Paper summary template

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   A. 제목: Estimation of Non-Normalized Statistical Models by Score Matching
   B. 저널: JMLR(Journal of Machine Learning)
   C. 저자: Aapo Hyvärinen
   D. 출판연도: 2005
   E. Keywords: non-normalized densities, Score Function

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**
   A. $P(x,\theta) = \frac{1}{Z(X)}q(x,\theta)$, 우리는 $z(x)$를 알아야, PDF($P(x,\theta)$)를 구할 수 있음
   B. 그러나 $z(x)$는 $q(x,\theta)$를 적분한 값으로, $q(x,\theta)$의 dimension이 커질수록 구하기 어렵다는 문제가 있음 (n=dimension 크기라고 할 때, n>2인 경우 문제발생): **문제점 1**
   C. 즉, 이는 $P(x,\theta)$값을 구하는 것을 불가능하게 함.
   D. 과거 논문에서는 해당 과정을 MCMC, 혹은 근사하는 방식으로 우회
   E. 이 논문에서는 확률모델을 log-density의 미분 값으로 표현 (Score Function)
   F. 해당 과정의 경우, gradient of log-density의 distance를 최소화하는 것을 목적으로 함 (즉, model – data 사이의 분포 차이를 최소화함)
   G. 이 과정에서, log를 취해주고, $\delta\theta$ 에 대해서 미분하는 과정에서, $z(x)$값을 고려하지 않고 수식을 전개해 나갈 수 있음
   H. 해당과정에서 data 1개는 PDF가 아닌 단편적인 정보(=0)이고, p_{data}에 대한 Score Function을 구하는 과정이 불가능함: **문제점 2**
   I. p_{data}의 Score Function 값을 $q(x)$의 Score Function에 대한 미분으로 구하는 과정을 증명함으로써, 결과적으로 Object Function을 q에 대한 Score Function만으로 표현함
   J. 해당과정을 통해 얻은 Score Object Function을 Multivariate Gaussian Density함수에 적용해서 확인(문제 X) ICA(Independent Component Analysis) Model에 적용하는 과정을 통해 Maximum likelihood estimation과의 차이를 비교(차이 존재)함
   K. Pseudo-Likelihood Estimation과의 비교를 통해 일관성 측면에서 한계를 가지는 것을 보여줌

L. 결과적으로는 Contrastive Divergence의 계산이 상대적으로 까다롭지만 Intractable models에도 사용이 가능하다는 장점 때문에 계속해서 사용될 것이라는 주장과, MCMC 기반의 목적함수보다 Consistency측면에서 우수하다는 장점을 가졌음을 언급함.

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. Normalized Density Function을 얻기 위해 필요한 Z(Normalized Constant)의 계산이 불가능할 정도로 어렵다는 것은 2005년 이전에도 존재하던 사실로, 이러한 한계를 해결하기 위해 선행 연구에서는 MCMC, 나 Gibbs sampling 같은 기법들을 통해 P(x)를 추정해 왔음.

   B. 그런데 이 논문에서는 Z를 구하지 않고도 Non-normalized(논문에서는 q(x)) 만으로 P(x)를 추정함으로써 새로운 방법론을 제안하고 있다는 점에서 상당히 우수하다고 생각함.

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. 문제해결을 우회하던 기존 연구와 달리 문제자체를 없애는 모습에서 상당히 인상깊었음.

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. 미분을 함으로써 Z에 대한 점을 고려하지 않아도 된다는 장점도 있지만, Local Consistency (Score 정보, 즉 미분 값)을 기준으로 식을 전개한다는 점에서, $P(x,\theta)$의 정확한 정보를 파악하지 못한다는 한계가 있음.
   B. 이 논문에서는 T가 Inf에 수렴하게 된다면 $P(x,\theta)$와 일치함을 보이지만, 이는 실질적으로 이론에 불과함 (무한개의 데이터 = 불가능)
6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
   A. Diffusion 기초 논문

7. 본 Summary를 작성하는 과정에서 생성형AI를 사용했나요?
   A. 아니요

날짜: 2025.06.24

이름: 신준원

서명: 

6.24. 화 Score matching (2005)

Abstract

확률모델의 추정 → $p(x)$ 는 알고있는데 Z 는 모름 (Z, normalization constant)
명시적으로는 구하는게 불가능에 가깝다는 문제가 있음 문제인식.

↳ 기존연구에서는.. Markov chain Monte Carlo (MCMC)로 근사 하는 형태로 근사 파악
선행연구 방안.

이 논문에서는 . model 의 "log-density 미분", 데이터의 log-density 미분 의 차이를 최소화 시키는 방식으로 확률 모델의 형태를 추정함.
objective function

그런데 사실 log-density 미분은 매우어려움 (non-parametric, 비모수)
↳ Simple formula for this object function → contribution.

denotation

$p_x(\cdot)$ = data. $p(\cdot;\theta)$ = model, $\hat{\theta}$ = estimated parameter value

Introduction

문제 $p(\xi;\theta) = \frac{1}{Z(\theta)} \cdot q(\xi;\theta)$ (계산 쉬운 구조) $Z(\theta)$ → 안돼.

$$Z(\theta) = \int_{\xi \in R^n} q(\xi;\theta)\, d\xi$$ ($n>2$ 인 경우, 불가능!!)

(Gaussian 예외 → normalized 가능) → PDF 공식 존재

따라서, MCMC 사용 (or 근사) → 성능이 안좋은 경우도 있음.
또한, Z X → non-normalized model 사용 방법도 역시 존재함.
↳ 아, normalization 필요없는 → 새가 새안 방법로.

min (Score function (x) − score function (model))^2.

이때 score function = gradient of log-density.

with simple formula

Estimation by score function

확률분포에 대한 loss function을 이용 추정 방법.

고려하면 $p(\xi;\theta)$에 대해 log 취 후 미분하면.

$$p(\xi;\theta) = \frac{1}{A(\theta)} g(\xi;\theta) \quad \cdots ①$$

$$\log p(\xi;\theta) = \log g(\xi;\theta) - \log A(\theta) \quad \cdots ②$$

여기서 θ에 대한 항. 즉. 1. data와 무관하므로

이용시 고려하지 않음 → $A(\theta)$에 의존하지 않아도 됨!

각 data $\xi_n \in R^n$

$$\psi(\xi;\theta) = \begin{pmatrix} \frac{\partial \log p(\xi,\theta)}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(\xi,\theta)}{\partial \xi_n} \end{pmatrix} = \begin{pmatrix} \psi_1(\xi,\theta) \\ \vdots \\ \psi_n(\xi,\theta) \end{pmatrix} = \nabla_\xi \log p(\xi;\theta)$$

결과적으로 $$\psi(\xi;\theta) = \nabla_\xi \log p(\xi;\theta) = \nabla_\xi \log g(\xi;\theta)$$

Data $$\psi_x(\cdot) = \nabla_\xi \log p_x(\cdot)$$ ⇒ Data 분포 기반 계산가능.

with non-parametric estimation of PDF.

→ 그러나 해당 과정 필요X (여사님)

loss function

$$J(\theta) = \frac{1}{n} \int_{\xi \in R^n} p_x(\xi) \| \psi(\xi;\theta) - \psi_x(\xi) \|^2 d\xi$$

→ $$\hat{\theta} = \arg\min_\theta J(\theta)$$ J를 최소화 만드는 $\theta = \hat{\theta}$

해당함수 $A(\theta)$는 계산 안됨 해석적 접근법 두 있다는 장점이 있음.

결과, 여기의 ψ_x를 구하기 위해서는 non-parametric estimation problem ⇒ 해결해야 함 (문제2 ☆)

↳ (제안) → Theorem 1 ☆.

간단하게!

Theorem 1 가정 1) model score function $\psi(\xi, \theta)$ 미분가능함.

가정 2) Some weak regularity conditions

→ 부분 적분법을 사용하기 위한 조건

→ $p_x(\xi) \cdot \nabla_\xi \psi(\xi;\theta)$ ☆ → 0

→ 제곱전개

$$J(\theta) = \frac{1}{2}\int_{\xi \in R^n} p_x(\xi) \| \psi(\xi;\theta) - \psi_x(\xi) \|^2 d\xi \quad \cdots \text{①}$$

$$= \frac{1}{2}\int_{\xi \in R^n} p_x(\xi) \| \psi(\xi;\theta)^2 + \psi_x(\xi)^2 - 2 \cdot \psi(\xi;\theta) \cdot \psi_x(\xi) \| d\xi \quad \cdots \text{②}$$

$J(\theta)$, θ와 관계없으므로 상수항 취급 → 제외

$$J(\theta) = \frac{1}{2}\int_{\xi \in R^n} p_x(\xi) \| \psi(\xi;\theta)^2 - 2\psi(\xi;\theta) \cdot \psi_x(\xi) \| d\xi \quad \cdots \text{③}.$$

샘플링 계산가능 문제의 항 → $\psi_x(\xi)$를 구하는것이 불가능

문제의 항

$$\frac{1}{2} \cdot (-2) \sum_i \int p_x(\xi) \cdot \psi_{x,i}(\xi)\, \psi_i(\xi;\theta) d\xi \quad \cdots \text{④}$$

→ $-\sum_i \int p_x(\xi) \cdot \frac{\partial \log p_x(\xi)}{\partial \xi_i} \cdot \psi_i(\xi;\theta) d\xi \Rightarrow -\sum_i \int p(\xi)(\log p(\xi))' f(x) dx$

미분시 $\frac{p'(x)}{p(x)}$

$$\int p(x) \cdot \frac{p'(x)}{p(x)} \cdot f(x) dx = \int p'(x) f(x) dx = [p(x) f(x)] - \int p(x) \cdot f'(x) dx$$

→ 가정 2에 해당함을 0 처리

즉, 문제의 항 $\sum_{i=1} \int p(x) \cdot \psi_i'(\xi;\theta) d\xi$ 부분적분시 → $\psi_x(\xi)$ 없이 $J(\theta)$ 표현가능

$$J(\theta) = \int_{\xi \in R^n} p_x(\xi) \sum_{i=1}^{n} \left[\partial_i \psi_i(\xi;\theta) + \frac{1}{2}\psi_i(\xi;\theta)^2 \right] d\xi + \text{Const} \quad \cdots \text{⑤}$$

$$\psi_i(\xi;\theta) = \frac{\partial \log q(\xi;\theta)}{\partial \xi_i}, \quad \partial \psi_i(\xi;\theta) = \frac{\partial^2 \log q(\xi;\theta)}{\partial \xi_i^2}$$

결과적으로 score function 항은 0으로 수렴 (급수의 법칙)

⑥ $\tilde{J}(\theta) \approx$ Const

n = 차원수, t = 데이터 개수

Theorem2 $J(\theta)=0 \Rightarrow \nabla_\xi p_{x_1}(\xi) = \nabla_\xi p(\xi;\theta)$ 이때 $\theta=\theta^*$

증명 생략. $\Rightarrow \ln p_{x_1} = \ln p(\xi;\theta) + C$ (이때 p_{x_1}, p_θ는 p_{x_1}이므로 $C=0$)

즉, $p_{x_1} = p(\xi;\theta)$ 이다.

Example ⇒ 자세한 증명은 생략하고 확인하는 정도

① Gaussian Density (multivariate)

$$p(x;\mu,\Sigma) = \frac{1}{Z(\mu,\Sigma)} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)\right)$$

Σ : Symmetric positive-define matrix ⇒ $\nabla\Sigma(x-\mu)(-\frac{1}{2})$

$$g(x) = \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)\right)$$

$$\nabla g(x) = \psi(x;\mu,\Sigma) = -\Sigma(x-\mu)$$

$$\partial\psi(x;\mu,\Sigma) = -\Sigma_{ij}$$

$$\tilde{J}(\mu,\Sigma) = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\phi=1}^{n} -m_{ij} + \frac{1}{n}(x(t)-\mu)^T\Sigma\Sigma(x(t)-\mu)\right]$$

이때 각 information ⇒ maximum likelihood estimation과 동일.

즉, Consistency 하다고 볼 수 있음. (증명생략!)

Intuitive Interpretation

ex) non-normalized log-likelihood 의 maximizing

↳ ______ 균등한 분포, $\mu=0$ / variance = infinite.

——→ which is → $-m_{ii}$ 는 샘플.

Estimation of Basic Independent Component Analysis Model (ICA)

기본 형태 $\log p(x) = \sum_{i=1}^{n} G(w_i^T x) + Z(w_1, \ldots, w_n)$

normalization constant $= -\log|\det W|$ (matrix W)

여기서 해당 모델을 다루는 이유 = Sampling이 쉽다.