Fairness

[NIPS 2010 : Discriminative Clustering by Regularized Information Maximization (RIM)]

Problem : learn a probabilistic discriminative classifier from an unlabeled dataset

$X = (x_1, \cdots, x_N)$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{iD})^T \in \mathbb{R}^D \xrightarrow{\text{learn}} p(y|\mathbf{x}, W)$

RIM : $F(p(y|x, W); X; \lambda) \Rightarrow$ evaluate the suitability of $p(y|x, W)$

① cluster assumption (decision boundaries $\times$ dense) → confidence level

conditional entropy $\frac{1}{N} \sum_i H\{p(y|x_i, W)\}$

(On unsupervised assumption, it can be reduced by removing decision boundaries)

② class balance (avoid degenerate solutions) → distribution level

empirical label distribution : $\hat{p}(y; W) = \frac{1}{N} \sum_i p(y|x_i, W)$

entropy $H\{\hat{p}(y; W)\}$

Combine ① + ② $I_W\{y; x\} = H\{\hat{p}(y; W)\} - \frac{1}{N} \sum_i H\{p(y|x_i, W)\}$

mutual information

($I_W\{y; x\}$ may be trivially maximized by a conditional model that classifies each data point x_i into its own category y_i)

③ classifer complexity (penalty)

$$F(W; X, \lambda) = I_W\{y; x\} - R(W; \lambda)$$

Learning a conditional distributional distribution for y that preserves information from the data set, subject to a complexity penalty.

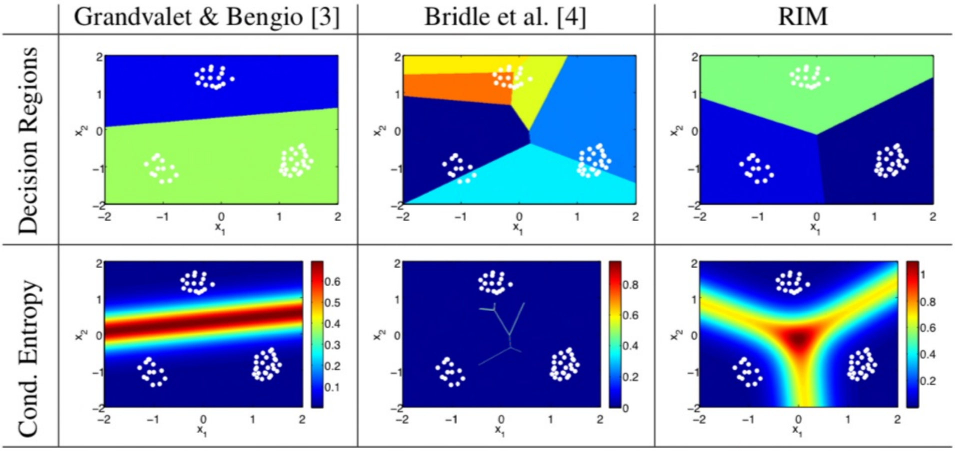

Figure 1: Example unsupervised multilogit regression solutions on a simple dataset with three clusters. The top and bottom rows show the category label $\arg\max_y p(y|\mathbf{x}, \mathbf{W})$ and conditional entropy $H\{p(y|\mathbf{x}, \mathbf{W})\}$ at each point $\mathbf{x}$, respectively. We find that both class balance and regularization terms are necessary to learn unsupervised classifiers suitable for multi-class clustering.

Since $H\{\hat{p}(y;W)\} = \log K - KL\{\hat{p}(y;W)\|U\}$, then

$$F(W;X,\lambda) = -\frac{1}{N}\sum_i H\{p(y|x_i,W)\} - \underbrace{KL\{\hat{p}(y;W)\|U\}}_{\text{Class balance} \to D(y;\gamma)} - R(W;\lambda)$$

Others

$$F(W;X,\lambda) = I_W(x;y) - H\{\hat{p}(y;W)\|D(y;\gamma)\} - R(W;\lambda)$$

In SSL, $S(W;\tau,\lambda) = \underbrace{\tau I_W\{y;x\} - R(W;\lambda)}_{D_U} + \underbrace{\sum_i \log p(y_i|x_i^L, W)}_{D_L}$

[IJCNN 2020: Pseudo-Labeling and Confirmation Bias in Deep SSL]

Two Reg' to improve convergence.

① $R_A = \sum_{c=1}^{C} p_c \log\left(\frac{p_c}{h_c}\right)$ distribution level $D_{KL}(U\|\bar{h}_\theta)$

p_c → prior distribution of c

h_c → mean softmax probability of model for c

② $R_H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} h_\theta^c(x_i)\log(h_\theta^c(x_i))$ confidence level (entropy regularization)

mix up loss

$l = l^* + \lambda_A R_A + \lambda_H R_H$

[ICLR 2023 : FreeMatch]

Self-adaptive fairness (distribution level)

$\tilde{p}_t(c) \longrightarrow$ estimate of the expectation of prediction distribution over D_U.

In RIM, we use $H(\hat{p}(y;W))$ to realize class balance. ($H(\mathbb{E}_u[p_m(y|u)])$ max)

We optimize CE of $\tilde{p}_t$ and $\bar{p} = \mathbb{E}_{\mu B}[p_m(y|\Omega(u_b))]$ as an estimate.

(We expect $D_{KL}(\tilde{p}_t \| \bar{p}) \simeq 0$, that is $H(\tilde{p}_t, \bar{p}) \simeq H(\tilde{p}_t) \simeq H(\bar{p})$)

? The underlying pseudo-label distribution may not be uniform ($\tilde{p}_t, \bar{p} \not\sim U$)

(max $H(\tilde{p}_t, \bar{p})$ $\Longrightarrow$ tend to be uniform distribution)

☆ modulate the fairness objective in a self-adaptive way (normalize)

$$\bar{p} = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}(\max(q_b) \geq \tau_t(\arg\max(q_b))) Q_b$$

$$\bar{h} = \mathrm{Hist}_{\mu B}(\mathbb{1}(\max(q_b) \geq \tau_t(\arg\max(q_b))) \hat{Q}_b)$$

$$\tilde{h}_t = \lambda \tilde{h}_{t-1} + (1-\lambda)\,\mathrm{Hist}_{\mu B}(\hat{q}_b)$$

The self-adaptive fairness (SAF) $\mathcal{L}_f$ at the t-th iteration is :

$$\mathcal{L}_f = -\mathcal{H}\Big(\underbrace{\mathrm{SumNorm}\Big(\frac{\tilde{p}_t}{\tilde{h}_t}\Big)}_{\text{Nearly}}, \underbrace{\mathrm{SumNorm}\Big(\frac{\bar{p}}{\bar{h}}\Big)}_{\text{Uniform}}\Big)$$

Nearly Uniform