Deep Generative Models for Discriminative Tasks

Dong Wang

2020/02/10

- All things considered here are for classification

MAP criterion for classification

- What is the OPTIMAL decision for a classification task, given x?
- Take a binary classification task as an example, and target for the minimum decision loss:

dec	A	B
loss	$loss_{B->A}\ p(B\|x)$	$loss_{A->B}\ p(A\|x)$

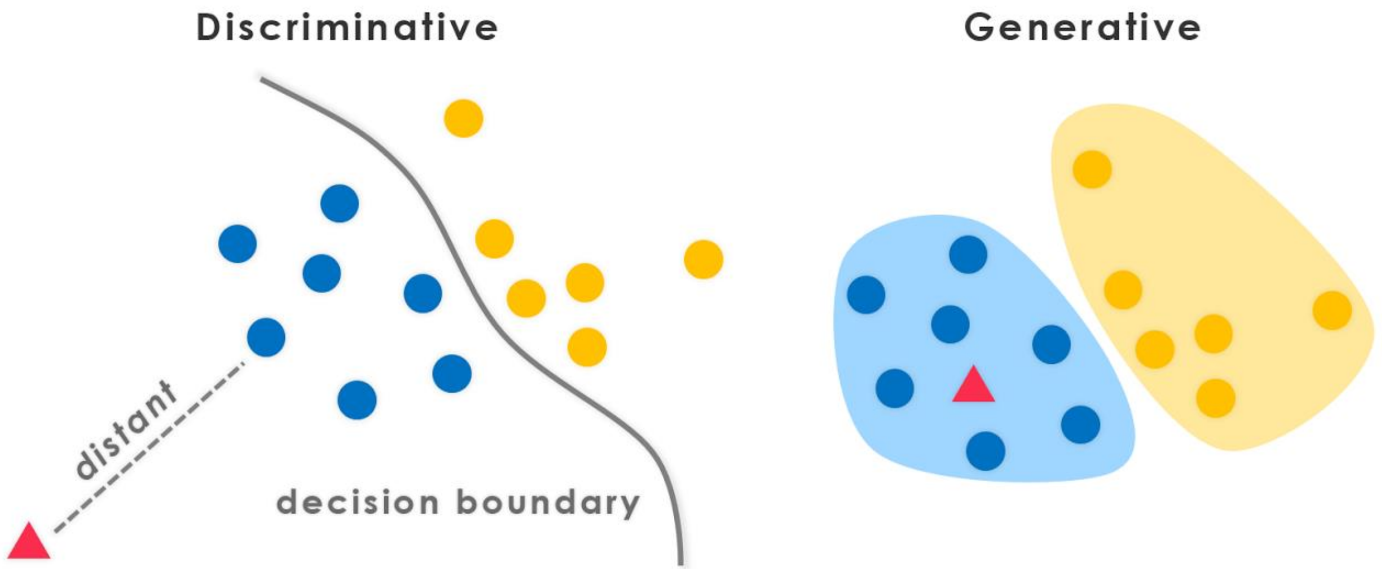
- When the two loss are the same, decision should be based on the posterior $p(A|x)$ or $p(B|x)$.

Modeling for MAP

- For any classification task, MAP is theoretical optimal.
- Therefore all the important is $p(c|x)$
- Two ways to compute $p(c|x)$:

  - Model $p(c|x)$ directly, the discriminative model
  - Model $p(c,x)$ or $p(x|c)$ and $p(c)$, the generative model

Two types of models



$P(c|x) \leftrightarrow c_x$ $P(x|c) \leftrightarrow x_c$

https://datawarrior.wordpress.com/2016/05/08/generative-discriminative-pairs/

Some typical models

- Generative models:
  - HMM, GMM, other Bayesian models
  - RBM, CRF, other MRF models
- Discriminative models
  - Logistic regression, DNN
  - SVM, DT, and other non-probabilistic models

Table 1. High-level comparisons between deep neural networks, a most popular form of deep discriminative models (mid column), and deep generative models (right column), in terms of 15 attributes (left column)

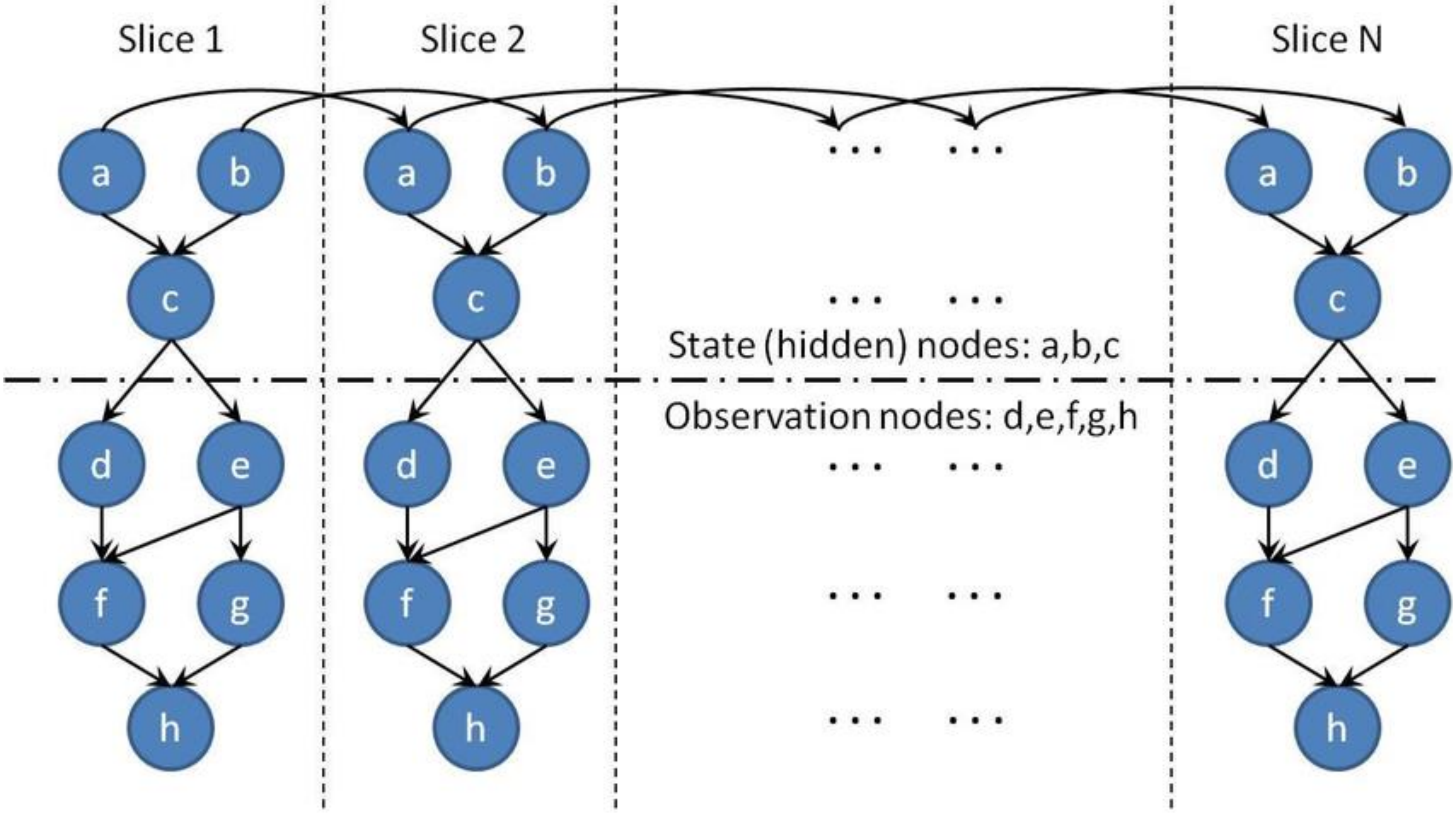
	Deep Neural Networks	Deep Generative Models
Structure	Graphical; info flow: bottom-up	Graphical; info flow: top-down
Domain knowledge	Hard	Easy
Semi/unsupervised	Harder	Easier
Interpretation	Harder	Easy (generative "story")
Representation	Distributed	Local or Distributed
Inference/decode	Easy	Harder (but note recent progress in Section 5.2)
Scalability/compute	Easier (regular computes/GPU)	Harder (but note recent progress)
Incorp. uncertainty	Hard	Easy
Empirical goal	Classification, feature learning, etc.	Classification (via Bayes rule), latent variable inference, etc.
Terminology	Neurons, activation/gate functions, weights, etc.	Random variables, stochastic "neurons", potential function, parameters, etc.
Learning algorithm	Almost a single, unchallenged algorithm — Backprop	A major focus of open research, many algorithms, & more to come
Evaluation	On a black-box score — end performance	On almost every intermediate quantity
Implementation	Hard, but increasingly easier	Standardized methods exist, but some tricks and insights needed
Experiments	Massive, real data	Modest, often simulated data
Parameterization	Dense matrices	Sparse (often); Conditional PDFs

Li Deng, Navdeep Jaitly**, Deep Discriminative and Generative Models for Speech Pattern Recognition,** in *Handbook of Pattern Recognition and Computer Vision (Ed. C.H. Chen), 2015.*

Conventional deep generative models

- Using complex hierarchical structure and simple conditionals to reach complex joint distributions.
- Semantic rich.
- Conjugate priors are often required.
- MCMC or variational methods for training and inference.

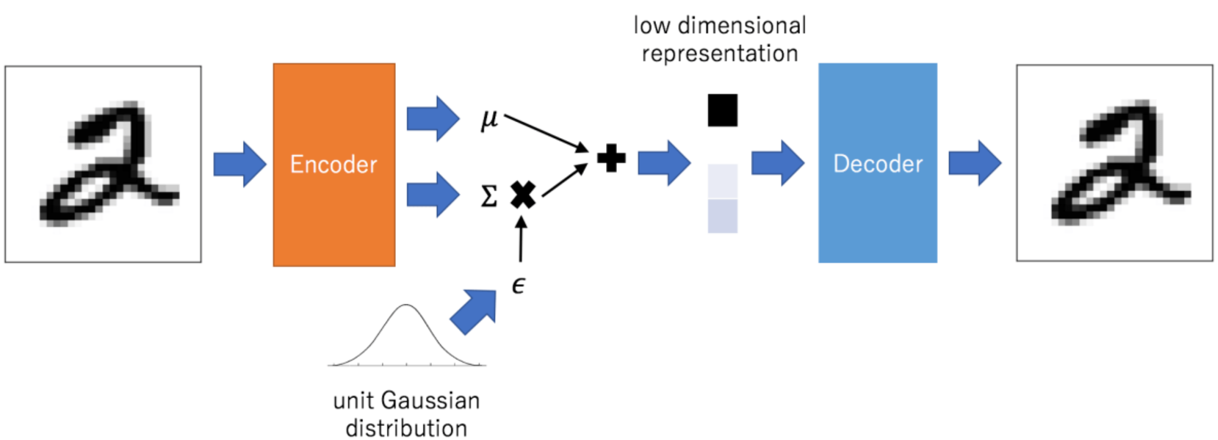
An example



Ke et al., A Review on Video-Based Human Activity Recognition.https://www.researchgate.net/figure/An-example-of-a-dynamic-Bayesian-network-DBN-is-unrolled-in-time-axis-with-state_fig14_285197344

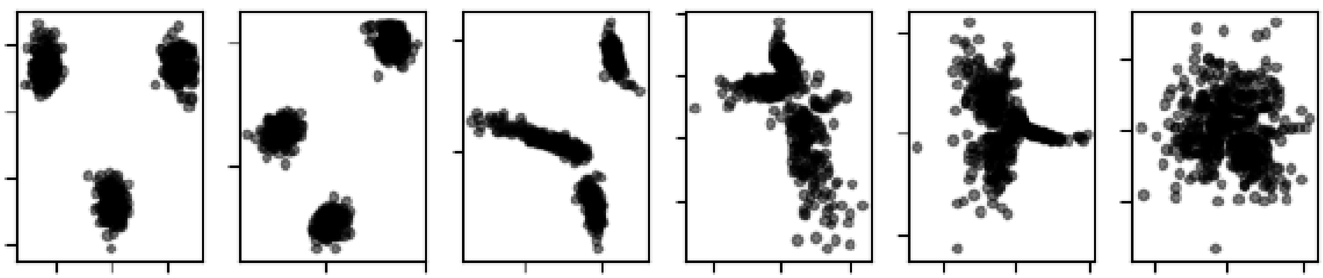
Neural deep generative models

- GAN, VAE and NF, some energy model
- Use complex feature mapping to disentangle the correlation among dimensions, and simplify the distribution form.
- Human designed conditions is replaced by conditional variables collected by a black box.



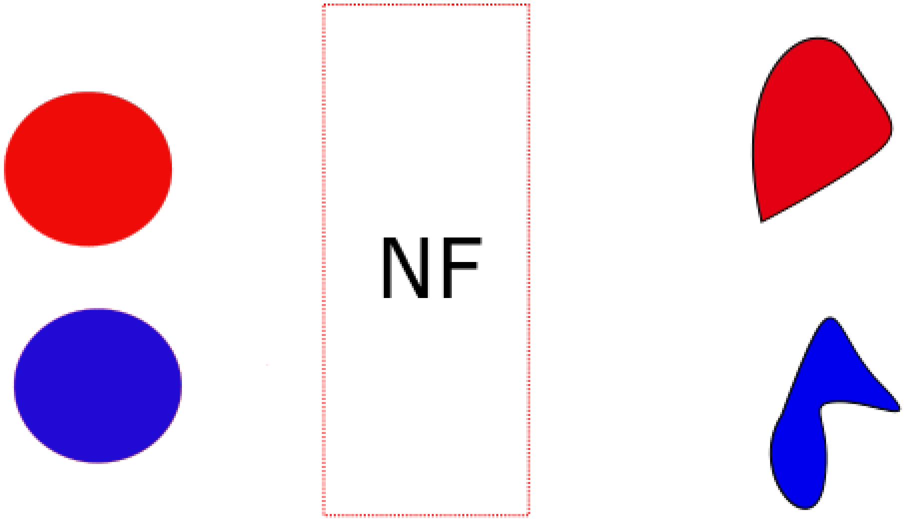
Neural deep generative models

- From general discriminant function to general probabilistic estimator.
- Better to view as a probability transform.



Discriminative NF

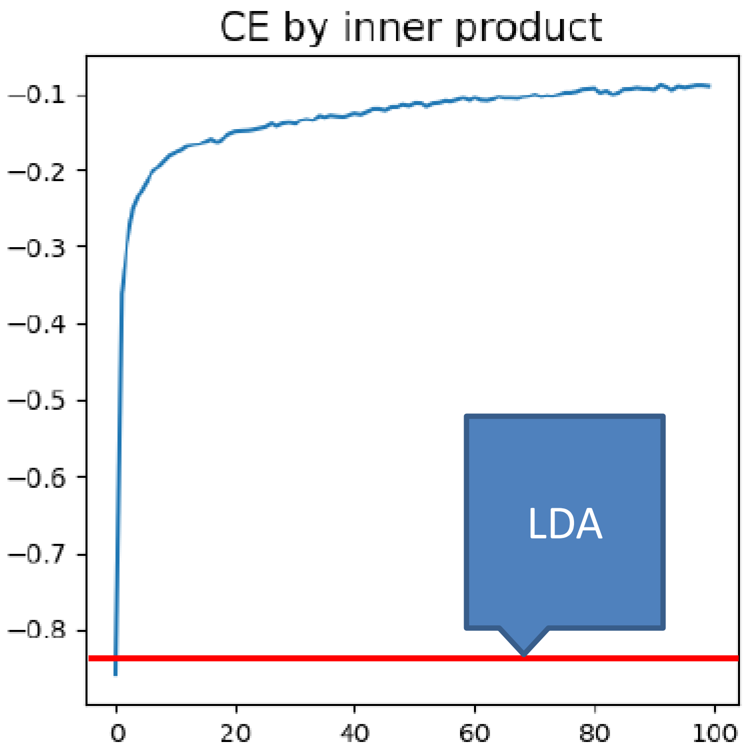
- Principle:
  - Learn $p(x|c)$, then infer $p(c|x)$ by Bayesian rule.
  - No assumption on $p(x|c)$, but homogeneous Gaussian assumption on $p(z|c)$.

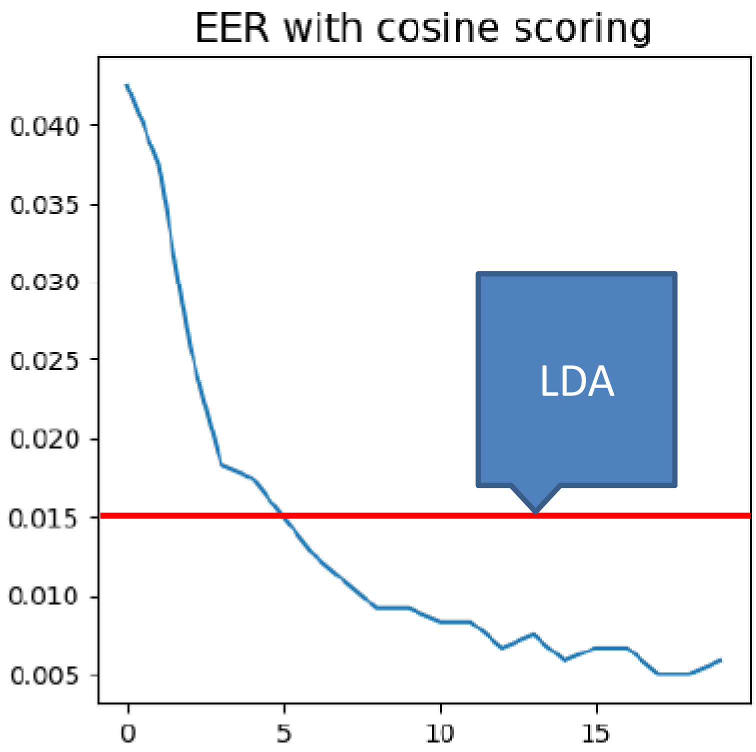


What is different from conventional generative models?

- No assumptions on the distribution form, data driven.
- No limit on both inference and training. Very complex data can be well addressed in theory.
- Mostly black box, not white box.

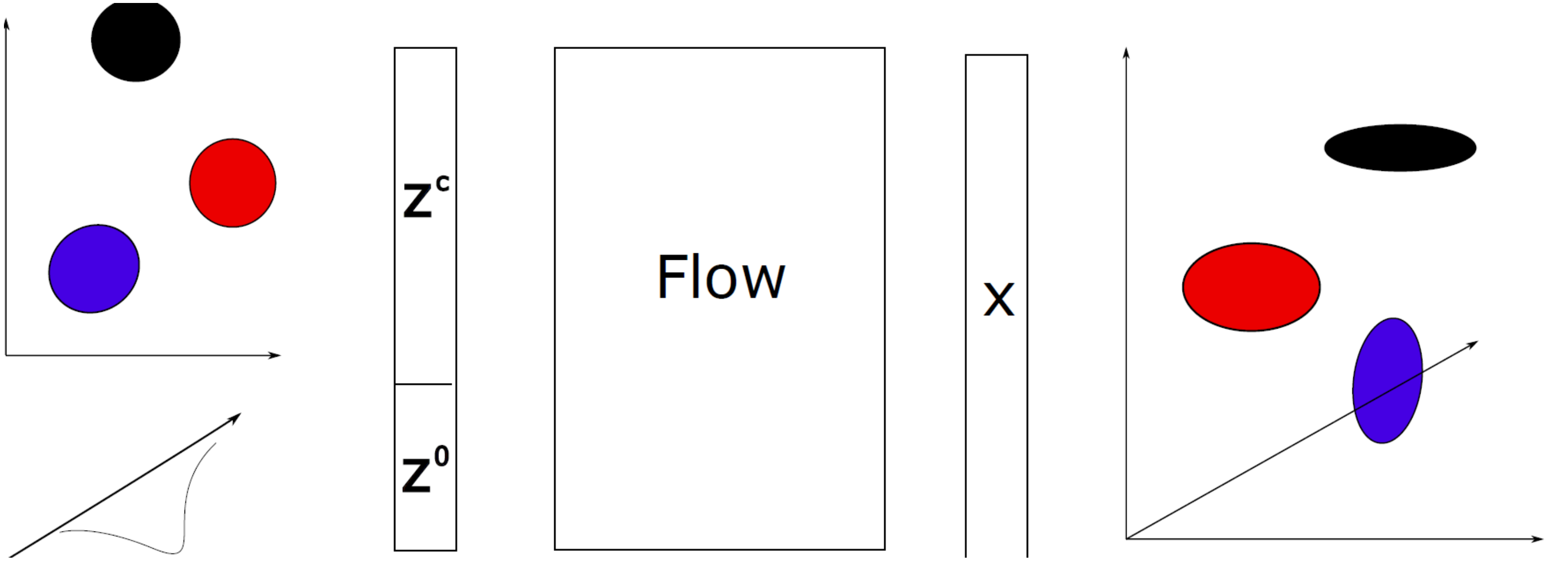
Test on speaker recognition





Subspace DNF

- Discriminative on informative dimension, with homogeneous prior
- Shared prior for residual



Best applications

- Mostly suitable for problems with large number of classes.
- Mostly suitable for generalization (open-set) and adaptation.
- Mostly suitable for weakly-supervised training.