대학원 전공 소개서

박태준

통계학과

Contents

1 전공 학업

1.1 전공 코스웍 (Coursework)

통계학과 석사과정을 수료하며 수학적으로 깊은 수준의 통계 이론과 다양한 분야의 연구 최신 동향 및 방법론적 수업을 들었습니다. 이를 통해 해당 분야의 이론적 지식과 기술을 습득하고, 학위논문 작성에 활용할 수 있었습니다. 다음은 대학원 석사과정 기간 동안 수강하였던 주요 수업과 사용 교재 목록입니다.

- **확률론**, Durrett, R. (2019). Probability: Theory and examples (5th ed.). Cambridge University Press.
- **통계이론**, Peter J. Bickel & Kjell A. Doksum. (2006). Mathematical Statistics, Basic Ideas and Selected Topics, Volume 1 (2nd ed.). Pearson.

- **응용통계**, Rencher, A. C., & Schaalje, G. B. (2018). Linear models in statistics (2nd ed.). John Wiley & Sons.
- **고급적 통계 방법론**, Efron, B., Hastie, T., & Tibshirani, R. (2021). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press.
- **범주형 자료분석**, Agresti, A. (2019). An Introduction to Categorical Data Analysis. John Wiley & Sons.
- **공간자료의 통계분석**, Cressie, N., & Wikle, C. K. (2011). Statistics for Spatio-Temporal Data. John Wiley & Sons.
- **통계상담 및 실습**
- **통계적 기계학습**, Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
- **딥러닝의 통계적 이해**, Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

1.2 딥러닝 세미나

제가 속해있는 공간통계연구실에서 2022년 7월부터 9월까지 딥러닝 세미나를 진행했습니다. 엄선한 딥러닝 논문들을 수학적과 통계적으로 이해하고 발표하는 활동을 했습니다. 공부하며 처음 딥러닝을 접하고 관심 갖게되는 계기가 되었습니다. 세미나의 전체적인 계획과 관련 논문 선정은 개인적으로 공부에 많은 도움을 준 유호준 박사(Ph.D.2022, Postdoctoral Fellow in University of Houston, USA)가 진행했습니다. 선정한 논문 중 일부 목록은 다음과 같습니다.

- Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.
- Goodfellow et al. (2014). Generative Adversarial Nets. In Advances in neural information processing systems (pp. 2672-2680).
- Arjovsky et al. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd

International Conference on Machine Learning (ICML 2016) (pp. 1050-1059). JMLR.org.

- Wu et al. (2020). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24.

2 논문

공간통계연구실에서는 공간적 혹은 시공간적 의존성이 있는 데이터에 대한 연구가 이루어집니다. 저는 학과 수업과 세미나에서 접한 논문 공부 등을 통해 머신러닝과 딥러닝의 통계적 이해에 대한 관심이 많았습니다. 그래서 저는 딥러닝을 활용해 시공간적 의존성을 학습하는 연구를 하며 이것을 석사 학위 논문으로 작성하였습니다. 논문의 목적은 딥러닝을 활용하여 성능이 좋은 시공간 예측 통계적 방법론을 제안하며 그 이론적 성질을 보이는 것이었습니다. 저는 제 방법론을 통계적 학습이론 관점에서 바라보아 복잡한 시공간적 의존성을 가진 데이터에 대하여 좋은 성능을 보일 수 있음을 수학적으로 증명하였습니다. 추가로 실제 우리나라 미세먼지 데이터에 방법론을 적용해 교차검증을 통해 실제 데이터에 적용해 기존의 방법보다 더 좋은 성능을 보임을 확인했습니다.

2.1 국문 초록

- **제목**: 딥러닝을 활용한 시공간 데이터 예측
- **주요어**: 시공간 데이터, 크리깅, 딥러닝, 심층 신경망, 다층 퍼셉트론, 통계 학습 이론

크리깅은 관측한 시공간 데이터를 이용해 관측되지 않은 위치를 예측하는 통계적 기법이다. 데이터를 예측하여 보간하는데 시공간 데이터의 시공간 의존성을 이용한다. 그러나 복잡한 데이터의 경우 크리깅은 최적의 예측값이 되지 않을 수 있다. 최근 심층 신경망을 이용한 딥러닝은 많은 분야에서 활용되고 있다. 다층 퍼셉트론을 회귀에 활용할 수 있다는 점을 이용해 본 논문에서는 이 신경망 구조를 이용한 새로운 크리깅 방법을 제안한다. 이 방법은 더 복잡한 시공간 확률 과정을 학습할 수 있다. 그다음, 통계 학습 이론의 관점에서 기존의 크리깅 방법과 제안된 방법을 비교한다. 마지막으로, 실제 한국의 미세먼지 농도 데이터에 제안된 방법을 활용하여 이것의 성능을 평가한다. 여기서 교차 검증 방법을 사용한다.

2.2 요약

다음의 그림(Figure 1)은 시공간 확률 과정 $\{Y(\mathbf{s};t)\}$ 을 학습하기 위한 딥러닝 구조입니다.

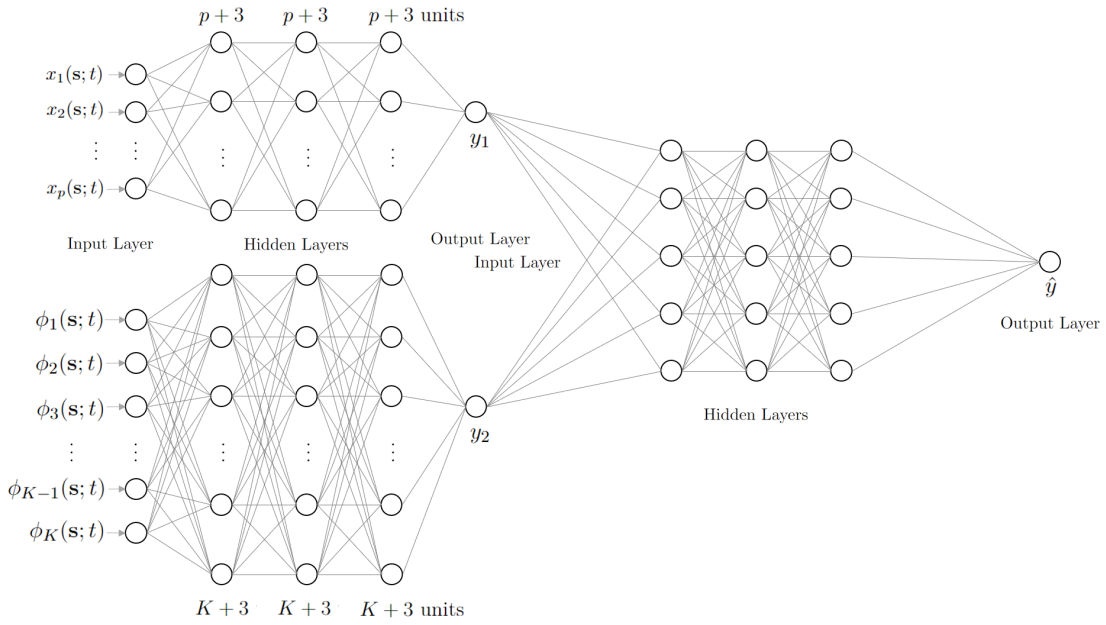

Figure 1: 방법론 시각화

이 구조는 다음의 정리 1(Theorem 1)과같은 수학적 성질이 있습니다. 이는 복잡한 시공간적 의존성을 가진 데이터에 대하여 좋은 성능을 보일 수 있음을 보여줍니다.

Theorem 1. *Let* $\mathcal{F}_{all} = C(V;\mathbb{R})$ *for some compact set* $V \subset \mathbb{R}^{p+K}$, $\mathcal{F}_{MLP} = \mathcal{NN}^{\rho}_{(p+K),1,(p+K)+3}$ *for the ReLU activation function* ρ, *and* $\mathcal{F}_{STUK} \subset \mathcal{C}(\mathbf{x}(\mathbf{s};t)) \cap \mathcal{F}_{all}$. *L is the squared error loss function. Assume that* $\hat{f}_{Bayes} = \underset{f\in\mathcal{F}_{all}}{\operatorname{argmin}} \mathbb{E}L(Y, f(\mathbf{x}))$, $\hat{f}_{Bayes} \notin \mathcal{F}_{STUK}$, *and* $\mathbb{E}[\{L(Y, f(\mathbf{x}))\}^2]$ *is bounded for all* $f \in \mathcal{F}_{all}$. *Then* $R(\hat{f}_{\mathcal{F}_{MLP}}) < R(\hat{f}_{\mathcal{F}_{STUK}})$.

실제 우리나라의 미세먼지($PM_{2.5}$) 농도(μ/m^3)가 높은 날, 보통인 날 그리고 낮은 날에 대해 시공간 데이터에 알맞은 교차검증을 수행했습니다. 그 결과 기존의 전통적인 시공간 데이터 예측 기법인 크리깅 (Kriging) 보다 더 높은 성능을 보임을 확인했습니다. AirKorea(https://www.airkorea.or.kr)의 2022년 1월 9일, 2022년 2월 1일, 2022년 6월 29일 미세먼지 데이터를 사용했습니다. R, Python, Tensorflow 환경을 사용했습니다.

3 프로젝트

대학원 석사과정 중에 학교 내에서 수행한 프로젝트 목록입니다. 통계적 의뢰 프로젝트 같은 경우 서울대학교 통계연구소의 통계 의뢰를 받아 팀을 구성해 수행했습니다. 시공간 데이터 미래 예측 프로젝트는 학위 논문에서 좀 더 발전된 형태를 생각하며 진행한 프로젝트입니다.

- 섀플리 초은하단의 공간 데이터 분석, 2021.09 - 2021.12
- 서울대학교 치과대학 연구 관련 의뢰 통계적 분석 및 자문, 2022.03 - 2022.06
- 삼성서울병원 데이터 분석 관련 의뢰 통계적 자문, 2022.03 - 2022.06
- 수의과대학 학위 논문 관련 통계적 자문, 2022.03 - 2022.06
- 순환 신경망을 활용한 미세먼지 시공간 데이터 미래 예측, 2022.09 - 2022.12