A unifying MI view of metic learning cross-entropy us pairwise losses

Notations:

- Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$
- input feat. $\mathcal{X}$
- Embedded features $\mathcal{Z} \in \mathbb{R}^d$
- label space $\mathcal{Y} \in \mathbb{R}^K$
- Encoder $\phi_w : \mathcal{X} \to \mathcal{Z}$
- Soft classifier $f_\theta : \mathcal{Z} \to [0,1]^K$

Information measurement:

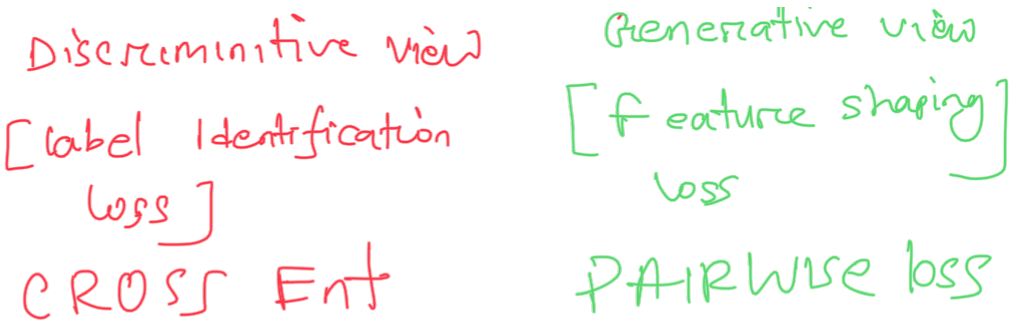
Entropy Y; $H(Y) := \mathbb{E}_{P_Y}[-\log P_Y(y)]$

Conditional Ent. $Y|\hat{Z}$ $H(Y|\hat{Z}) = \mathbb{E}_{P_{Y\hat{Z}}}[-\log P_{Y|\hat{Z}}(Y|\hat{Z})]$

MI between Y & $\hat{Z}$; $I(\hat{Z};Y) = H(Y) - H(Y|\hat{Z})$

Two views of MI

$$I(\hat{Z};Y) = \underbrace{H(Y) - H(Y|\hat{Z})}_{} = \underbrace{H(\hat{Z}) - H(\hat{Z}|Y)}_{}$$



Paige wise loss and generative view of MI

1. Contrastive loss

$$\mathcal{L}_{cont} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j:\, y_j = y_i} D_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j:\, y_j \neq y_i}\left[m - D_{ij}\right]_+^2$$

$\mathcal{L}_{Contrast}$ — Tightness term

$\mathcal{L}_{Contrast}$ — contrastive term

Here,

$$\mathcal{L}_{Contrast} \overset{c}{=} \underbrace{H(\hat{z}, \hat{z}|y)}_{\text{cross entropy}} = H(\hat{z}|y) + D_{KL}(\hat{z}\|\hat{z}|y)$$

$$\geq H(\hat{z}|y)$$

$\mathcal{L}_{Contrast}$ ⇓ valid only when same label

we want them to go to unique point

$H(\hat{z}|y)$ ⇓ go to unique solution if not given other contrastive term.

we can show that $\mathcal{L}_{Contrast}$ is related to $H(\hat{z})$

Finally:

$$\mathcal{L}_{cont} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{j:\, y_i \neq y_j}\left(D_{ij}^2 + 2m D_{ij}\right)}_{\propto H(\hat{z}|y)} - \underbrace{\frac{2m}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}}_{\propto H(\hat{z})} \quad \propto -I(\hat{z};y)$$

Table 2. Several well-known and/or recent DML losses broken into a *tightness* term and a *contrastive* term. Minimizing the cross-entropy corresponds to an approximate bound optimization of PCE.

Loss	Tightness part $\propto \mathcal{H}(\widehat{Z}\|Y)$	Contrastive part $\propto \mathcal{H}(\widehat{Z})$
Center [42]	$\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{z}_i - \boldsymbol{c}_{y_i}\|^2$	$-\frac{1}{n}\sum_{i=1}^{n}\log p_{iy_i}$
Contrast [7]	$\frac{1}{n}\sum_{i=1}^{n}\sum_{j:y_j=y_i} D_{ij}^2$	$\frac{1}{n}\sum_{i=1}^{n}\sum_{j:y_j\neq y_i}[m - D_{ij}]_+^2$
SNCA [43]	$-\frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{j:y_j=y_i}\exp\frac{D_{ij}^{\cos}}{\sigma}\right]$	$\frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{k\neq i}\exp\frac{D_{ik}^{\cos}}{\sigma}\right]$
MS [40]	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\log\left[1+\sum_{j:y_j=y_i} e^{-\alpha(D_{ij}^{\cos}-m)}\right]$	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\beta}\log\left[1+\sum_{j:y_j\neq y_i} e^{\beta(D_{ij}^{\cos}-m)}\right]$
PCE Prop. 1	$-\frac{1}{2\lambda n^2}\sum_{i=1}^{n}\sum_{j:y_j=y_i}\boldsymbol{z}_i^\intercal\boldsymbol{z}_j$	$\frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}\exp\left[\frac{1}{\lambda n}\sum_{j=1}^{n}p_{jk}\boldsymbol{z}_i^\intercal\boldsymbol{z}_j\right]\right] - \frac{1}{2K^2\lambda^2}\sum_{k=1}^{K}\|\boldsymbol{c}_k^s\|^2$

Lemma 1. *Let T_A denote the tightness part of the loss from method A. Assuming that features are ℓ_2-normalized, and that classes are balanced, the following relations between Center [42], Contrastive [7], SNCA [43] and MS [40] losses hold:*

$$T_{SNCA} \overset{c}{\leq} T_{Center} \overset{c}{=} T_{Contrastive} \overset{c}{\leq} T_{MS} \tag{8}$$

Where $\overset{c}{\leq}$ stands for lower than, up to a multiplicative and an additive constant.

Notation :

$\underline{\underline{\leq}}$ means equal upto additive and multiplicative constant.