CS 391D - Data Mining: A Mathematical Perspective **Spring 2020**

Project Suggestions

Several possible topics for your course project is listed here.

1. **The NetFlix Prize — Regression with Deep Learning**

   Descrption: Netflix has released movie review data with the goal of improving movie recommendations. They have offered a 1 million prize to the first team to sufficiently improve the recommendations. The problem is a large-scale regression problem: training data, consisting of user ID, date, movie ID, and rating, is given. The test data has user ID, date, and movie ID (but no rating). The goal is to predict the ratings on the test data such that the RMSE (root mean squared error) improves 10 percent over Netflix's system. This project would explore running large-scale regression algorithms on this data, with the goal of achieving as low an RMSE as possible. Try to apply deep learning on this task. You may need to design (or select) an appropriate network structure, propose (or pick) suitable regularization techniques, and so on. As the test data is not available now, you may need to segment the training dataset to examine the performance for your method.

   Reference:

   (1) The paper for the best solution:
   https://www.netflixprize.com/assets/ProgressPrize2007_KorBell.pdf

   (2) About dataset:
   https://opendata.stackexchange.com/questions/7883/netflix-data-set

   (3) Reference Project 1:
   https://github.com/eti-p-doray/log6308

   (4) Reference Project 2:
   https://karthkk.wordpress.com/2016/03/22/deep-learning-solution-for-netflix-prize/

2. **Variational Autoencoder**

   Description: Generative models are cool. A Generative Model is a powerful way of learning any kind of data distribution using unsupervised learning and it has achieved tremendous success in just few years. All types of generative models aim at learning the true data distribution of the training set so as to generate new data points with some variations. Variational Autoencoders (VAEs) are powerful generative models, now having applications as diverse as from generating fake human faces, to producing purely synthetic music.

   Reference:

   (1) Reference paper 1:
   https://arxiv.org/pdf/1606.05908.pdf

   (2) Reference paper 2:
   https://openreview.net/pdf?id=Sy2fzU9gl

(3) Reference paper 3:
https://arxiv.org/pdf/1711.00937.pdf

3. **Generative Adversarial Network**

   Descriptions: Everybody's talking about Generative Adversarial Network (GAN) in the past few years. Try some GANs, apply it to interesting datasets, and see what the result we can get.

   Reference:

   (1) Reference paper 1:
   https://arxiv.org/pdf/1406.2661.pdf

   (2) Reference paper 2:
   https://arxiv.org/pdf/1906.01529.pdf

   (3) Reference paper 3:
   https://arxiv.org/pdf/1703.10593.pdf

4. **Deep Reinforcement Learning**

   Descrption: Deep reinforcement learning is the combination of reinforcement learning (RL) and deep learning. This field of research has been able to solve a wide range of complex decision-making tasks that were previously out of reach for a machine. Thus, deep RL opens up many new applications in domains such as healthcare, robotics, smart grids, finance, and many more. Go find some interesting environments, and train your own intelligent agent. Due to the constraint of computational resources, I suggest using lightweight environments like Mujoco.

   Reference:

   (1) Reference paper 1:
   https://arxiv.org/pdf/1708.05866.pdf

   (2) Reference paper 2:
   https://arxiv.org/pdf/1801.01290.pdf

   (3) Reference paper 3:
   https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf

5. **A Deep Network for Language Modelling—Transformer**

   Description: With the power of deep learning, natural language processing (NLP) develops rapidly. Beyond convolutional networks and recurrent networks, Transformer start a new era of language modelling. The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. Test transformer on an NLP task, then you can try to find how its performance is influenced by the different parts in its structure.

   Reference:

   (1) Reference paper 1:
   https://arxiv.org/pdf/1706.03762.pdf

   (2) Reference paper 2:
   https://ai.tencent.com/ailab/nlp/papers/emnlp2018_deep_representations.pdf

(3) Reference paper 3:
https://arxiv.org/pdf/1807.03819.pdf

6. **Gaussian Process in Deep Learning Era**

Description: A different flavor of learning algorithms are Bayesian methods. Unlike classical learning algorithm, Bayesian algorithms do not attempt to identify "best-fit" models of the data (or similarly, make "best guess" predictions for new test inputs). Instead, they compute a posterior distribution over models (or similarly, compute posterior predictive distributions for new test inputs). Learning with Gaussian process (GP) is one of these methods. a GP is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution. A machine-learning algorithm that involves a GP uses lazy learning and a measure of the similarity between points (the kernel function) to predict the value for an unseen point from training data. The prediction is not just an estimate for that point, but also has uncertainty information—it is a one-dimensional Gaussian distribution (which is the marginal distribution at that point). The combination of (or the relationship between) GP and deep learning is a activate research field.

Reference:

(1) Reference Book:
http://www.gaussianprocess.org/gpml/chapters/

(2) Reference paper 1:
https://arxiv.org/pdf/1807.01622.pdf

(3) Reference paper 2:
https://arxiv.org/pdf/1711.00165.pdf

(4) Reference paper 3:
https://arxiv.org/pdf/1807.01613.pdf

7. **Adversarial Attack and Defense**

Description: With rapid progress and significant successes in a wide spectrum of applications, deep learning is being applied in many safety-critical environments. However, deep neural networks have been recently found vulnerable to well-designed input samples, called adversarial examples. Adversarial examples are imperceptible to human but can easily fool deep neural networks in the testing/deploying stage. The vulnerability to adversarial examples becomes one of the major risks for applying deep neural networks in safety-critical environments. Therefore, attacks and defenses on adversarial examples draw great attention. This topic contains many potential projects: investigating the performance of different attack and defense algorithms; theoretically analysis of the influence of small perturbation to simple networks (like 2-layer ReLU network); improving existing algorithms (e.g., a naive idea is to find an optimal way to combine different adversarial attacks, which may be able to beat powerful defense algorithms).

Reference:

(1) Reference paper 1:
https://arxiv.org/pdf/1801.00553.pdf

(2) Reference paper 2:
https://arxiv.org/pdf/1810.00069.pdf

(3) Reference paper 3:
https://arxiv.org/pdf/1802.00420.pdf

(4) Reference paper 4:
https://arxiv.org/pdf/1412.6572.pdf

8. **Neural Network Certification**

Description: Beyond attack and defense, people expect theoretical guarantees of how neural networks perform under small distortion. Neural network certification provides bounds for the output of a neural networks when the input is slightly perturbed. This is a new research field, and project for this topic can be (1) implementing an algorithm and evaluate its performance on different datasets and models; (2) attempting to derive better bounds for network certification.

Reference:

(1) Reference paper 1:
http://www.cs.utexas.edu/~inderjit/public_papers/fast_robustness_relu_icml18.pdf

(2) Reference paper 2:
https://papers.nips.cc/paper/7742-efficient-neural-network-robustness-certification-with-general-activation-functions.pdf

(3) Reference paper 3:
https://openreview.net/pdf?id=Skg8gJBFvr