10주차 - Advanced Topics

- Online Learning

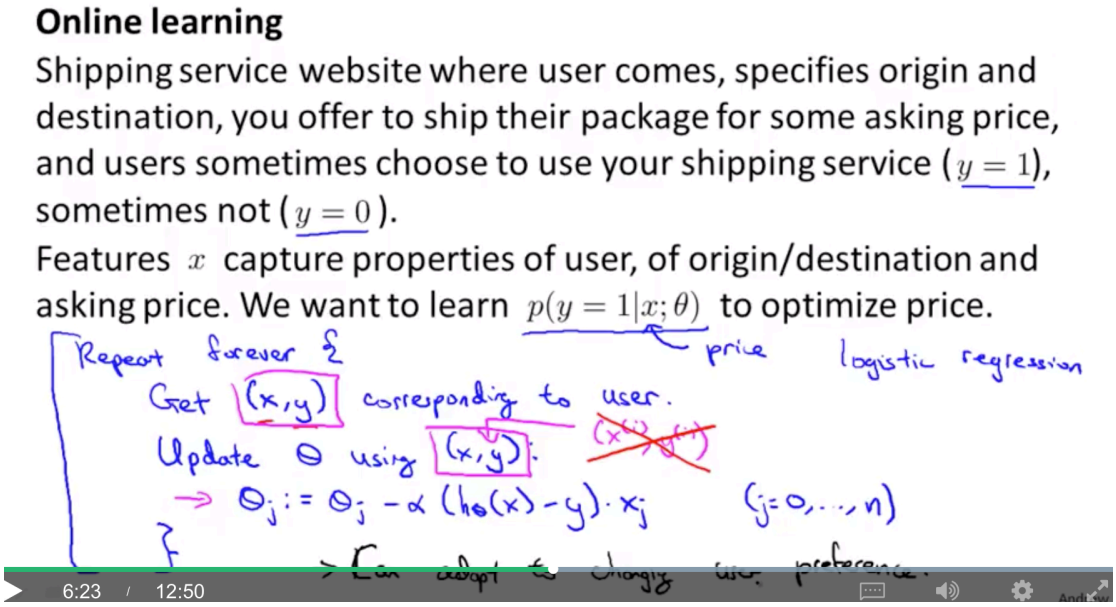
온라인 학습법은 끊임없이 주어지는 데이터속에서 효율적인 학습법으로, 주어지는 데이터에 대해서 학습한뒤 그 데이터를 버린다.
온라인 학습법은 기존 고정 학습법(fixed learning : m의 크기가 변하지 않는 학습법)에 비해 2가지 장점이 있다.

1. 컴퓨터 용량을 아낄 수 있다.

-> 들어오는 데이터를 학습하는 데 사용하고 바로 버리므로 저장할 필요가 없다.

2. user의 선호도 변화에 민감하게 반응한다.

-> 들어오는 데이터에 대하여 학습하므로 최신의 트랜드를 학습한다.
온라인 학습법은 들어오는 데이터마다 처리하는 특성이 있으므로 Stochastic gradient descent 학습법을 사용한다. 다만 차이점은 한번 처리된 데이터는 더이상 확인하지 않는 다는 것이다.



물론 online learning을 적용하는 문제에 fixed learning을 사용해도 상관 없지만 회사에 경우 그 많은 데이터를 저장하기 힘든 경우에는 oneline learning이 효과적이다.
한편, 아래와 같은 사용자 맞춤 시연 시스템을 learning the predicted click-through rate(CTR)이라고 한다.

Other online learning example:

Product search (learning to search)

User searches for "Android phone 1080p camera"

Have 100 phones in store. Will return 10 results.

→ x = features of phone, how many words in user query match name of phone, how many words in query match description of phone, etc.

→ $y = 1$ if user clicks on link. $y = 0$ otherwise.

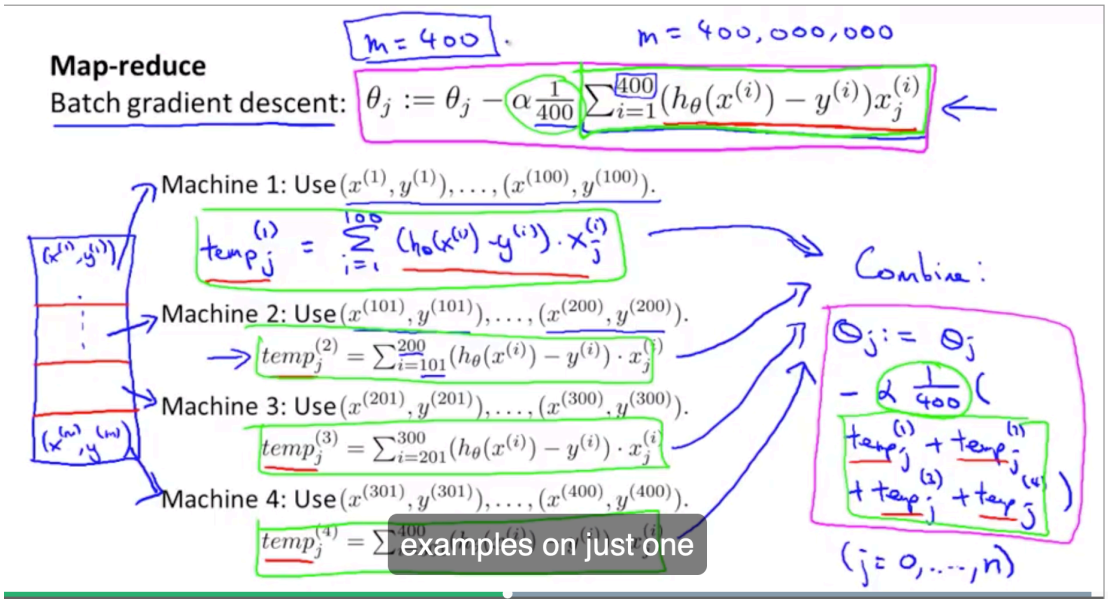
→ Learn $p(y = 1|x; \theta)$.

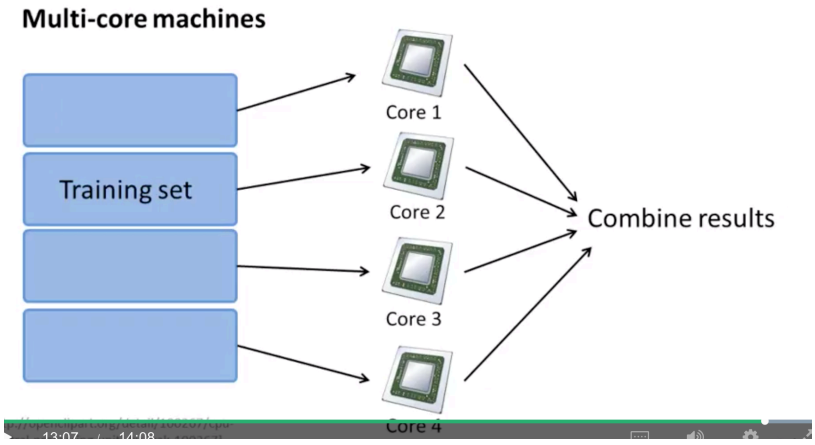
→ Use to show user the 10 phones they're most likely to click on.

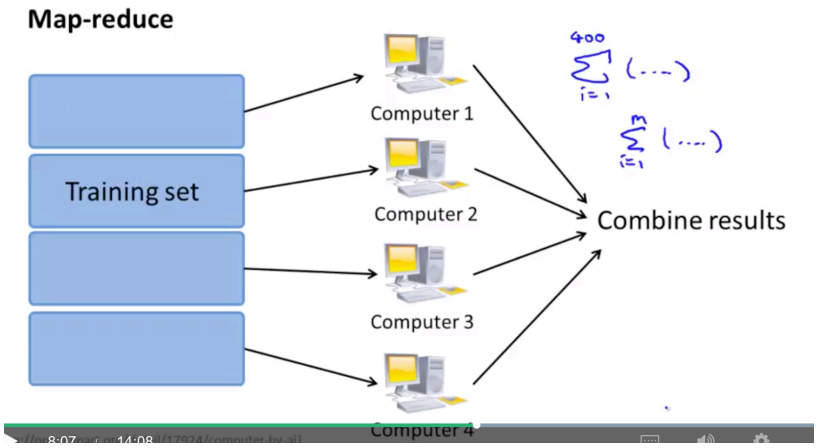
Other examples: Choosing special offers to show user; customized selection of news articles; product recommendation; ...

- # Map Reduce and Data Parallelism

맵 리듀스는 여러 컴퓨터로 데이터에 대한 비용함수의 합을 구할 때 유용한 방법으로 다음과 같이 데이터를 컴퓨터 개수에 해당하는 부분으로 쪼개서 합을 구한 뒤 전체 데이터의 합을 구한다.
맵 리듀스는 이론적으로 데이터를 b개의 덩어리로 쪼개어 병렬적으로 처리하면 b배의 처리속도를 기대할 수 있다.
하지만 데이터를 전송하는 네트워크 속도에 따라 b배 이하로 나올 수도 있다.
한편, 한 컴퓨터의 여러 코어를 사용하여 맵 리듀스를 적용할 경우는 네트워크 속도를 고려하지 않아도 되서 좀 더 b배에 가까운 속도를 기대할 수 있다.
또한, 다중 코어 컴퓨터에서 라이브러리를 이용한 선형 대수 계산의 경우 벡터화된 계산 식을 사용하면 자동으로 병렬 처리를 해준다.




Multi-core machines
Training set
Core 1
Core 2
Core 3
Core 4
Combine results


Map-reduce
Training set
Computer 1
Computer 2
Computer 3
Computer 4
Combine results