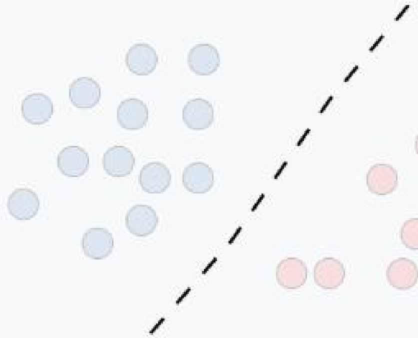
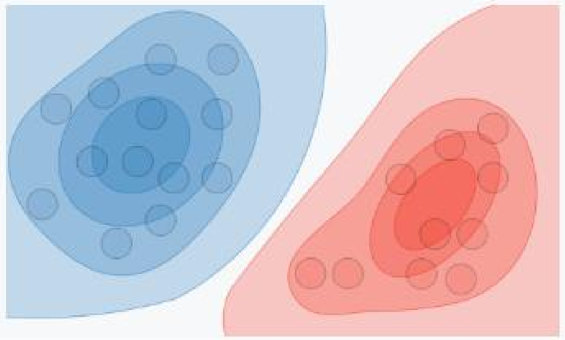
Real DNF

Dong Wang

2020/05/09

Discriminative model and generative model

- A model that focuses on the classification boundary is a discriminative model.
- A model that focuses on describing the class conditional is a generative model.

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-a-discriminative-algorithm/879591#879591

Sometimes, discriminative model is better

- If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

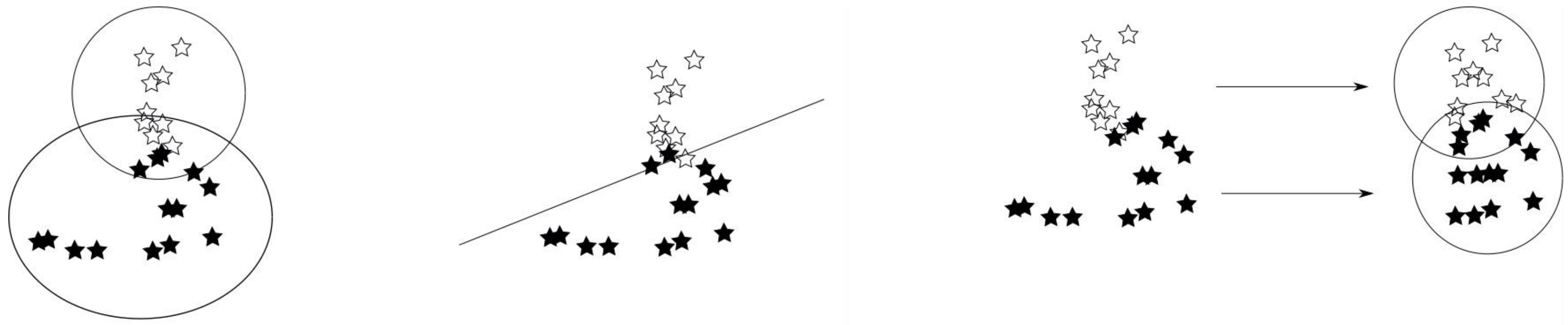
V.N. Vapnik, Statistical Learning Theory. New York: John Wiley & Sons, 1998

Put it in another word….

- If we don’t have enough knowledge to design a generative model, then design a discriminative model.
- Partly contribute to the success of deep learning
- However if we have enough knowledge, then design a generative model will have numerous benefits: generalization, adaptation, visualization, understanding, explanation….

Deep generative model with a weak prior and data learning

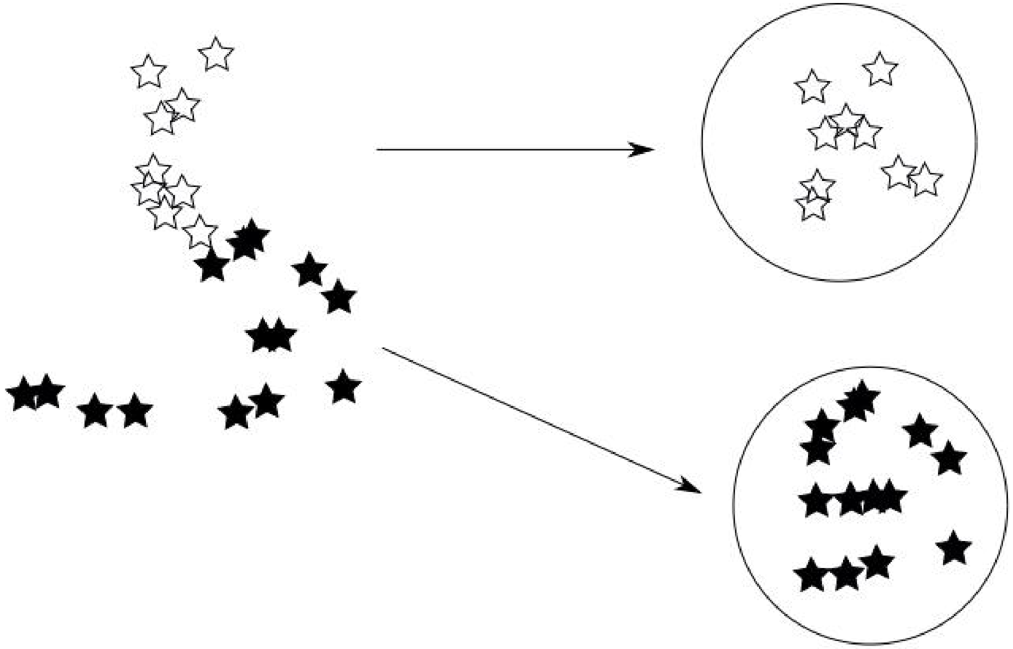
- Design a generative model with a very weak assumption, and make the form flexible enough, and let the data to materialize the model.
- That is the deep generative model.



Generative model Discriminative model Deep Generative model

Discriminative training for deep generative model

- We hope more discriminant power, but keep the data in a good probabilistic form
- A ‘global’ discrimination

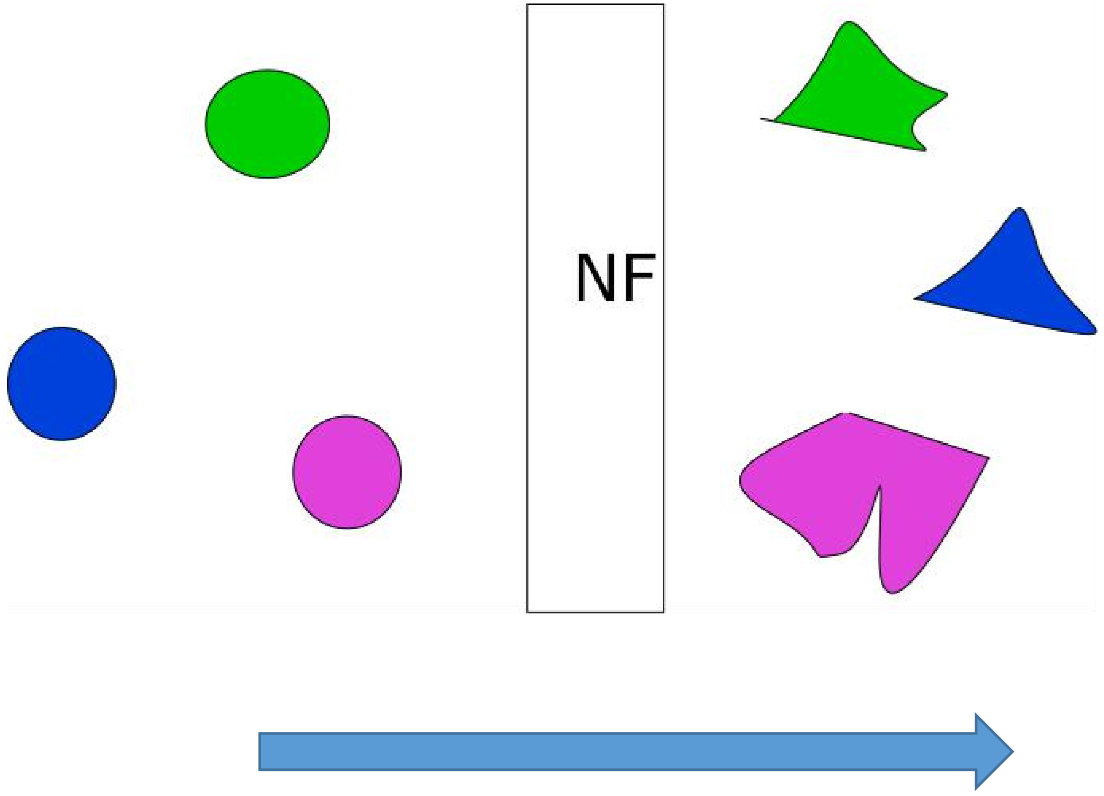


Discriminative Deep Generative model

Task formulation: Maximum likelihood training

$$p(x) = p_{c_x}(x) J_x$$

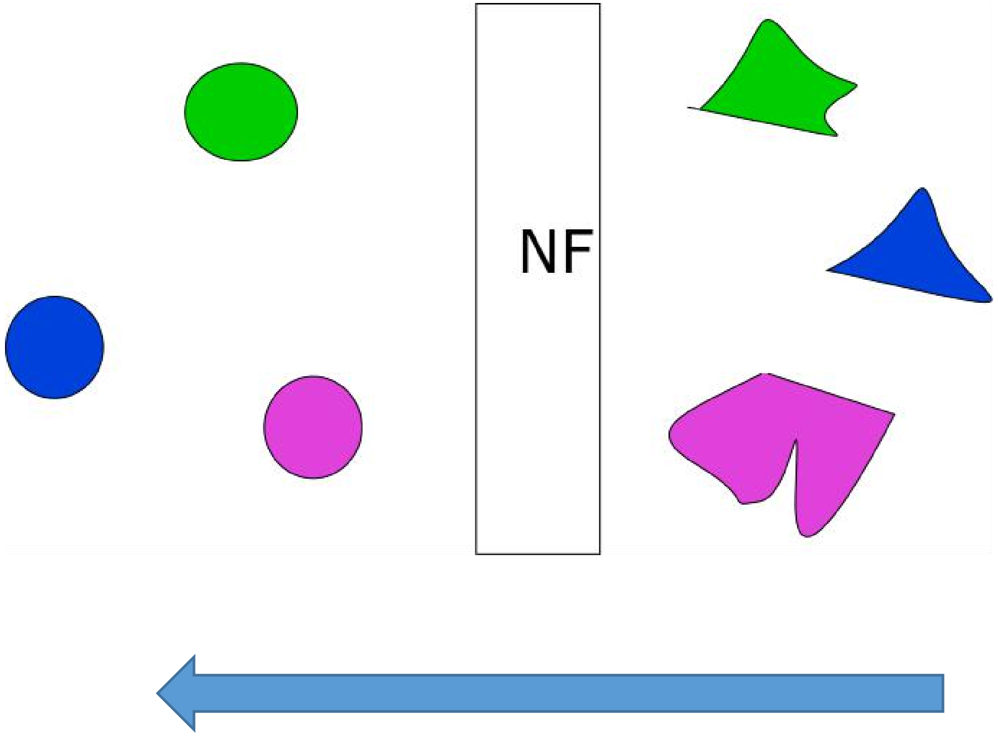
$$L = \prod_x p_{c_x}(x) J_x$$



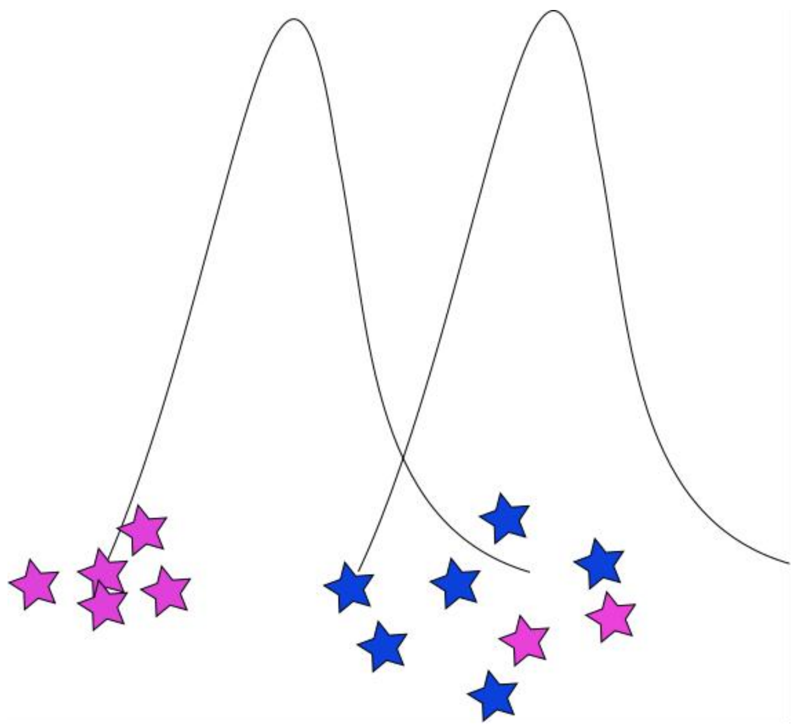
Discriminative training

$$p(c_x|x) = \frac{p_{c_x}(x)p(c_x)}{\sum_{c'} p_{c'}(x)p(c')} = \frac{p_{c_x}(z)p(c_x)}{\sum_{c'} p_{c'}(z)p(c')}$$

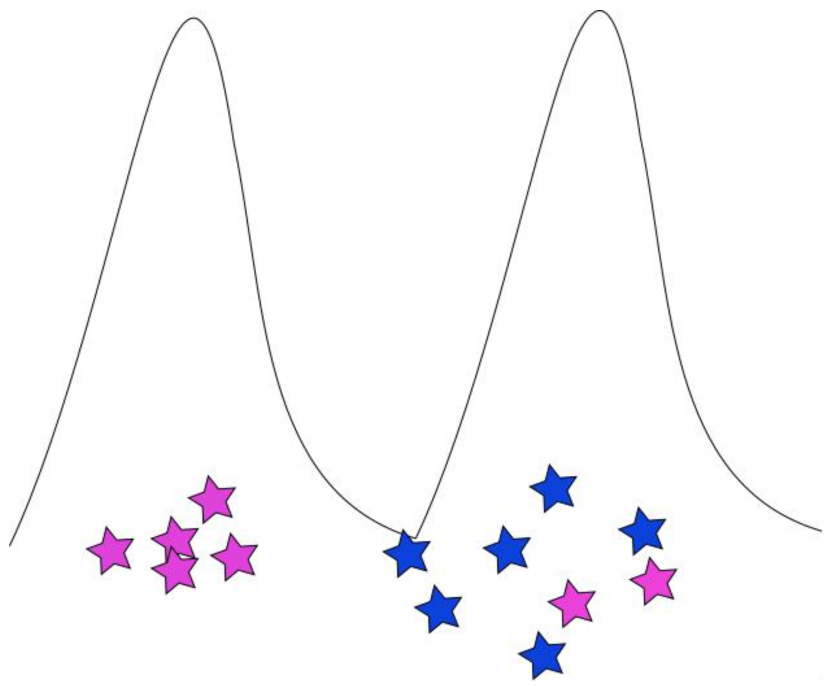
$$L = \prod_x p(c_x|x)$$



Why DT works?



ML training

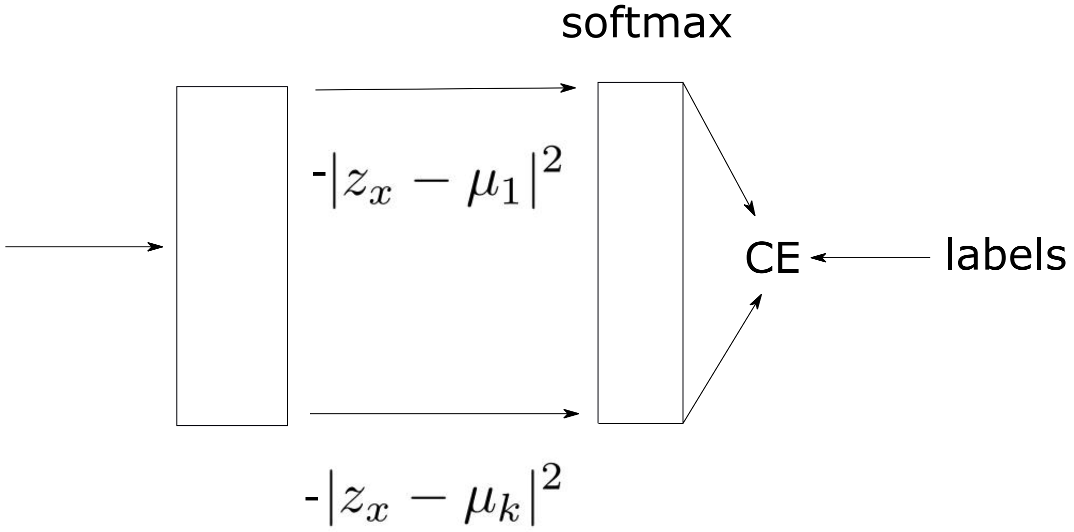


Discriminative training

Linked to cross entropy

$$p(c_x|x) \propto \frac{e^{-|z_x - \mu_{x_c}|^2}}{\sum_{c'} e^{-|z_x - \mu_{c'}|^2}} = softmax(-|z_x - \mu_c|^2)$$

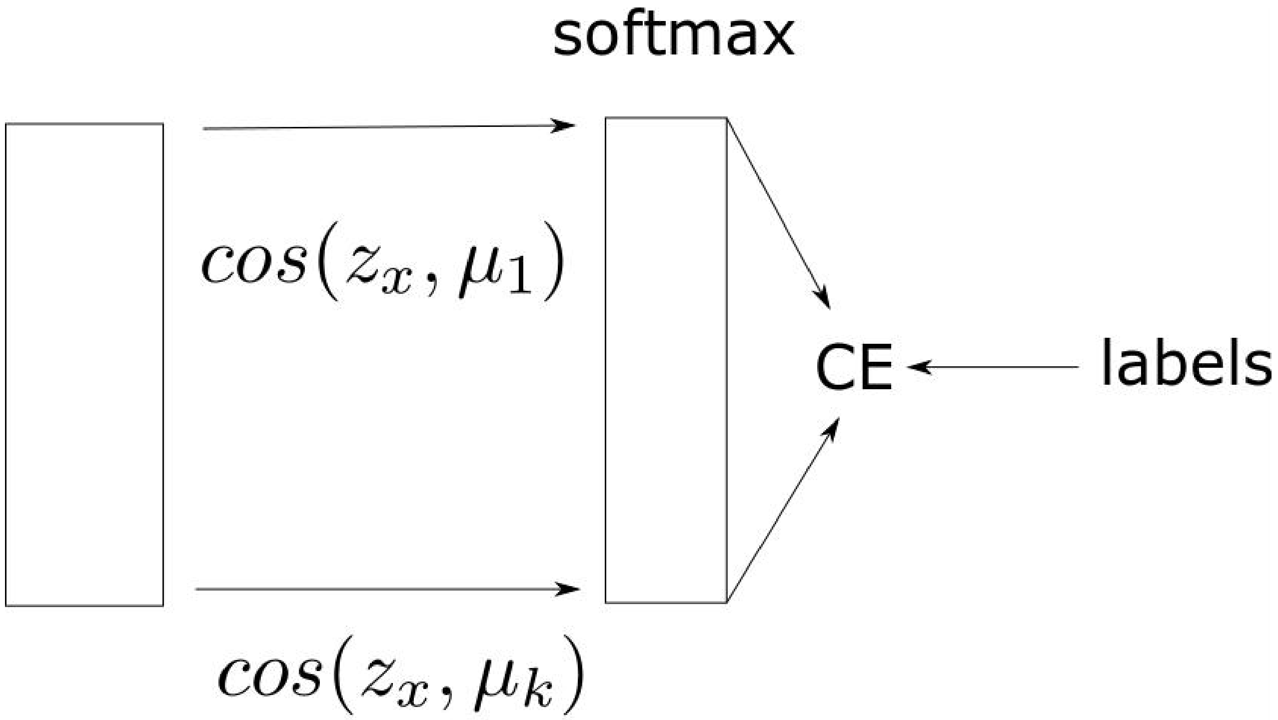
$$L = \sum_x \ln p(c_x|x) = \sum_x CE(\delta(c_x), p(c_x|x))$$

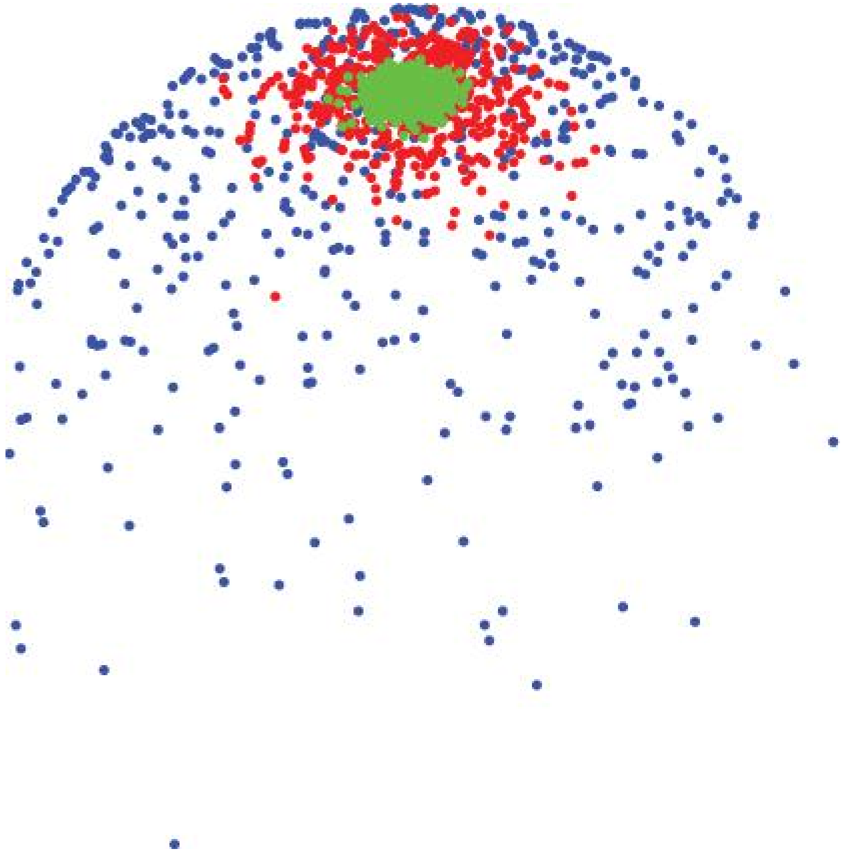


Compared to angular softmax

$$p(c_x|x) \propto \frac{e^{z_x^T \mu_{x_c}}}{\sum_{c'} e^{z_x^T \mu_{c'}}} = softmax(z_x^T \mu_c) \quad s.t. \ ||z_x|| = 1 \quad ||\mu_i|| = 1$$

$$L = \sum_x \ln p(c_x|x) = \sum_x CE(\delta(c_x), p(c_x|x))$$





z is assumed to be VMF distributed.

Remarks

- All the above derivation are based on the following assumption: p(x) = p(z)J(x). It requires a frame-level invertible function.
- X-vector training will not meet this request, so it is something mixed up.
- Recall google's paper that 'defines' the logit as the likelihood, we 'derive' that the logit is a likelihood, if some conditions are met. Specifically, we show that the likelihood will be based on VMF if x and w is regularized to 1, and the final layer operation is wx. In contrast, if the final layer operation is |x-w|^2, then the likelihood is based on Gaussian. Exact likelihood should consider the Jacobian.
- We hilight the form |x-w|^2 is a radio basis function with w as the mean of the RBFs. It does not reduce the expressive power of the entire neural net. This means that the net is not weak than any other regular NN, if the model is deep enough.

A simple test on x-vectors[cvss764]

	TR-cos(EER/IDR)	TR-NL (EER/IDR)	CN-cos(EER/IDR)	CN-NL(EER/IDR)
Init	0.07/0.766	0.01833/0.9525	0.1715/0.468	0.3225/0.2675
It 0	0.050833/0.88000	0.015833/0.953333	0.170000/0.53750	0.280000/0.290000
It 5	0.043333/0.88500	0.013333/0.957500	0.173000/0.52050	0.272500/0.290000
It 10	0.048333/0.89000	0.013333/0.960833	0.172000/0.51700	0.272500/0.290000
It 15	0.045000/0.90250	0.013333/0.967500	0.173500/0.51050	0.277500/0.262500
It 20	0.042500/0.91250	0.013333/0.968333	0.173000/0.50100	0.282500/0.272500
It 25	0.043333/0.91166	0.012500/0.969167	0.177500/0.50100	0.287500/0.255000
It 30	0.043333/0.91333	0.012500/0.970833	0.173500/0.50050	0.292500/0.250000
It 35	0.040000/0.92083	0.011667/0.970833	0.177000/0.50150	0.295000/0.240000
It 40	0.037500/0.92416	0.011667/0.975000	0.174500/0.49100	0.292500/0.242500
It 45	0.037500/0.92833	0.010833/0.976667	0.176000/0.49700	0.295000/0.242500
It 50	0.037500/0.92833	0.010833/0.976667	0.176000/0.49700	0.295000/0.242500
It 55	0.034167/0.93583	0.010000/0.979167	0.178000/0.49300	0.297500/0.255000
It 60	0.035000/0.93750	0.010833/0.977500	0.178500/0.49100	0.302500/0.250000

ML training

	TR-cos(EER/IDR)	TR-NL (EER/IDR)	CN-cos(EER/IDR)	CN-NL(EER/IDR)
Init	0.07/0.766	0.01833/0.9525	0.1715/0.468	0.3225/0.2675
It 0	0.022500/0.93416	0.005833/0.988333	0.177500/0.53500	0.280000/0.327500
It 5	0.014167/0.97750	0.000833/0.995833	0.193000/0.49950	0.260000/0.315000
It 10	0.008333/0.99416	0.000833/0.999167	0.199500/0.47500	0.245000/0.290000
It 15	0.005833/0.99750	0.000000/1.000000	0.203000/0.46350	0.245000/0.267500
It 20	0.004167/0.99750	0.000000/1.000000	0.207500/0.44850	0.232500/0.235000
It 25	0.003333/0.99833	0.000000/1.000000	0.209500/0.43600	0.237500/0.247500
It 30	0.002500/0.99833	0.000000/1.000000	0.213000/0.42850	0.245000/0.235000
It 35	0.002500/0.99833	0.000000/1.000000	0.215000/0.41700	0.242500/0.227500
It 40	0.002500/0.99833	0.000000/1.000000	0.217000/0.40600	0.237500/0.225000
It 45	0.002500/1.00000	0.000000/1.000000	0.220500/0.39650	0.237500/0.215000
It 50	0.002500/1.00000	0.000000/1.000000	0.220500/0.39650	0.237500/0.215000
It 55	0.002500/0.99916	0.000000/1.000000	0.221000/0.38750	0.235000/0.220000
It 60	0.002500/1.00000	0.000000/1.000000	0.224000/0.37300	0.230000/0.182500

Discriminative training

Compared to regular CE

	TR-cos(EER/IDR)	TR-NL (EER/IDR)	CN-cos(EER/IDR)	CN-NL(EER/IDR)
Init	0.07/0.766	0.01833/0.9525	0.1715/0.468	0.3225/0.2675
It 0	0.068333/0.77750	0.021667/0.946667	0.172000/0.47100	0.300000/0.285000
It 5	0.066667/0.78916	0.020833/0.946667	0.172000/0.47250	0.290000/0.280000
It 10	0.065833/0.78916	0.020000/0.947500	0.171000/0.47500	0.292500/0.275000
It 15	0.064167/0.79000	0.019167/0.946667	0.171000/0.47900	0.295000/0.277500
It 20	0.064167/0.79166	0.020000/0.947500	0.172000/0.47950	0.295000/0.277500
It 25	0.063333/0.79333	0.020000/0.947500	0.172500/0.47950	0.292500/0.280000
It 30	0.063333/0.79416	0.020000/0.948333	0.173000/0.48000	0.295000/0.280000
It 35	0.063333/0.79416	0.019167/0.950833	0.173500/0.48250	0.292500/0.275000
It 40	0.062500/0.79583	0.018333/0.951667	0.173000/0.48350	0.297500/0.272500
It 45	0.065000/0.76500	0.019167/0.935000	0.166500/0.47050	0.272500/0.272500
It 50	0.065000/0.76500	0.019167/0.935000	0.166500/0.47050	0.272500/0.272500
It 55	0.062500/0.76833	0.018333/0.936667	0.165500/0.47450	0.272500/0.282500
It 60	0.060833/0.77250	0.019167/0.939167	0.164000/0.48200	0.280000/0.272500

Extended to NDA

Gassian and Computable

$$p(c_x|x) = \frac{\int p(x|\mu)p(\mu|D_{c_x})\mathrm{d}\mu \;\; \mathrm{p}(\mathrm{c_x})}{\sum_{c'} \int p(x|\mu)p(\mu|D_{c'})\mathrm{d}\mu \;\; \mathrm{p}(\mathrm{c}')}$$

Discriminative training In PlDA

$$
\begin{aligned}
s &= \log \frac{p(\phi_1, \phi_2 | \mathcal{H}_s)}{p(\phi_1, \phi_2 | \mathcal{H}_d)} \\
&= \log \frac{\int p(\phi_1|\mathbf{y}) p(\phi_2|\mathbf{y}) p(\mathbf{y}) d\mathbf{y}}{p(\phi_1) p(\phi_2)}
\end{aligned}
$$

$$
\begin{aligned}
s &= \mathbf{w}^T \varphi(\phi_1, \phi_2) \\
&= \begin{bmatrix} \mathrm{vec}(\mathbf{\Lambda}) \\ \mathrm{vec}(\mathbf{\Gamma}) \\ \mathbf{c} \\ k \end{bmatrix}^T \begin{bmatrix} \mathrm{vec}(\phi_1 \phi_2^T + \phi_2 \phi_1^T) \\ \mathrm{vec}(\phi_1 \phi_1^T + \phi_2 \phi_2^T) \\ \phi_1 + \phi_2 \\ 1 \end{bmatrix}.
\end{aligned}
$$

Lost the probabilistic interpretation!!

Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification.
https://ieeexplore.ieee.org/abstract/document/5947437
Constrained discriminative PLDA training for speaker verification
https://ieeexplore.ieee.org/abstract/document/6853882

Conclusion

- Discriminative training for deep generative model opens a door. It is a real discriminative NF.
- It is not much different from the regular CE, however it offers a way to keep the probabilistic assumption.
- The discriminative training paves the way for a more explainable training approach for close-set tasks, e.g., ASR. It paves the way for multi-conditional training and adaptation.