生成模型 (generative model)

Discriminative vs Generative Models

Consider supervised learning task:

Given training data $\{x_n, y_n\}_{n=1}^{N}$ — features $x_n \in \mathbb{R}^D$, labels $y_n \in \mathcal{Y}$

Learn to predict $p(y_*|x_*)$ labels-given-features

Two Approaches to probabilistic modeling

Discriminative (Sec 4.3)

Model: $p(y|x) = \prod_{n=1}^{N} p_w(y_n|x_n)$

only y is r.v.

x treated as fixed/known

Parameters: w generates label given feature

Prediction: $p(y_*|x_*, w)$

directly use likelihood

Training: $\max_w \prod_n p(y_n|x_n, w)$

Pro:
- simpler. fewer parameters.
- directly solve supervised task

Con:
- cannot predict if x has missing values

Generative (Sec 4.2)

Model: $p(x, y) = \prod_{n=1}^{N} p_\theta(x_n|y_n)\, p_\pi(y_n)$

both x & y as r.v.

Parameters: θ generate x given y

π generates y

Prediction: $p(y_*|x_*) = \dfrac{p_\theta(x_*|y_*)\, p_\pi(y_*)}{\sum_{y' \in \mathcal{Y}} p_\theta(x_*|y')\, p_\pi(y')}$

via Bayes Rule

Training: $\arg\max_{\theta, \pi} \sum_n \log p_\theta(x_n|y_n) + \log p_\pi(y_n)$

Pro:
- can predict if x w/ missing values
- can sample new x from $p(x|y)$

Con:
- more complex prediction
- more parameters

unsupervised

$p(X, Z)$

任务：监督 vs. 非监督

监督 ↓ (分类，回归，标记)

非监督 ↘ (降维，聚类，特征学习，密度估计，生成数据)

监督
- 概率模型
  - 判别：logistics regression, CRF
  - 生成：
- 非概率模型：PLA(感知机)，SVM，KNN，NN(若差率的)，tree model

非监督
- 概率模型：生成模型
- 非概率模型：PCA，LSA，k-means，autoencoder

生成模型：

Naive Bayes,

Mixture model：GMM

time series model：HMM，kalman filter，Particle filter

Non-Parameter：Bayesian model
- Gaussian Process
- Dirichlet process

mixed memership model：LDA

Factorial model：Factor analysis，P-PCA

潜在生成模型：Energy-based model：Boltzmann machine (无向图)

VAE

GAN

Autoregressive model

Flow-based model：normalizing flow

PCA → P-PCA → FA

Kmeans → GMM

Autoencodes — VAE

LSA → P LSA → LDA

模型表示 (生成模型)

形神兼备

- 形
  - discrete vs continuous
  - directed vs undirected
  - latent variable model vs fully observed model
  - Shallow vs deep (层)
  - Sparse vs dense (连接)
- 神:
  - parametric vs. non-parametric
  - Implicit density (GAN) vs Explicit density $P(X)$

参数化方法：参数数量固定，不随着样本数量变化而变化

推断: tractable vs intractable (approximate inference)

学习:

likelihood-based model vs. likelihood-free model

Maximum likelihood

- based → explicit density
  - tractable
    - fully observed (autoregressive)
    - change of variable (Flow-based)
  - approximate inference
    - Variation (VAE)
    - MC (Energy-based)
- free → implicit density
  - direct (GAN)
  - MC (GSN)

概率图 vs 神经网络
(bayesian network) (neural network)

概率图: p(x)的表示
神经网络: 函数逼近器

Bayesian network		Neural network
表示:	结构化的, 稀疏, 浅层, 条件独立性假设；具备可解释性	深层. 稠密. 计算图；可解释性未知
推断:	精确. 近似 (mcmc, VI)	推断容易但无意义
学习:	likelihood maximum	梯度下降(BP) ↳ {链式求导 / 动态规划 (逆向传播结果)}
适合:	high level reasoning	表示学习, low level reasoning

Graphical models vs. Deep nets

Graphical models

- Representation for encoding meaningful knowledge and the associated uncertainty in a graphical form
- Learning and inference are based on a rich toolbox of well-studied (structure-dependent) techniques (e.g., EM, message passing, VI, MCMC, etc.)
- Graphs represent models

Deep neural networks

- Learn representations that facilitate computation and performance on the end-metric (intermediate representations are not guaranteed to be meaningful)
- Learning is predominantly based on the gradient descent method (aka backpropagation); Inference is often trivial and done via a "forward pass"
- Graphs represent computation

Graphical models vs. Deep nets

Graphical models

Utility of the graph

- A vehicle for synthesizing a global loss function from local structure
  - potential function, feature function, etc.
- A vehicle for designing sound and efficient inference algorithms
  - Sum-product, mean-field, etc.
- A vehicle to inspire approximation and penalization
  - Structured MF, Tree-approximation, etc.
- A vehicle for monitoring theoretical and empirical behavior and accuracy of inference

Utility of the loss function

- A major measure of quality of the learning algorithm and the model

Deep neural networks

Utility of the network

- A vehicle to conceptually synthesize complex decision hypothesis
  - stage-wise projection and aggregation
- A vehicle for organizing computational operations
  - stage-wise update of latent states
- A vehicle for designing processing steps and computing modules
  - Layer-wise parallelization
- No obvious utility in evaluating DL inference algorithms

Utility of the Loss Function

- Global loss? Well it is complex and non-convex...

Graphical Models vs. Deep Nets

- So far:
  - Graphical models are representations of probability distributions
  - Neural networks are function approximators (with no probabilistic meaning)
- Some of the neural nets are in fact proper graphical models (i.e., units/neurons represent random variables):
  - Boltzmann machines (Hinton & Sejnowsky, 1983)
  - Restricted Boltzmann machines (Smolensky, 1986)
  - Learning and Inference in sigmoid belief networks (Neal, 1992)
  - Fast learning in deep belief networks (Hinton, Osindero, Teh, 2006)
  - Deep Boltzmann machines (Salakhutdinov and Hinton, 2009)
- Let's go through these models one-by-one