Kullback-Leibler Divergence

by Alvin Wan . alvinwan.com/cs189

This document explores the implications of Kullback-Leibler (KL) Divergence and how it relates to both cross entropy and logistic regression. We will derive cross entropy from KL-divergence and coerce log loss with the derivation of logistic regression presented in Note 11: Logistic Regression.

1 Cross Entropy

In Note 11, we showed that our assumptions led to a log-linear model. As it turns out, this model is the least biased within constraints.

In logistic regression, we would like to find a probability distribution that accurately represents our model. In other words, we consider the best projection of empirical probabilities onto a log-linear model. To minimize the difference between an empirical probability distribution and the log-linear probability distribution, we need a measure of "divergence", which KL-divergence provides. In the derivation below, we will show how minimizing KL-divergence is equivalent to minimizing cross entropy.

Consider the definition of cross entropy, for the true distribution p_i and predicted distribution q_i. The entropy of p and q ($H(p,q)$) is the sum of the true distribution's entropy ($H(p)$) and the KL-divergence of q from p ($K(p||q)$).

$$H(p,q) = H(p) + K(p||q)$$

Note that $H(p)$ is a constant. To minimize cross entropy, we thus minimize the KL-divergence. We can re-express KL-divergence using the following. For the second step, we apply chain rule. In the third step, we note that all $p(\cdots)$ are constants.

$$
\begin{aligned}
K(p||q) &= \int_{(x,y)} p(x,y)\log(\frac{p(x,y)}{q(x,y)}) \\
&= \int_{(x,y)} p(x,y)\log(p(x,y)) - \int_{(x,y)} p(x,y)\log(q(x,y)) \\
&= \int_{(x,y)} p(x,y)\log(p(x,y)) - \int_{(x,y)} p(x,y)\log(q(y|x)p(x)) \\
&= \int_{(x,y)} p(x,y)\log(p(x,y)) - \int_{(x,y)} p(x,y)\log(p(x)) - \int_{(x,y)} p(x,y)\log(q(y|x)) \\
&= C - \int_{(x,y)} p(x,y)\log(q(y|x))
\end{aligned}
$$

The true distribution is unknown but can be estimated using the training data $\{x_i, y_i\}_{i=1}^n$. Thus we take the following.

$$\text{minimize}\, H(p,q) = \text{minimize}\, K(p||q) = \text{minimize} - \sum_i \log(q(y_i|x_i,\theta))$$

This is precisely our formulation for log loss, or cross entropy. Note that $p(x_i, y_i)$ is always 1. We take the average for our log loss definition:

$$-\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

2 Logistic Regression

In Note 11, we applied MLE, assuming a uniform prior and taking class-conditional probabilities to be the sigmoid. This yielded the following, where y_i are labels, and μ_i are the predicted labels.

$$\sum_i y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)$$

Note that the negative, normalized version of this quantity gives us the definition of cross entropy. Thus, maximizing this quantity is equivalent to minimizing the quantity in the previous section.