Assignment 5

EAS 504

Anish Shah

Computational Advertising is a new discipline that spans areas of computing, economics, and machine learning. Various features such as

Base Questions:

- What are the principal uses of data sciences in this domain?
  A. Value estimation of ads using generalized linear models, targeted delivery of ads using information retrieval, personalized content for each user using machine learning, recommendation system and dynamic pricing using game theory are the principal uses of data sciences in this domain.

- How are data and computing related methods used in the organizational workflow?
  A. Advertisers want to buy ads from publisher who have access to audience. In between these two parties there are various other parties involved such as ATDs, Demand Side Platforms(DSPs), AD networks, Ad exchanges and Supply Side Platforms(SSPs) which acquire user data from data aggregators. Demand side platform is where most of the machine learning is used because they have to accurately value every opportunity to show the ad to make money.

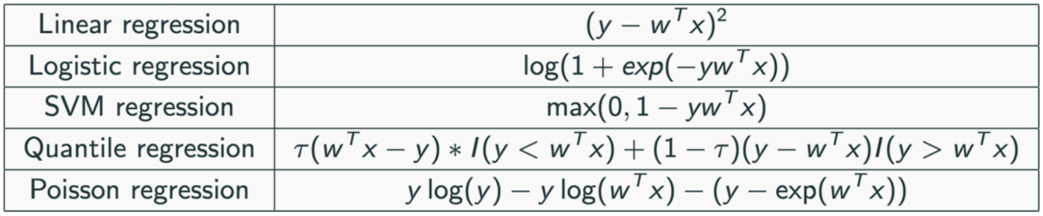
- What data science related skills and technologies are commonly used in this sector?
  A. Skills such as distributed computing to reach web-scale audience, various machine learning techniques for value estimation and personalized content, information retrieval for targeted delivery of ads, machine algorithms such as recommendation systems, gradient descent and dynamic pricing using game theory method are used in this sector. Technologies such as Vowpal Wobbit which is a scalable machine learning library, generalized linear models are used for value estimations and topic modeling using Latent Dirichlet Allocation are used.

- What are the primary opportunities for growth?
  A. Computational advertising has already surpassed TV advertising and has an annual growth rate of 11%, So in about 7 years the company's revenue will be doubled. Computational Advertising industry is projected to be worth $260 billion by 2020. As more and more people nowadays in developing countries or anywhere in the world have started to use smartphones, the audience for showing the ads keeps increasing, therefore increasing the revenue.

Gradient Methods in Machine Learning applications:

Vowpal Wobbit optimizes the loss function which depends on the linear dot product and the label. In order to add a new loss function to VW we need to write the expression for what the loss function is.

VW supports a variety of loss functions such as



Linear regression	$(y - w^T x)^2$
Logistic regression	$\log(1 + exp(-yw^T x))$
SVM regression	$\max(0, 1 - yw^T x)$
Quantile regression	$\tau(w^T x - y) * I(y < w^T x) + (1 - \tau)(y - w^T x) I(y > w^T x)$
Poisson regression	$y \log(y) - y \log(w^T x) - (y - \exp(w^T x))$

So, any function which is convex can be solved or minimize the risk by gradient descent. VW specializes in gradient descent.
There are two forms of gradient descent:

1. **Batch gradient descent**: where we update the weight based on the average gradient with a fixed learning rate. We're using all the data here. It is very efficient. We had a known model ahead of time, we learned the model in less amount of time and it converges in linear time. Here the learning rate is constant.
2. **Stochastic Gradient Descent**: In stochastic learning descent is the function of the time or of the iteration. The rate of convergence is much slower than batch gradient descent. There are different dimensions in which gradient descent might differ such as learning rate schedule, weight update, loss functions. Various stochastic gradient methods used in ML applications are
   - Bag of words accounts only for term frequency, but does not consider its relative frequency within the corpus.
   - TF-IDF (Term Frequency – Inverse document frequency): Here the idea is to weight each word by its relative rarity as we're going from text data to numeric data.
   - LSI (Latent Semantic Indexing): Compute the TF-IDF and SVD is used to get a lower dimensional representation. Each document is represented using its scale vector.
   - Probabilistic Topic Modeling which includes Topic Mixture models: Here the idea is every document belongs to a topic and the topic determines what words show up. It is similar to Gaussian Mixture Modeling(GMM) for numeric data.
   - Latent Dirichlet Allocation: every document has its own distribution on how the numbers are divided. Can't be solved using gradient descent due to its complexity.
   - Vector embedding methods such as word2vec and GloVe are also used in stochastic gradient modelling.