Week 10 Exercise Sheet Solutions

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time or effort to solve compared to an exercise with (***), which is a more complex exercise.

Generative Models

1. (**) Bayesian linear regression: Consider the example given in the lecture, where we are trying to fit $\tilde{y}_n = w_0 + w_1 x_n + e_n$ to our observations. w_0 and w_1 are the weights we are optimising, x_n and y_n are our observations, $\tilde{y}_n$ is our prediction of the observation and $e_n \sim \mathcal{N}(0, \sigma)$ is a Gaussian noise term with standard deviation σ. We assumed a Gaussian form for the likelihood

$$P(y_n|x_n, w_0, w_1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_n - \tilde{y}_n)^2}{2\sigma^2}\right) \tag{1}$$

and prior

$$P(\mathbf{w}|\bar{\mathbf{w}}, \mathbf{\Sigma}) = \frac{1}{\sqrt{2\pi|\mathbf{\Sigma})|}} \exp\left(-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})\mathbf{\Sigma}^{-1}(\mathbf{w} - \bar{\mathbf{w}})\right) \tag{2}$$

where $\mathbf{w} = (w_0, w_1)^T$, $\bar{\mathbf{w}}$ is the mean of the weight distribution and $\mathbf{\Sigma}$ is the covariance matrix of the weights.

a) If our first observed datapoint is $x = -0.9$, $y = -0.1$. Sketch the likelihood as a function of w_0 and w_1.

b) If the initial prior distribution is isotropic with $\bar{\mathbf{w}} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{I}$ ($\mathbf{I}$ is the identity matrix). Sketch what the posterior will be after multiplying the prior and likelihood from a).

Solution:

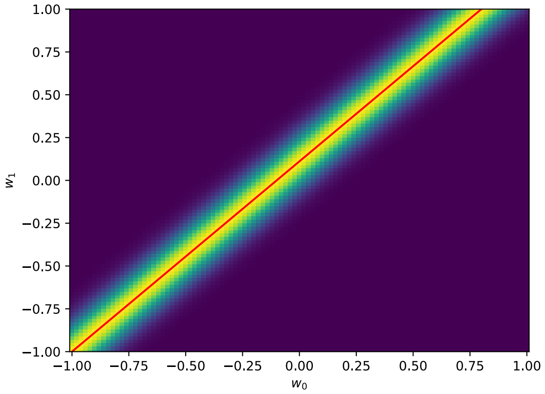
a) The likelihood is expressing the probability of observing a particular observation given the model parameters. Given our first observation that $y = -0.1$ when $x = -0.9$, we can use our linear model to determine which weights are likely to have produced this data. Since we assume a Gaussian noise term then the likelihood for any datapoint is

$$P(y_n|x_n, w_0, w_1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_n - w_0 - w_1 x_n)^2}{2\sigma^2}\right). \tag{3}$$

Note: We have inserted the expression for $\tilde{y}_n$ into the likelihood but we do not need to insert e_n as this is given by the Gaussian distribution. So using the measured values for x and y we have the likelihood function for the first observation as

$$P(y_1 = -0.1|x_1 = -0.9, w_0, w_1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(-0.1 - w_0 + w_1 0.9)^2}{2\sigma^2}\right) \tag{4}$$

which we looks as follows when $\sigma = 0.1$: The red line shows the maximum probability, which given we only have 1 observation defines a line. This line is given by



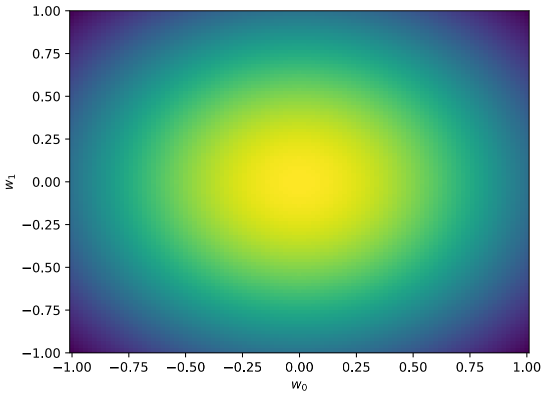
$$\begin{aligned} y_1 &= w_0 + w_1 x_1 \\ w_1 &= \frac{y_1 - w_0}{x_1} \qquad (5) \\ w_1 &= 0.111 + 1.111 w_0. \qquad (6) \end{aligned}$$

b) Using Bayes' rule the Posterior is given by (to simplify the expression w is written in vector form here):

$$P(\mathbf{w}|y_n, x_n) \propto P(y_n|x_n, \mathbf{w}, \sigma) P(\mathbf{w}|\bar{\mathbf{w}}, \mathbf{\Sigma}) \tag{7}$$

where we are assuming that the denominator (the evidence or marginal likelihood) is a normalising factor and is thus given as a proportionality. Here it is specified that the prior is isotropic with $\bar{\mathbf{w}} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{I}$. This means the equation is simplified to

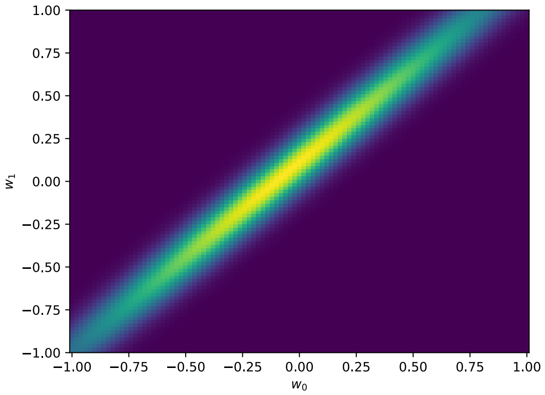
$$P(\mathbf{w}|\bar{\mathbf{w}}, \mathbf{\Sigma}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(w_0^2 + w_1^2)\right) \tag{8}$$



which looks as follows: If we multiply the likelihood in part a) with this function we get an expression for the posterior

$$P(\mathbf{w}|y_n, x_n) \propto \frac{1}{2\pi\sigma} \exp\left(-\frac{(-0.1 - w_0 + w_1 0.9)^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}(w_0^2 + w_1^2)\right) \tag{9}$$

which if we plot looks as follows where we note that is still follows the same line as



in part a) but since our initial assumption (from the prior) favoured weight values closer to zero then our new distribution does as well. Further observations will make the distribution narrower since we will have more information.

---

2. (**) What is/are the sufficient statistics for a Bernoulli distribution? (You will need to do further reading on sufficient statistics to find this).

Solution:
The Bernoulli distribution concerns the probability of binary outcomes (e.g yes or no) and the parameter π determines the number of position outcomes in the

distribution. The sufficient statistic provide all the information needed to estimate this parameter. Therefore, given a set of independent and identically distributed (iid) Bernoulli random variables $X_1, X_2, \ldots, X_n$ then the sufficient statistic of π is $\sum_{i=1}^{n} X_i$.

---

3. (**) Show how to obtain a variable z with a Gaussian distribution of mean μ and standard deviation (std) σ from a standard Gaussian distribution with a mean of zero and std of 1. Verify that the mean and std of z are indeed μ and σ respectively.

Solution:
If z is a Guassian random variable with a mean μ and standard deviation σ, such that $z \sim \mathcal{N}(\mu, \sigma)$ then it can be expressed as

$$z = \mu + \sigma x \tag{10}$$

where $x \sim \mathcal{N}(0, 1)$ (i.e Gaussian RV with zero mean and std of 1) We can verify this; first consider the mean:

$$\begin{aligned}\mathbb{E}[z] &= \mathbb{E}[\mu + \sigma x] \\ &= \mu + \sigma \mathbb{E}[x] \\ &= \mu \end{aligned} \tag{11}$$

since $\mathbb{E}[x] = 0$. For the standard deviation we use the definition of the variance:

$$\begin{aligned}\text{Var}(z) &= \mathbb{E}[z^2] - \mathbb{E}[z]^2 \\ &= \mathbb{E}[(\mu + \sigma x)^2] - \mu^2 \\ &= \mathbb{E}[\mu^2 + 2\sigma\mu x + \sigma^2 x^2] - \mu^2 \\ &= 2\sigma\mu\mathbb{E}[x] + \sigma^2\mathbb{E}[x^2] \\ &= \sigma^2\mathbb{E}[x^2]. \end{aligned} \tag{12}$$

Since the standard deviation of x is 1 then

$$\begin{aligned}\text{Var}(x) &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\ &= \mathbb{E}[x^2] = 1. \end{aligned} \tag{13}$$

This means that the variance of z is

$$\text{Var}(z) = \sigma^2 \tag{14}$$

which verifies the re-parameterisation of z.
