MLSS+ICML

2014년 7월 11일

2014년 6월 15일부터 22일에 거쳐 MLSS와 ICML에 참석하였다. 허충길 교수님과 진행중인 probabilistic programming language 연구를 위해서 다녀오는것이다보니 프로그래밍 학회가 아닌 머신러닝 학회를 방문하게 된 것이다. 그 경험을 공유하고자 한다.

1 학회소개

머신러닝 분야에서 가장 큰 학회는 ICML(International Conference on Machine Learning)과 NIPS(Neural Information Processing Systems Conference)이다. ICML은 이름대로 기계학습을 다루고 있고, NIPS는 인지과학이나 기계학습 응용분야 등 좀 더 넓은 주제에 대한 발표도 많은 편이다.

한편 머신러닝을 다루는 여름학교로는 GSS(Graduate Summer School), MLSS(Machine Learning Summer School)이 있다. GSS의 경우 머신러닝만 다룬다기보다는 NIPS에 나올법한 주제들을 폭 넓게 다루는 편이다. 2011년에는 Probabilistic Models of Cognition, 2012년에는 Deep Learning, Feature Learning, 2013년에는 Computer Vision이 주제였다. MLSS는 머신러닝의 기본과 최근의 연구화제를 주로 다루는 편이다. MLSS는 1년에도 여러번 열리는데 그중 한번은 ICML학회 열리는 전 주에 열려서 여름학교와 학회를 둘다 참석하기 좋게끔 되어있다.

허충길 교수님과 확률적 프로그래밍 연구를 하는 중, 확률적 프로그래밍이 널리 쓰일 가능성이 있는 분야중 하나인 머신러닝 분야에 대한 역량을 쌓고자

관련 학회나 여름학교에 참가해 보기로 하였다. 올해 NIPS는 캐나다 몬트리올에서 12월에 열리고 ICML은 중국 베이징에서 6월에 열리기에 올 여름에 ICML과 MLSS에 지원하였다. 발표 논문 없이 참석만 하러 가는것인데 MLSS와 ICML을 전부 참석하기엔 2주 이상의 시간이 소요되는것이 부담스러워 MLSS은 전부 참가하고 ICML은 하루, 튜토리얼 세션만 참석하였다.

2 가는길

2.1 준비

중국에 입국하려면 비자를 받아야 하는데 개인이 직접 발급 받기에는 과정이 복잡한것 같았다. 여행사를 통해 대행하는것을 추천하기에, 연구실 비서장님이 알려주신 우먼여행사를 통해 비행기표 예약과 비자 발급을 함께 해결하였다. 우먼여행사 측에서 여권을 학교에 방문수령 해줘서 편했다. 기간은 일주일 정도 정도 걸린것 같다.

2.2 호텔예약

숙소는 여름학교 홈페이지에서 추천한 호텔리스트 중에서 하나를 골라 예약하였다. 나는 호텔홈페이지에서 예약을 하는것 보다는 가능하면 주로 해당 호텔을 hotels.com이나 booking.com 같은 호텔예약사이트를 통해 예약하는 것을 선호하는 편이다. hotelscombined.com같은 가격비교 사이트를 통하면 간편하게 검색, 예약하는것이 가능하다. 이렇게 하는 편이 간편하기도 하고 가격 비교 하기도 좋은것 같다. 구글맵과 연동되서 검색도 매우 편하다. 게다가 이번엔 중국 호텔이라서 그런지 영문홈페이지가 매우 부실했는데, hotelscombined.com을 이용해 별 어려움 없이 원 하는 호텔에 예약을 할 수 있었다.

2.3 베이징 가는 길

베이징 자체는 도시 구조나 지하철 같은 대중교통 시스템이 서울과 매우 유사하였다. 중국에 처음 가보는것이라 걱정이 됐었는데 막상 가서 보니 딱히

미리 준비하거나 하지는 않아도 되겠다는 생각이 들었다. 신용카드 사용이 어려운 경우가 많으니 위안화 현금만 어느정도 챙기는 것으로 충분한것 같다.

공항에서 베이징 도심으로 가는데는 공항철도를 이용하였다. 공항에서 표지판을 따라 걸어서 공항철도역으로 이동후 자동발권기에에서 지폐를 넣고 기차표를 사면 된다. 2,30분 정도 걸려서 베이징에 도착할 수 있었다. 공항철도과 지하철역이 연결은 되어 있지만 발권은 따로 또 해야 한다. 지하철이 한번 타는데 한화로 약 300원 정도 했던것 같다.

3 학회 분위기

3.1 MLSS

이번 MLSS은 중국인민대학교에서 열렸다. 강의실은 서울대 문화관 대강당과 비슷한 구조의 건물이었는데 화장실이 협소한것만 제외하고는 매우 만족스러웠다. 일정은 꽤 빡빡한 편이었다. 아침 여덟시반 부터 오후 다섯시반 까지 하는 강행군(?)이 6일간 진행되었다.

참가 인원은 300여명 가량 되었고 대부분 중국인 대학원생들이었다. 나는 MLSS에 지원하면서 내심 회사의 업계 종사자도 많이 오고 외국인 비율도 높을거라 기대하고 있었는데, 그렇지 않아서 약간 아쉬었다. 개최지가 중국이라는 점에 큰 영향을 받았을듯 싶다.

참석자간에 교류할 기회도 그다지 많지는 않았다. 강의 사이의 30분 휴식시간과 한시간 뿐인 점심시간이 전부였다. MLSS의 경우 규모가 워낙 커서 그런지 숙소도 알아서 근처 호텔에 예약해야 했기에 강의가 끝나면 전부 뿔뿔이 헤어지는 분위기였다.

3.2 ICML

ICML은 베이징의 Beijing International Convention Center 라는곳에서 열렸다. 시설은 서울의 코엑스와 비슷해서 겉보기에는 그럴싸해 보이는 곳이었는데, 겉보기만 그런것이었다. 컨퍼런스 장소는 넓기만 하지 의자도 불편하고 스크린도 낮아서 앞사람에 가려 잘 안보이는 구조였다. 휴식시간에 주는 간

식도 사람수에 비에 양도 적고 부실하였다. 그래서 학회 시설은 전반적으로 매우 부실하다는 느낌이었다. MLSS가 열렸던 인민대학교 시설이 그리워질 정도였다. 다행으도 점심은 다양한 식당이 모여있는 곳이 근처 건물에 있어서 거기서 저렴하게 사먹을 수 있어 좋았다.

참석자들을 살펴보니 역시 대학교 뿐만이 아니라 회사에서 오는 경우도 꽤 많은것 같았다. 실제로 최근에 머신러닝이 이슈가 되는것은 미국 실리콘밸리 등에서 머신러닝 기술에 기반해 성공한 기업이 많이 나왔기에 많은 기업들에서 관심을 가지고 참여하고 있는것 같다. 우리나라 회사에서 온 사람들도 간간히 보였다. 특히 구글, 아마존, 애플, 마이크로소프트, 페이스북, 트위터 등에서는 머신러닝 이론에도 기여한 바가 크지만, 상업적으로도 온라인 광고 등에서 큰 성공을 거두었기에 머신러닝 학계에서 업계가 차지하는 비율은 꽤 큰 편 같다.

4 인상 깊었던 발표

4.1 Graphical Models and Kernel Methods

확률추론이나 머신러닝에 널리 쓰이는 방법이지만 서로 상반된 이론을 기반으로 하는 두 분야, 확률기반의 PGM(Probabilistic Graphical Models)와 확률에 기반하지 않는 방법인 Kernel Methods에 관한 강의였다.

PGM강의는 Belief Network, Markov Network등 대표적인 PGM에 관한 기본적인 지식부터 시작해, 복잡한 PGM을 Monte Carlo방법이나 Sum-product 알고리듬으로 근사적 계산을 하는 방법에 관한 내용이었다.

그다음 Kernel Method 강의에서는 SVM(Support Vector Machine)과 PCA(Principal Component Analysis)와 이 알고리듬에 Kernel Method를 적용시킨 알고리듬인 kernel SVM, kernel PCA을 다루었다.

Kernel Method는 대표적인 non-probabilistic machine learning 알고리듬 이고 반대로 PGM은 probabilistic machine learning이 기반을 두는 이론이다. 양 극단에 있는 두 분야에 대해 고루 배울 수 있어 많은 도움이 된 강의였다.

4.2 Introduction to Computational Linguistics and Natural Language Processing

자연어 처리는 Deep learning이 흥하고 있는 대표분야중 하나다. 연관 분야로 음성 인식도 있다. 둘다 오래전부터 Gaussian mixture model, hidden Markov model등이 쓰여 왔는데, 문제는 10여년이 넘도록 여기서 큰 발전이 없었다는 것이다. 그러다가 2006년에 deep learning이 고안된 이후로 다시금 주목받고 급격히 발전중인 분야이다. 구글 보이스, 애플의 시리 등도 전부 이러한 머신러닝 알고리즘에 기반하고 있다.

Computational linguistics는 사람이 어떠한 원리로 언어를 학습하고 사용할 수 있는것인지에 초점을 맞추고 있어서, 언어를 이해하는데 초점을 맞추는 NLP와는 약간 다르다. 자연어에 머신러닝을 적용하기 위해서는 적절한 모델을 잡을 수 있어야 하는점에서 CL은 실용적인 측면에서도 자연어 처리와 밀접한 연관을 가지고 있는것 같다.

자연어 처리는 머신러닝의 대표적인 성공사례중 하나다. 처음에는 컴퓨터 인공지능 연구처럼 NLP도 문법구조를 연구하고 그것을 언어 분석에 적용하는 방식이었는데, 머신러닝으로 사전지식 없이 그저 많은 언어자료를 가지고 hidden Markov model을 배우는 방식으로 훨씬 더 좋은 인식율을 보인 덕분이다. 최근에는 HMM보다도 월등한 성능을 보여준 deep learning이 큰 주목을 받고 있다고 한다.

강의 후반은 주로 NLP의 한 분야라고도 할 수 있는 topic model을 설명하는데 할애 하였다. Topic model이라는 분야의 목표는 문서에서 주요 주제를 파악하고 비슷한 주제의 문서끼리 분류해두거나 주제별 문서 검색을 해주는것다. Topic model은 latent Dirichlet allocation이라는 mixture model이 시작이었는데 그 이후로 hierarchical topic model, Bayesian nonparametric topic models등 여러 모델이 나오고 있다. 최근에는 여기에도 deep learning을 사용해서 주목할 만한 성과를 보인 연구도 많이 나오고 있다.

Topic model 분야는 언어의 복잡석이라는 특성상 여러 모델들이 끊임없이 나오고 있고, 앞으로도 계속 나올 분야이다. 또한 방대한 데이터를 대상으로 계산을 해야 하기에 빠른 성능 또한 중요하다. 대부분의 topic model들이 PGM으로 표현할 수 있기에 다양한 PGM들을 편히 구현할 수 있고 빠른

속도로 inference 해줄 수 있는 확률적 프로그래밍 언어(PPL)의 필요성이 큰 분야이다. Topic model 은 그 자체로도 매력적인 연구대상이지만, PPL의 주요 벤치마크 대상으로도 좋은것 같다.

4.3 Emerging Systems for Large-Scale Machine Learning

ICML 튜토리얼 세션중 가장 인상깊은 발표였다. 아주 큰 데이터 셋에 대한 빠른 분석을 위해 분산 in-memory graph processing 프레임워크인 GraphX에 대한 소개였다. 몇년전 UC Berkeley의 AMPLab에서 GraphLAB이라는 인메모리 그래프 분석 프레임워크를 발표했었는데, 이번에 새로이 GraphX를 발표한 것이다. GraphX는 GraphLAB보다는 두배정도 느리지만 GraphLAB이 단일 노드 안에서의 병렬프로세싱에 최적화된것에 반해 GraphX는 Hadoop의 인메모리 프로세싱 시스템인 Apache Spark기반이라서 분산시스템에서 scailability가 좋다는 장점이 있다고 한다. GraphLAB을 중단한것은 아니라고 하는데, 아무래도 GraphX에 더 무게감이 실리는 모양새였다.

그래프 프로세싱은 PGM에 대한 확률추론이나, 소셜 네트워크 분석등 그래프에 기반한 분석에서 아주 중요하다. 한 컴퓨터에 담을 수 없는 큰 데이터셋을 가지고 분석을 하려면 결국 분산시스템에서 계산을 할 수 밖에 없는데 예전에는 이럴때 MPI에 기반한 방식이 주로 쓰였는데 이제는 Hadoop이나 Apache Spark같은 새로운 옵션들도 널리 쓰이기 시작하는것 같다.

4.4 기억남는 다른 발표, 강의

Deterministic algorithms에 무작위적인 요소를 접합시킨 연구들이 여럿 보였다. Randomized Linear Algebra의 경우는 선형대수 연산을 할때 계산시간 대비 정확도를 향상시키기 위해 stochastic한 요소를 도입하였다. SVM, PCA, RBM등 많은 머신러닝 알고리듬들이 선용대수 계산을 주로 하기때문에 머신러닝 시스템의 성능향상에도 큰 도움이 되는 연구가 될듯 하다.

Stochastic Variational Bayes 혹은 stochastic variational inference는 확률추론에서 널리 쓰이는 detereministic한 방법인 variational inference의 성능을 향상시키기 위해 추가로 stochastic한 근사방법을 도입한 것이다.

ICML튜토리얼 세션중 "Bayesian Posterior Inference in the Big Data Arena"은 아주 큰 데이터를 대상으로 계산을 할때 scalability를 높이기 위한 근사방법에 연구였다. 보통 데이터가 아주 크면 분산시스템에 데이터를 저장, 계산을 하게 된다. 이러한 상황에서 Bayesian posterior inference를 계산할 경우 전체 데이터셋을 사용하는대신에 각 노드별로 부분적인 데이터(mini batch of data)만 사용해서 posterier를 계산해서 이를 도합하는 방법에 대한 내용이었다. Monte-Carlo방법에 기반한 근사적 확률추론 알고리듬들은 unbiased estimatior라는 장점이 있었는데, mini-batch 기반으로 계산할때도 이러한 성징을 유지 할 수 있을지도 좋은 연구주제가 될것 같다는 생각이 들었다. 특히 PPL에서 사용자한테 정확도-계산시간 간에 균형을 사용자가 조절할 수 있도록 하는데에 중요한 이슈가 될 것이기 때문이다.

5 정리

이번 ICML학회에서는 학계나 업계 모두 새로운 기계학습 알고리듬보다 알려진 기계학습 알고리듬을 더 큰 데이터에 대해 적용시키는 방법에 대한 연구가 더 주목을 받는것 같은 인상을 받았다. 사실 데이터셋의 크기를 늘리는 것이 성능을 늘리는 가장 손쉬운 방법이라는 점이 벤치마크나 실제 사례들을 통해서 널리 알려져서 그런것 같다. 특히 업계에서는 데이터는 넘쳐나는데 분석을 못하는 경우가 대부분이라고 하니 업계의 수요도 높을듯 싶다.

확률 프로그래밍 언어를 연구하는데 있어서도 이러한 이슈들을 다루는것이 필요해보인다. 연산성능이나 최신 알고리듬의 경우 이미 잘 알려진 유명 기계학습 라이브러리들이 이미 경쟁력을 갖추고 있고 대규모 데이터에 적용하기 위한 확장성의 경우 Hadoop계열의 시스템들이 경쟁력을 갖추고 있는 상황이다. 사용편의성의 경우 Python, R, MATLAB 등의 언어가 가지고 있는 방대한 라이브러리들과 개발툴들이 잘 제공해주고 있다. 확률 프로그래밍을 머신러닝쪽에도 적용시킬 수 있으려면 어디에 특화된 경쟁력을 갖춰야 좋을지 고민이 필요할것 같다는 생각이 들었다.

6 맺음말

처음으로 참석한 머신러닝 학회와 여름학교에서 많은것을 배울 수 있었다.
이러한 기회를 제공해주신 이광근 교수님과 허충길 교수님께 감사드린다.