CS294-158 (Spring 2020)

Deep unsupervised learning from @ucb

class video from(SP19) https://www.bilibili.com/video/BV1Eb411Y7J5?p=1

(SP20)https://www.bilibili.com/video/BV1oE411F7iz?p=2

Notes by Haotian Xue from @sjtu

homepage : https://htxue.info 😄

email : xavihart@sjtu.edu.cn ✉

Week 1

1.Motivation

- likelihood-based models:
  estimate p_{data} from samples $\{x^{(i)}\}$
- trade-off(to get the data distribution):
  - Efficient training and model representation
  - Expressiveness and generalization
  - Sampling quality and speed
  - Compression rate and speed

2.Simple generative models

- Just count it
  - JUST A histogram
  - fail in high dimension, poor in generalization
  - Solutions : function approximation $p_\theta(x)$
- To get $p_\theta(x)$, maximum likelihood:

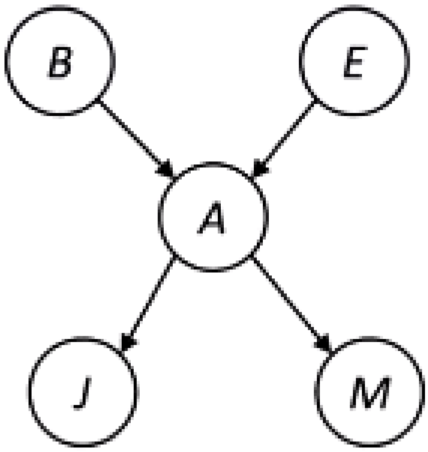
$$argmin_\theta loss(\theta, x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} -log(p_\theta(x^{(i)}))$$

  等价于计算数据01分布和$p_\theta(x)$的KL散度最小

  -> Maximum likelihood + SGD

- (*) Bayes Network(Belif net / causal net)
  - DAG: vertex->property & edge->dependency & define parents and children
  - PGM(probability graph model) = Markov(无向) Net + Bayes Net(有向)
  - sparsity the 2^i sized tabular

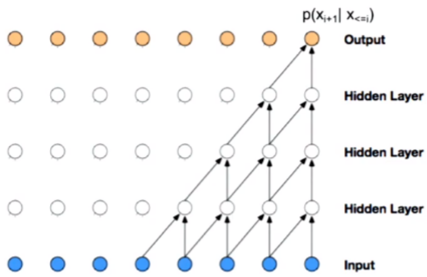
B	P(B)
+b	0.001
-b	0.999

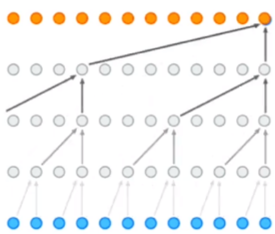


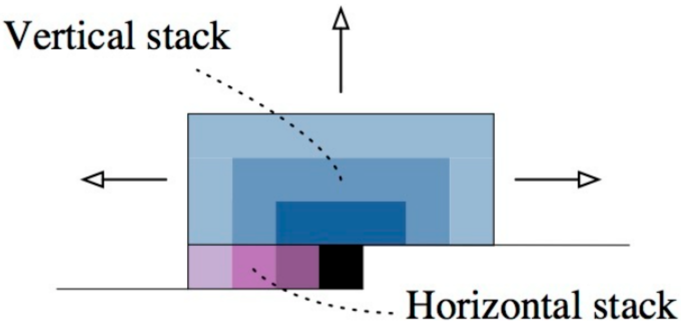
E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

- Autoregressive Models
  - a fully expressive Bayes Net (just a chain rule model)
  - $logp(x) = \sum logp(x_i|x_{1:i-1})$
  - A toy example: p(x1, x2) = p(x1)p(x2|x1)
    - p(x1) : histogram
    - p(x2|x1): MLP with input x1 and output joint distribution of p(x2|x1)
    - Extent to high dimensions:
      - only need O(d) Param instead of O(e^d) tabular Param
      - no share of information between different conditional distribution
  - popular models:
    - RNN
    - Mask
      - masked MLP (MADE[masked auto encoder for distribution estimation])
        - satisfy the **autoregressive property**, the output of d dimension is only related to the input before the d dimension.
        - **more** to referred to in the MADE arxiv paper
      - masked convolutions
        - use the convolutional kernel 

        - limited receptive filed ; faster
        - Wave-Net  dilated convolution

            - pixel-CNN (2016) :
                - combines two kind of convs together: vertical + horizontal
                - 

            - gated pixel CNN :
                - with improved conv structure : Gate Residual Block
        - self-attention 注意力机制

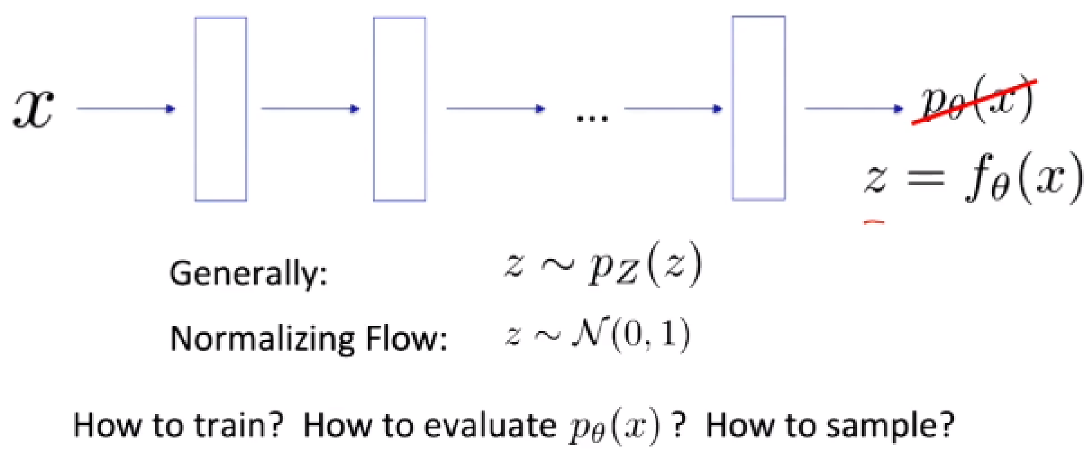
3.Modern NN-based autoregressive models

Week 2

- Foundation of flows (1-D)
    - how to fit a density model
        - mixture of gaussians ?

$$p_\theta(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x; \mu_i; \sigma_i)$$

  not right for high dimensional data !



    - x -> z , can calculate and get a bridge between p(x) and p(z)
    - After SGD optimization to get z, we sample z and project back to x to get the real sample
    - 有点类似DIP里的直方图均衡
- 2-D flow： the same as 1-D
- N-D flow
    - Autoregressive flows and inverse autoregressive flow
        -

    - RealNVP-like arch
    - Glow, Flow++, FFJORD
- dequantization

--------------- TO P3 60:00 ---------------------