1 ELBO diagonal Gaussian

The general form of the ELBO is:

$$\mathbb{E}_{q(w_l)}[\sum_{i=1}^{n} p(y_i|f(x_i)) + D_{KL}(q(w_l)\|p(w_l))]$$

1.1 KL-divergence

We set the prior $p(w_l)$ as well as q to a diagonal gaussian

$$\begin{aligned}
D_{KL}(q(w_l)\|p(w_l)) &= \mathbb{E}_{q(w_l)}[log(\frac{p(w_l)}{q(w_l)})] \\
&= \mathbb{E}_{q(w_l)}[log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_q)}} exp(-\frac{1}{2}(w_l - \mu_q)^T \Sigma_q^{-1}(w_l - \mu_q))) \\
&- log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_p)}} exp(-\frac{1}{2}(w_l - \mu_p)^T \Sigma_p^{-1}(w_l - \mu_p)))] \\
&= \mathbb{E}_{q(w_l)}[log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_q)}}) - \frac{1}{2}(w_l - \mu_q)^T \Sigma_q^{-1}(w_l - \mu_q) \\
&- log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_p)}}) + \frac{1}{2}(w_l - \mu_p)^T \Sigma_p^{-1}(w_l - \mu_p)] \\
&= \mathbb{E}_{q(w_l)}[\frac{1}{2} log(\frac{det(\Sigma_p)}{det(\Sigma_q)}) - \frac{1}{2}(w_l - \mu_q)^T \Sigma_q^{-1}(w_l - \mu_q) \\
&+ \frac{1}{2}(w_l - \mu_p)^T \Sigma_p^{-1}(w_l - \mu_p)] \\
&= \frac{1}{2}(log(\frac{det(\Sigma_p)}{det(\Sigma_q)}) - n + tr(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)^T \Sigma_p^{-1}(\mu_p - \mu_q))
\end{aligned}$$

As we assumed Σ is diagonal:

$$= \frac{1}{2}\sum_{i=1}^{n} log(\frac{\Sigma_{p_i}}{\Sigma_{q_i}}) - n + \frac{\Sigma_{q_i}}{\Sigma_{p_i}} + (\mu_{p_i} - \mu_{q_i})^2 \Sigma_{p_i}^{-1} \tag{1}$$

If we use a prior of $\mu_p = 0$ and $\Sigma_p = Id$, then

$$= \frac{1}{2}\sum_{i=1}^{n} -log(\Sigma_{q_i}) - n + \Sigma_{q_i} + \mu_{q_i}^2 \tag{2}$$

1.2 Likelihood

Assuming $f_w(x)$ outputs a probability distribution over the classes, we use a categorial distribution for the likelihood

$$p(y_i\|f_w(x_i)) = \prod_{j=1}^{k} f_w(x_{i_j})^{y_{i_j}}$$

If we have just one true class l and one-hot encoded, we have:

$$= f_w(x_{i_l})$$

2 Predictions

$$\begin{aligned} p(y|f_w(x)) &= \mathbb{E}_{p(w)}[p(y|f_w(x))] \\ &\approx E_{q(w)}[p(y|f_w(x))] \\ &\approx \frac{1}{M}\sum_{i=1}^{m} P(y|f_{w_i}(x)) \end{aligned}$$