**강의개요**

Shrinkage Methods and Tree Ensembles for High-dimensional Sparse Data

생물정보학에서 다루는 많은 데이터들은 변수의 개수는 많지만 표본 크기는 “상대적으로 작은” 고차원 희박 데이터(high-dimensional sparse data)이다. 예를 들어 마이크로어레이나 RNA 시퀀싱으로 얻어지는 유전자 발현 데이터는 수천 ~ 수만 개의 유전자에 대한 발현 정보를 가지고 있지만 표본의 크기는 대부분 수백 ~ 수만에 지나지 않는다.

본 강의에서는 고차원 희박 데이터가 기계학습에 어떠한 악영향을 미치는지를 직관적으로 설명하고, 이러한 데이터를 분석하는 데 널리 사용되는 shrinkage 방법과 tree ensemble에 대해 설명한다. 선형회귀 및 로지스틱 회귀 기반의 shrinkage 방법이 어떠한 전략으로 고차원 희박 데이터 문제를 해결하는지 설명하고, 그 구체적인 활용 방법에 대해 강의한다. 또한, 고차원 희박 데이터를 다룰 수 있는 비선형 방법인 결정트리(decision tree) 기반의 tree ensemble도 상세히 다룬다.

강의는 다음의 내용을 포함한다:

* Bias-Variance Trade-Off
* 고차원 희박 데이터의 문제점
* Shrinkage 방법 (Ridge, Lasso, Elastic Net)
* Tree Ensemble (Bagging, Random Forest, Boosting)

\*참고강의교재:

An Introduction to Statistical Learning: with Applications in R (Springer, 2013)

\*교육생준비물:

노트북 (동영상 강의 시청용)

\* 강의 난이도: 초급

\* 강의: 황규백 교수 (숭실대학교 컴퓨터학부)

**Curriculum Vitae**

**Speaker Name: Kyu-Baek Hwang, Ph.D.**

****

▶**Personal Info**

Name Kyu-Baek Hwang

Title Professor

Affiliation Soongsil University

▶**Contact Information**

Address 369 Sangdo-ro, Dongjak-gu, Soongsil University, Seoul 06978

Email kbhwang@ssu.ac.kr

Phone Number 02-820-0925

**Research Interests:** Machine learning and bioinformatics

**Educational Experience**

1997 B.S.E. in Computer Engineering, Seoul National University, Korea

1999 M.S.E. in Computer Engineering, Seoul National University, Korea

2005 Ph.D. in Computer Science and Engineering, Seoul National University, Korea

**Professional Experience**

2004 Short-term Visiting Scholar, Children’s Hospital Boston, USA

2012 Visiting Research Associate, Boston Children’s Hospital, USA

2006 - Professor, Soongsil University, Korea

**Selected Publications (5 maximum)**

1. Hwang, K.-B.+, Lee, I.-H.+, Li, H., Won, D.-G., Hernandez-Ferrer, C., Negron, J.A., and Kong, S.W., Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings, Scientific Reports, vol. 9, p. 3219, 2019.
2. Li, H.+, Park, J.+, Kim, H., Hwang, K.-B.\*, and Paek, E.\*, Systematic comparison of false-discovery-rate-controlling strategies for proteogenomic search using spike-in experiments, Journal of Proteome Research, vol. 16, no. 6, pp. 2231-2239, 2017.
3. Li, H., Joh, Y.S., Kim, H., Paek, E., Lee, S.-W., and Hwang, K.-B., Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification, BMC Genomics, vol., 17, no. Suppl 13, p. 3327, 2016.
4. Seok, H.-S., Song, T., Kong, S.W., and Hwang, K.-B., An efficient search algorithm for finding genomic-range overlaps based on the maximum range length, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 4, pp. 778-784, 2015.