Logistic Regression

Pattern Recognition And Machine Learning

Indian Institute of Technology Madras

November 4, 2019

Entropy and Cross Entropy

Entropy

Entropy is a measure of average information in the system. For discrete random variable X, the entropy is defined as follows:

$$H(P(X)) = E_{P(X)}\left[-\log P(X)\right] \tag{1}$$

Cross Entropy

The cross entropy between the actual and the predicted labels can be used as a loss function to train classifiers. Its defined as follows

$$H(P(X), Q(X)) = E_{P(X)}\left[-\log Q(x)\right] \tag{2}$$

Why cross entropy?

Let P_i be the true distribution of the data, and Q_i be estimated distribution. Ideally we would like $KLD(P, Q) \approx 0$, when estimated distribution Q matches the true distribution P. The Kullack Leibler Divergence between two discrete distributions is given by (which is always positive):

$$\begin{aligned} KLD(P, Q) &= \sum_i P_i \log \frac{P_i}{Q_i} \\ &= \sum_i P_i \log P_i + \sum_i P_i log Q_i \end{aligned}$$

The first term is the input entropy which cannot be reduced. We therefore reduce the second term which is the cross entropy between the true distribution and the distribution estimated from the data.

This is equivalent to the minimizing the negative of log likelihood in maximum likelihood estimation technique, where we try to minimize the cross entropy between $\frac{1}{N}$, and distribution $p(\bar{x})$ estimated from the data.

Logistic Regression for 2 Class problem

$$P(c_1|\bar{x}) = \frac{P(\bar{x}|c_1)\ P(\bar{c_1})}{P(\bar{x}|c_1)\ P(\bar{c_1}) + P(\bar{x}|c_2)\ P(\bar{c_2})} \tag{3}$$

$$P(c_1|\bar{x}) = \sigma(a) = \frac{1}{1 + e^{-a}} \tag{4}$$

Logistic Regression for Multi Class problem

$$P(c_k|\bar{x}) = \frac{P(\bar{x}|c_k)\ P(\bar{c_k})}{\sum_{j=1}^{K} P(\bar{x}|c_j)\ P(\bar{c_j})} \tag{5}$$

$$P(c_1|\bar{x}) = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}} \tag{6}$$

a is modelled as function of input features defined by weights W (recall linear regression):

$$a = \mathbf{w}^t \Phi(\bar{x}) \tag{7}$$

Logistic regression (estimate parameters of the discriminant function for classification)

We use Logistic Regression to estimate the boundary between a PAIR of classes, where $t_n = \{0, 1\} \forall n = 1 \; to \; N$ is the class label for each example.
Let the training data set be $D = \{\bar{x}_n, t_n\}_{n=1}^{N}$, where $t_n = \{0, 1\} \forall n = 1 \; to \; N$.
Let $\phi(\bar{x})$ be a set of basis function defined as follows:

$$\phi(\bar{x}) = [\phi_0(\bar{x}), \; \phi_1(\bar{x}), \ldots, \; \phi_m(\bar{x})] \tag{8}$$

The objective is to find the weights $\mathbf{w}$ such that likelihood of observing t_n given W is maximum. The likelihood function is given below:

$$P(t|\mathbf{w}) = \prod_{n=1}^{N} P(t_n|\mathbf{w}) \tag{9}$$

$$= \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n} \tag{10}$$

Logistic regression (estimating parameters)(contd..)

Minimizing the negative log-likelihood will maximize the likelihood given in Eq 9. Therefore, the objective is to minimize

$$\Sigma(W) = -\log P(t|\mathbf{w}) \tag{11}$$

$$= -\log\left(\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\right) \tag{12}$$

$$= -\sum_{n=1}^{N} \log\left(y_n^{t_n}(1-y_n)^{1-t_n}\right) \tag{13}$$

$$= -\sum_{n=1}^{N} \log\left(y_n^{t_n}\right) + \log\left((1-y_n)^{1-t_n}\right) \tag{14}$$

$$= -\sum_{n=1}^{N} t_n \log\left(y_n\right) + (1-t_n)\log\left((1-y_n)\right) \tag{15}$$

Logistic regression (estimating parameters)(contd..)

To maximize the likelihood given in Eq 9, we need to take the derivative Eq 15 with respect to W.

$$\frac{\partial \Sigma(\mathbf{w})}{\partial \mathbf{w}} = -\sum_{n=1}^{N} \frac{t_n}{y_n}\frac{\partial y_n}{\partial \mathbf{w}} - \frac{1-t_n}{1-y_n}\frac{\partial y_n}{\partial \mathbf{w}} \tag{16}$$

$$\frac{\partial y_n}{\partial \mathbf{w}} = \frac{\partial \sigma(\mathbf{w}^t \phi(\bar{x}_n))}{\partial \mathbf{w}} \tag{17}$$

$$= \sigma(\mathbf{w}^t \phi(\bar{x}_n))(1 - \sigma(W^t \phi(\bar{x}_n)))\phi(\bar{x}_n) \tag{18}$$

$$= y_n(1-y_n)\phi(\bar{x}_n) \tag{19}$$

$$\therefore \frac{\partial \Sigma(\mathbf{w})}{\partial \mathbf{w}} = -\sum_{n=1}^{N}(t_n - y_n)\phi(\bar{x}) \tag{20}$$

$$= \sum_{n=1}^{N}\left(\frac{1}{1+e^{\mathbf{w}^t\phi(\bar{x})}} - t_n\right)\phi(x) \tag{21}$$

Logistic regression (estimating parameters)(contd..)

Since there is no closed form solution for Eq 20, we approximate the solution using gradient descent as demonstrated below:

$$\mathbf{w}^{new} = \mathbf{w}^{old} - \eta \frac{\partial \Sigma(\mathbf{w})}{\partial \mathbf{w}} \tag{22}$$

Where η is the learning rate. η is set empirically.