We are interested in interpreting models like regression or logistic regression, where we minimize

$$\frac{1}{N}\sum_{n=1}^{N}\|t_n - (w^\intercal x_n + b)\|^2 \tag{1}$$

or for logistic regression

$$\frac{1}{N}\log p(y_n|x_n, w) \tag{2}$$

For regression, we note that this is equivalent to the likelihood of a linear Gaussian distribution, $t_n \sim \mathcal{N}\left(w^\intercal x_n + b, \sigma^2\right)$, which has a likelihood of

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{1}{2}\frac{\|t_n-(w^\intercal x_n+b)\|^2}{\sigma^2}} \tag{3}$$

and a log likelihood of

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N}\left(-\frac{\|t_n - (w^\intercal x_n + b)\|^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)\\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{N}\sum_{n=1}^{N}\|t_n - (w^\intercal x_n + b)\|^2\end{aligned} \tag{4}$$

which is regression with constant scaling and a constant offset, neither of which is affected by the parameters w, b.

Knowing that we are once again maximizing log likelihood, we will extend our result that maximum likelihood is the same as KL minimization, to conditional distributions / discriminative models.

$$\arg\min_w KL\left(p_{data}(y|x)\|p(y|x,w)\right) \approx \arg\max_w \frac{1}{N}\sum_{n=1}^{N}\log p(y_n|x_n, w) \tag{5}$$

1 Information Theory

$KL\left(p_{data}(y|x)\|p(y|x,w)\right)$ is known as conditional KL or conditional relative entropy.

To be formal, we should define conditional entropy, which involves an outer expectation and an inner conditional expectation

$$H(p(y|x) = -E_{x\sim p(x)}\left[E_{y\sim p(y|x)}\left[\log p(y|x)\right]\right] \tag{6}$$

Conditional KL similarly modifies KL by adding an additional outer expectation

$$KL\left(p(y|x)\|q(y|x)\right) = E_{x\sim p(x)}\left[E_{y\sim p(y|x)}\left[\log\frac{p(y|x)}{q(y|x)}\right]\right] \tag{7}$$

In fact, this is the same as $KL\left(p(x)p(y|x)\|p(x)q(y|x)\right)$, the KL of the joint if they share the same marginal distribution over x.

Like before, we choose p to be our data distribution and q to be our model distribution, which is parameterized by w.

Then minimizing KL is once again minimizing cross entropy

$$\arg\min_w KL\left(p_{data}(y|x)\|p(y|x,w)\right) = \arg\min_w -E_{x\sim p_{data}(x)}\left[E_{y\sim p_{data}(y|x)}\left[\log p(y|x,w)\right]\right] \tag{8}$$

Once again we can do Monte Carlo estimation of the cross entropy (which we could also write as an expectation over an empirical distribution), yielding a log likelihood estimator

$$\frac{1}{N}\sum_{n=1}^{N}\log p(y_n|x_n,w) \to E_{x\sim p_{data}(x)}\left[E_{y\sim p_{data}(y|x)}\left[\log p(y|x,w)\right]\right] \tag{9}$$

Therefore maximizing log likelihood for a discriminative model is equivalent to minimizing conditional KL.

2 Gradient updates for linear regression / generalized linear models

If we wanted to perform gradient updates on the weights, we could use the chain rule on our previous result about the gradient of the log likelihood with respect to natural parameters, noting that $\eta(x_n, w, b) = w^\intercal x_n + b$

$$\begin{aligned}\nabla_w \log p(y_n|x_n,w) &= \nabla_\eta \log p(y_n|\eta(x_n,w,b)) \cdot \nabla_w \eta(x_n,w,b) \\ &= \left(y_n - E_{y\sim p(y|\eta(x_n,w,b))}\left[y\right]\right)\cdot w\end{aligned} \tag{10}$$