Complementary Label

- Collecting true label is expensive, time-consuming. For this reason we make complementary label

- Complementary label are widely used in our daily lives. For example, to identify a language we do not know, we may say "not english".

- K = number of output
  j = class of complementary label
  n = number of data
  i = indexed of x
  n_j = number of samples that has j as the complementary label

$$P_{i\bar{j}} = \begin{cases} P(y=j \mid x=x_i) \Rightarrow \text{ordinary label} \\ P(\bar{y}=j \mid x=x_i) \Rightarrow \text{complementary label} \end{cases}$$

$$= \sum_{k \neq j}^{|y|} P(y=k \mid x),$$

For $|y|=3$, number of output class $=3$

If $y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ then $\bar{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

If $p(y|x) = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$ then $p(\bar{y}|x) = \begin{bmatrix} 0.8 \\ 0.7 \\ 0.5 \end{bmatrix}$

And we can see that

$$P(\bar{y}=j \mid x) = \sum_{k \neq j}^{|y|} P(y=k \mid x),$$

$$P(\bar{y}=2 \mid x) = p(y=1 \mid x) + p(y=3 \mid x)$$

Date

Maximum Entropy loss with complementary label

constraint lagrangian of maximum entropy

$$\max \sum_{ij} -P_{ij} \log P_{ij}$$

s.t (i) $P_{ij} \geq 0$

(ii) $\sum_j P_{ij} = 1 \quad , \forall i = 1, 2, \cdots, n$

If we have $|y| = 3$, so
$P(y=1|x_i) + P(y=2|x_i) + P(y=3|x_i) = 1 \quad , \forall i = 1$

(iii) For each class of $\bar{y}$ (complementary)

$$\frac{1}{n} \sum_i P(\bar{y}=j|x) x_i = \frac{1}{n_j} \sum_{x \in X_j} z(\bar{y}=j|x) x_i \quad , \forall j$$

$\Leftrightarrow$

$$\frac{1}{n} \sum_i \sum_k I(k \neq j) P(y=k|x) x_i = \frac{1}{n_j} \sum_{x \in X_j} z(\bar{y}=j|x) x_i$$

Deff: $P_{ij} = P(\bar{y}=j|x=x_i)$ (because we're talking complement label)

$$z(\bar{y}=j|x=x_i) = \begin{cases} 1 & , y \neq j \\ 0 & y = j \end{cases}$$

$$P(\bar{y}=j|x=x_i) = \sum_{k \neq j}^{|y|} P(y=k|x)$$

If $|y| = 3$ then $P(\bar{y}=3|x) = P(y=2|x) + P(y=1|x)$

Scanned with CamScanner

For simple example of maximum entropy loss, you can see my previous file that I've given two days ago.

- Based on constraint lagrangian that I've given, we can get formula of maxent

We have

$$P_{ij} = p(\bar{y}=j \mid x=x_i) = \frac{\exp\{\sum_k I(k \neq j)\, \theta_j^T x_i\}}{\sum_{j'} (\exp\{\sum_k I(k \neq j')\, \theta_j^T x_i\})}$$

Let $Z = \sum_{j'} (\exp\{\sum_k I(k \neq j')\, \theta_j^T x_i\})$

So, we have

$$P_{ij} = \frac{\exp\{\sum_k I(k \neq j)\theta_j^T x_i\}}{Z_\theta(x)} \quad \ldots \text{(i)}$$

And, we put the P_{ij} and lagrangian, to maximum entropy

$$L(\theta) = \min_\theta \sum_{ij} -P_{ij}\log P_{ij} + \sum_i \alpha_i \left(\sum_j P_{ij} - 1\right)$$

$$+ \sum_j \sum_i \sum_k \theta_j \left[I(k \neq j)\, p(\bar{y}=k \mid x=x_i)\, x_i \right]$$

$$- \sum_j \left[\frac{n}{n_j} \sum_{x \in X_i} \theta_j^T z(\bar{y}=j \mid x_i)\, x_i \right] \quad \ldots \text{(ii)}$$

Consider

$$\sum_{ij} -P_{ij}\log P_{ij} = \sum_{ij}\left[-P_{ij}(-\log Z_\theta(x))\right] - \sum_{ij} P_{ij}\left[\sum_k I(k \neq j)\theta_j^T x_i\right]$$

from (i)

$$= \sum_i \log Z_\theta(x) - \sum_{ijk} P_{ij}\, I(k \neq j)\,\theta_j^T x_i \quad \ldots \text{(iii)}$$

Date

from (ii) and (iii)

$$\mathcal{L}(\theta) = \min_{\theta} \sum_i \log Z_\theta(x_i) - \sum_{ijk} p_{ij}\, I(k \neq j)\, \theta_j^T x_i + \sum_i \sum_j \sum_k \theta_j\, I(k\neq j)\, q(\bar{y}|x)\, x_i$$

$$- \sum_j \left[\frac{n}{n_j} \sum_{x \in X_i} \theta_j\, z(\bar{y}=j \mid x_i)\, x_i \right]$$

$$\mathcal{L}(\theta) = \min_{\theta} \sum_i \log Z_\theta(x_i) - \sum_j \left[\frac{n}{n_j} \sum_{x\in X_i} \theta_j\, z(\bar{y}=j|x_i)\, x_i \right]$$

$$\boxed{= \min_{\theta} \sum_i \log \left[\sum_{j'} \exp\left\{ \sum_k I(k \neq j')\, \theta_j^T x_i \right] - \sum_j \frac{n}{n_j} \sum_{x \in X} \theta_j\, z(\bar{y}=j|x_i)\, x_i}$$

And then, we give first derivation of $\mathcal{L}(\theta)$ over θ_j

$$\frac{\partial L}{\partial \theta_j} = \sum_i \frac{1}{\sum_j \exp\left\{ \sum_k I(k\neq j)\, \theta_j^T x_i \right.} \cdot \sum_k I(k \neq j) . x_i . \exp\left\{ \sum_k I(k\neq j)\, \theta_j \right\}.$$

$$- \frac{n}{n_j} \sum_{x \in X_i} z(\bar{y}=j \mid X = x_i)\, x_i$$

The gradient of loss

$$\boxed{\frac{\partial L}{\partial \theta_j} = \left(\sum_i P(\bar{y}=j \mid x_i) \sum_k I(k\neq j)\, x_i - \frac{n}{n_j} \sum_{x\in X_j} z(\bar{y}=j \mid x)\, x_i \right) = 0}$$