Appendix

A. Proof

Theorem 1. *The relative efficiency of logistic regression to linear discriminant analysis is*

$$\mathrm{Eff}_p(\zeta,\Delta) = (Q_1 + (p-1)Q_2)/(Q_3 + (p-1)Q_4),$$

where $Q_2 = 1 + \pi_0\pi_1\Delta^2$, $Q_4 = \frac{1}{A_0}$ *and*

$$\begin{aligned} Q_1 &= \begin{pmatrix} 1 & \frac{\zeta}{\Delta} \end{pmatrix} \begin{bmatrix} 1+\frac{\Delta^2}{4} & (\pi_0-\pi_1)\frac{\Delta}{2} \\ (\pi_0-\pi_1)\frac{\Delta}{2} & 1+2\pi_0\pi_1\Delta^2 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{\zeta}{\Delta} \end{bmatrix}, \\ Q_3 &= \begin{pmatrix} 1 & \frac{\zeta}{\Delta} \end{pmatrix} \frac{1}{A_0A_2 - A_1^2} \begin{bmatrix} A_2 & A_1 \\ A_1 & A_0 \end{bmatrix} \begin{bmatrix} 1 \\ \frac{\zeta}{\Delta} \end{bmatrix}. \end{aligned}$$

Proof. The proof can be found in Efron (1975). □

Theorem 2. *If* $\pi_i = \pi_j$*, the expectation of the distance* $d_{(i,j)}$ *is a function of the Mahalanobis distance* $\Delta_{i,j}$*:*

$$\mathbb{E}[d_{(i,j)}] = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\Delta_{i,j}^2}{8}\right) + \frac{1}{2}\Delta_{i,j}\left[1 - 2\Phi(-\frac{\Delta_{i,j}}{2})\right],$$

where $\Phi(\cdot)$ *is the normal cumulative distribution function.*

Proof. Since $d_{(i,j)}$ is the distance of $x_{(i)}$ to the decision boundary between class i and j decided by the Fisher's linear discriminant function $\lambda_{i,j}(x) = \beta_{i,j} + \alpha_{i,j}^\top x = 0$, where

$$\begin{aligned} \beta_{i,j} &= \log(\pi_i/\pi_j) + \frac{1}{2}(\|\mu_j\|_2^2 - \|\mu_i\|_2^2), \\ \alpha_{i,j}^\top &= (\mu_i - \mu_j)^\top. \end{aligned}$$

We have

$$d_{(i,j)} = \frac{|\beta_{i,j} + \alpha_{i,j}^\top x_{(i)}|}{\|\alpha_{i,j}\|_2} = \frac{|\beta_{i,j} + \alpha_{i,j}^\top x_{(i)}|}{\Delta_{i,j}}.$$

Because $x_{(i)}$ is sampled from the conditional Gaussian distribution of class i, there is

$$x_{(i)} \sim \mathcal{N}(\mu_i, I).$$

Let $H_{i,j} = \beta_{i,j} + \alpha_{i,j}^\top x_{(i)}$, and $\zeta_{i,j} = \log(\pi_i/\pi_j)$, then according to the property of Gaussian distribution we can know that

$$H_{i,j} \sim \mathcal{N}(\mu'_{i,j}, \sigma_{i,j}^2),$$

where

$$\begin{aligned} \mu'_{i,j} &= \beta_{i,j} + \alpha_{i,j}^\top \mu_i \\ &= \zeta_{i,j} + \frac{1}{2}(\|\mu_j\|_2^2 - \|\mu_i\|_2^2) + (\mu_i - \mu_j)^\top \mu_i \\ &= \zeta_{i,j} + \frac{1}{2}\|\mu_i - \mu_j\|_2^2 \\ &= \zeta_{i,j} + \frac{1}{2}\Delta_{i,j}^2, \end{aligned}$$

and $\sigma_{i,j}^2 = \alpha_{i,j}^\top I \alpha_{i,j} = \Delta_{i,j}^2$. Thus $|H_{i,j}|$ distributes as a *Folded Gaussian (Normal) Distribution*. From the property of folded Gaussian distribution we know that

$$\mathbb{E}[|H_{i,j}|] = \sqrt{\frac{2}{\pi}}\sigma_{i,j}\exp(-\frac{{\mu'_{i,j}}^2}{2{\sigma_{i,j}}^2}) + \mu'_{i,j}[1 - 2\Phi(-\frac{\mu'_{i,j}}{\sigma_{i,j}})].$$

For notation clarity, we let

$$\alpha_{i,j} = \frac{\mu'_{i,j}}{\sigma_{i,j}} = \frac{1}{2}\Delta_{i,j} + \zeta_{i,j}/\Delta_{i,j}.$$

Since $\mathbb{E}[d_{(i,j)}] = \mathbb{E}[|H_{i,j}|]/\Delta_{i,j}$, we have

$$\mathbb{E}[d_{(i,j)}] = \sqrt{\frac{2}{\pi}}\exp(-\frac{\alpha_{i,j}^2}{2}) + \alpha_{i,j}[1 - 2\Phi(-\alpha_{i,j})].$$

The derivative of $\mathbb{E}[d_{(i,j)}]$ to $\Delta_{i,j}$ is

$$\begin{aligned} \frac{\partial\mathbb{E}[d_{(i,j)}]}{\partial\Delta_{i,j}} &= \frac{\partial\mathbb{E}[d_{(i,j)}]}{\partial\alpha_{i,j}} \cdot \frac{\partial\alpha_{i,j}}{\partial\Delta_{i,j}} \\ &= [1 - 2\Phi(-\frac{1}{2}\Delta_{i,j} - \frac{\zeta_{i,j}}{\Delta_{i,j}})](\frac{1}{2} - \frac{\zeta_{i,j}}{\Delta_{i,j}^2}) \\ &= [1 - 2\Phi(-\Delta_{i,j}(\frac{1}{2} - \frac{\zeta_{i,j}}{\Delta_{i,j}^2}))](\frac{1}{2} - \frac{\zeta_{i,j}}{\Delta_{i,j}^2}) \\ &\geq 0, \end{aligned}$$

where the Mahalanobis distance $\Delta_{i,j}$ is non-negative. Specially, when $\pi_i = \pi_j$, i.e., $\zeta_{i,j} = 0$, there is

$$\mathbb{E}[d_{(i,j)}] = \sqrt{\frac{2}{\pi}} \exp(-\frac{\Delta_{i,j}^2}{8}) + \frac{1}{2}\Delta_{i,j}[1 - 2\Phi(-\frac{\Delta_{i,j}}{2})].$$

□

Theorem 3. *Assume that* $\sum_{i=1}^{L} \mu_i = 0$ *and* $\|\mu\|_2^2 = C$. *Then there has*

$$\overline{\mathrm{RB}} \leq \sqrt{\frac{LC}{2(L-1)}}.$$

The equality holds if and only if

$$\mu_i^\top \mu_j = \begin{cases} C, & i = j, \\ C/(1-L), & i \neq j, \end{cases}$$

where $i, j \in [L]$ *and* $\mu_i, \mu_j \in \mu$.

Proof. According to the definition of $\overline{\mathrm{RB}}$, we have

$$\begin{aligned}
\overline{\mathrm{RB}} &= \frac{1}{2} \min_{i,j \in [L]} \Delta_{i,j} \\
&= \frac{1}{2} \sqrt{\min_{i,j \in [L]} \Delta_{i,j}^2} \\
&\leq \frac{1}{2} \sqrt{\frac{1}{L(L-1)} \sum_{i \neq j} \Delta_{i,j}^2} \\
&= \frac{1}{2} \sqrt{\frac{1}{L(L-1)} \sum_{i \neq j} (\|\mu_i\|_2^2 + \|\mu_j\|_2^2 - 2\mu_i^\top \mu_j)} \\
&= \frac{1}{2} \sqrt{\frac{2}{L} \sum_{i \in [L]} \|\mu_i\|_2^2 - \frac{1}{L(L-1)} \sum_{i \neq j} 2\mu_i^\top \mu_j} \\
&= \frac{1}{2} \sqrt{\frac{2}{L-1} \sum_{i \in [L]} \|\mu_i\|_2^2 - \frac{1}{L(L-1)} (\sum_{i \in [L]} \mu_i)^2},
\end{aligned}$$

Since $\sum_{i=1}^{L} \mu_i = 0$ and $\|\mu\|_2^2 = C$, we further have

$$\begin{aligned}
\overline{\mathrm{RB}} &\leq \sqrt{\frac{1}{2(L-1)} \sum_{i \in [L]} \|\mu_i\|_2^2} \\
&\leq \sqrt{\frac{LC}{2(L-1)}}.
\end{aligned}$$

Note that the final equality holds if and only if all the equalities hold, i.e., there are

$$\|\mu_i\|_2^2 = C, \forall i \in [L],$$

and

$$\Delta_{i,j} = \text{constant}, \forall i \neq j.$$

Thus we can easily derive that the final equality holds if and only if

$$\mu_i^\top \mu_j = \begin{cases} C, & i = j, \\ C/(1-L), & i \neq j, \end{cases}$$

where $i, j \in [L]$ and $\mu_i, \mu_j \in \mu$.

□

B. More Discussions and Details

In this section we discuss more on the loss function of the MM-LDA network, and the choice of the square norm of C of MMD. Besides, we provide technical details of the adversarial training methods we use in our experiments.

B.1. The Loss Function of the MM-LDA Network

Considering that the network with parameters θ induces a joint distribution on the latent feature z and the label y as $Q_\theta(z, y)$. We denote the MMD as $P(z, y)$, $\mathbb{H}(P, Q)$ as the cross-entropy for the distributions P and Q. Then the training objective could be designed as

$$\begin{aligned}
\mathbb{H}(Q_\theta, P) &= \mathbb{E}_{(z,y)\sim Q_\theta}[-\log P(y|z) - \log P(z)] \\
&= \mathbb{E}_{(z,y)\sim Q_\theta}[-\log P(y|z)] + \mathbb{E}_{z\sim Q_\theta^{'}}[-\log P(z)].
\end{aligned}$$

Here $Q_\theta^{'}$ is the marginal distribution of Q_θ for z. Since we are focusing on classification tasks, we assume for tractability that the marginal distribution $Q_\theta^{'}(z)$ is consistent with it of the MMD, i.e., $P(z)$. Therefore, minimizing $\mathbb{H}(Q_\theta, P)$ equals to minimizing $\mathbb{E}_{(z,y)\sim Q_\theta}[-\log P(y|z)]$, which further leads to the loss function $\mathcal{L}_{\mathrm{MM}}$ under the Monte Carlo approximation. In practice, the gap between $Q_\theta^{'}(z)$ and $P(z)$ would not influence the performance, as shown in our experiment results.

In order to better gather the latent feature vectors to their corresponding conditional Gaussian distributions, the loss function of the MM-LDA network should be

$$\mathcal{L}_{\mathrm{MM}} = -F_{\mathrm{MM}}(\mu_y^*)^\top \log F_{\mathrm{MM}}(x),$$

where $F_{\mathrm{MM}}(\mu_y^*)_k = P(y = k|z = \mu_y^*), k \in [L]$. $F_{\mathrm{MM}}(\mu_y^*)$ is the prediction vector on the mean vector μ_y^* in MMD. We further have

$$\begin{aligned}F_{\text{MM}}(\mu_y^*)_y &= \frac{\exp(2{\mu_y^*}^\top \mu_y^*)}{\sum_{i\in[L]}\exp(2{\mu_y^*}^\top \mu_i^*)}\\ &= \frac{1}{1+\sum_{i\neq y}\exp[2{\mu_y^*}^\top(\mu_i^*-\mu_y^*)]}\\ &= \frac{1}{1+(L-1)\exp(-\frac{2LC}{L-1})}.\end{aligned}$$

Thus the L_∞-distance between the one-hot label vector 1_y and $F_{\text{MM}}(\mu_y^*)$ is

$$\begin{aligned}\|1_y - F_{\text{MM}}(\mu_y^*)\|_\infty &\leq \left|1 - F_{\text{MM}}(\mu_y^*)_y\right|\\ &= \frac{1}{1+\frac{1}{L-1}\exp(\frac{2LC}{L-1})}.\end{aligned}$$

It is easy to see that the gap between 1_y and $F_{\text{MM}}(\mu_y^*)$ rapidly decreases w.r.t C. For instance, when $C > 10$, $L = 10$, we numerically have $\|1_y - F_{\text{MM}}(\mu_y^*)\|_\infty \leq 10^{-8}$. Therefore we use the one-hot label 1_y in the loss function because of its simplicity.

B.2. Technical Details of Adversarial Training

Our experiments are done on NVIDIA Tesla P100 GPUs. The number of the adversarial fine-tuning steps on MNIST is 10,000 and on CIFAR-10 is 30,000. We apply a constant learning rate of 0.01 on both datasets. The mixing ratio of normal examples and adversarial examples is 1:1. Averagely the time cost to craft an adversarial example using FGSM, BIM and ILCM is less than 0.1 seconds, while using JSMA is around 5 seconds. This makes adversarial training on JSMA computationally expensive.

References

Efron, Bradley. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.