KL Divergence between 2 Gaussian distributions:-

KL Divergence between 2 distributions P and Q of a continous random variable is given by:-

$$D_{KL}(P||Q) = \int_{\mathbb{X}} p(x) \log \frac{p(x)}{q(x)}$$

And probability density function of multivariate Normal distribution is given by:-

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

Now, let our 2 normal distributions be $\mathcal{N}(\mu_p, \Sigma_p)$ and $\mathcal{N}(\mu_q, \Sigma_q)$ both k dimensional.

$$D_{KL}(P||Q) = E_p[\log p(x) - \log q(x)]$$

$$E_p\left[\frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p) + \frac{1}{2}(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)\right]$$

$$\frac{1}{2}E_p\left[\log\frac{|\Sigma_q|}{|\Sigma_q|}\right] - \frac{1}{2}E_p[(x-\mu_q)^T\Sigma_p^{-1}(x-\mu_p)] + \frac{1}{2}E_p[(x-\mu_q)\Sigma_q^{-1}(x-\mu_q)]$$

Scanned with CamScanner

$$= \frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}E_p[(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)] + \frac{1}{2}E_p[(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q)]$$

Now since $(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)$ in the second term $\in \mathbb{R}$, we can write it as $tr\{(x-\mu_p)^T\Sigma_p^{-1}(x-\mu_p)\}$ where $tr\{\}$ is the trace operator.

And we can write it as :-

$$tr\{(x-\mu_p)(x-\mu_p)^T\Sigma_p^{-1}\}$$

The second term now is :-

$$= \frac{1}{2}E_p[tr\{(x-\mu_p)(x-\mu_p)^T\Sigma_p^{-1}\}]$$

The expectation and trace can be interchanged to get

$$= \frac{1}{2}tr\{E_p[(x-\mu_p)(x-\mu_p)^T\Sigma_p^{-1}]\}$$

$$= \frac{1}{2}tr\{E_p[(x-\mu_p)(x-\mu_p)^T]\Sigma_p^{-1}\}$$

We know $E_p[(x-\mu_p)(x-\mu_p)^T] = \Sigma_p$ Simplifying it to.

Scanned with CamScanner

$$= \frac{1}{2} \text{tr}\{\Sigma_p \Sigma_p^{-1}\}$$

$$= \frac{1}{2} \text{tr}\{I_k\}$$

$$= \frac{k}{2}$$

We can simplify the third term, we get:-

$$\mathbb{E}_p[(x-\mu_q)^T \Sigma_q^{-1} (x-\mu_q)] = (\mu_p-\mu_q)^T \Sigma_q^{-1} (\mu_p-\mu_q) + \text{tr}\{\Sigma_q^{-1}\Sigma_p\}$$

Combining all this we get:-

$$D_{KL}(p||q) = \frac{1}{2}\left[\log\frac{|\Sigma_q|}{|\Sigma_p|} - k + (\mu_p-\mu_q)^T \Sigma_q^{-1}(\mu_p-\mu_q) + \text{tr}\{\Sigma_q^{-1}\Sigma_p\}\right]$$

When q is $N(0, I)$, we get

$$D_{KL}(p||q) = \frac{1}{2}\left[\mu_p^T \mu_p + \text{tr}\{\Sigma_p\} - k - \log|\Sigma_p|\right]$$

Scanned with CamScanner