct as long as you can sketch correctly to show what the posterior looks like typically. You can also get the likelihood as above and multiply with the current prior.*

The posterior is a PDF (probability density function) of the parameters of model conditioned by the evidence (observed data $\mathcal{D}$), represented by $p(\mathbf{w} \mid \mathcal{D})$ or $p(\mathbf{w} \mid \mathbf{y}, \mathbf{x})$. It tells the probability of the model given the observed evidence.

Under the assumption of a linear model defined by Eqn. (1), if we observed two data points $(x = -0.7, y = 0.8)$ and $(x = 0.9, y = 0.1)$, we could draw a straight line connecting the two points to justify this is exactly the model we get after seeing the data points.



Therefore, the parameters of the model can be calculated by

$$\begin{aligned} -0.7w_1 + w_0 &= 0.8 \\ 0.9w_1 + w_0 &= 0.1 \end{aligned} \tag{5}$$

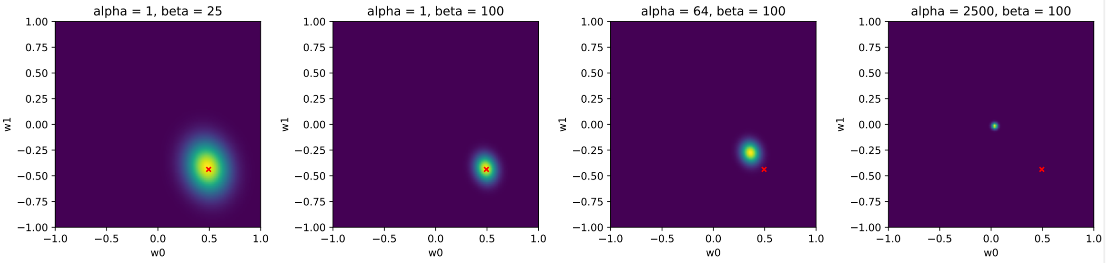
resulting in $w_0 = 0.4938, w_1 = -0.4375$. We can draw this point in the space formed by w_0 and w_1 to illustrate the posterior PDF at this point has a probability density of infinite value (CDF is 1 under an infinitely small area).

Similar to the first question, we believe the model should include noise and be represented by Eqn. (3). Therefore, the parameters of model can be relaxed from a fixed line. Intuitively, the posterior will look like to have an area centred by the point $w_0 = 0.4938, w_1 = -0.4375$ with high probability density. Next, let us show if our intuition is correct in a diagram.

With a Gaussian prior and Gaussian noise assumption:

$$\begin{aligned} p(\mathbf{w} \mid \alpha) &= \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}\right) \\ p(\epsilon) &= \mathcal{N}\left(\epsilon \mid 0, \beta^{-1}\right) \end{aligned} \tag{6}$$

where β^{-1} is the inverse variance, we can obtain a closed-form of the posterior. We borrow the formulas and implementation from A2 and A3 in lab9, using two data points $(x = -0.7, y = 0.8)$ and $(x = 0.9, y = 0.1)$, the posterior in the space of w_0 and w_1 is shown in the following diagram. The red cross dot represent the point $w_0 = 0.4938, w_1 = -0.4375$. Different combinations of α and β are used in the evaluation.



The diagram shows that with a larger β (inverse variance), the area has a smaller size. Because the output of the model is less uncertain with a large β, the possible models are more similar to the fixed line represented by the red cross dot. Moreover, the diagram shows that our aforementioned intuition is incorrect when α is large. This is because the prior $p(\mathbf{w} \mid \alpha) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}\right)$ with a large α will become a dominant factor over the likelihood in the posterior expressed by Eqn. (7). The possible parameters generated by the Gaussian prior is mostly confined to a small area with mean value as centre given a large α. The results show that our belief on the prior is critical to the model learned from the evidence.

$$p(\mathbf{w} \mid \mathbf{y}, \mathbf{x}, \alpha) \propto p(\mathbf{y} \mid \mathbf{w}, \mathbf{x}) p(\mathbf{w} \mid \alpha) \tag{7}$$

$\boxed{1}$ 3. Slide 26: What is/are the sufficient statistics for a Bernoulli distribution?

Solution:
Let $X_1, \cdots, X_n$ be iid Bernoulli random variables with parameter π, $0 < \pi < 1$. Then $\sum_{i=1}^{n} X_i$ is a sufficient statistic for π.

$\boxed{3}$ 4. Slide 36: show how to obtain a variable z with a normal distribution of mean μ and standard deviation (std) σ from a standard normal distribution with a mean of zero and std of 1 and verify the mean and std of z are indeed μ and σ respectively.

Solution: $x \sim \mathcal{N}(0, 1)$
We know that:

$\mathbb{E}[x] = 0$
$Var(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1$
We can obtain $z \sim \mathcal{N}(\mu, \sigma)$ by setting $z = \mu + \sigma x$.

Let us prove it now.

Mean:
$\mathbb{E}[z] = \mu + \sigma\mathbb{E}[x]$
$= \mu + \sigma(0) = \mu$
$Var(z) = \mathbb{E}[z^2] - \mathbb{E}[z]^2$
$= \mathbb{E}[(\mu + \sigma x)^2] - (\mu)^2$
$= (\mu^2 + 2\mu\sigma\mathbb{E}[x]) + \sigma^2\mathbb{E}[x^2] - (\mu)^2$
Looking at the defined formula for $Var(x)$ above. We know that:
$\mathbb{E}[x^2] = 1 - \mathbb{E}[x]^2 = 1$
Therefore:
$Var(z) = \mu^2 + 2\mu\sigma(0) + \sigma^2(1) - \mu^2$
$= \sigma^2$