Beginner Intro

ML Club Officers

September 2021

1 Introduction

TJ Machine Learning Club's purpose is to teach students about both the theory and practical applications of machine learning. Throughout the year, we'll give lectures that teach the foundations of machine learning and incorporate coding competitions. Feel free to ask questions anytime during the meeting.

1.1 Leadership and Contact Info

- Captain: Aarav Khanna
- Captain: Ron Nachum
- Teaching Coordinator: Irfan Nafi (Nafi)
- Teaching Coordinator: Tarushii Goel
- Teaching Coordinator: Sauman Das
- Sponsors: Mr. Jurj and Mrs. Anderson

If you have any questions, you can reach out to any of us on Facebook or at tjmachinelearning@gmail.com. You should also join our discord! (Invite link: https://discord.gg/Qara7ma4)

1.2 Lectures

Lectures will begin with standard machine learning topics before delving into deep learning. We cover not only classical machine learning and deep learning algorithms, but also new and exciting advances in the field. The lectures will be split into a beginner series for those new to machine learning and an advanced series for those who already know the fundamentals.

1.3 Problem Sets

Problem sets will be given occasionally to help students learn and retain material. You will be given one week to complete them.

1.4 Competitions

Machine Learning Club will be holding in-house contests through Kaggle. Students will be ranked based on their achievement in these contests.

For all competitions, we will be using Python 3. If you're unfamiliar with Python or programming, don't worry because it is fairly simple to learn. We recommend starting here: `https://www.python.org/about/gettingstarted/`.

As the year progresses, Machine Learning Club members can participate in real-world Kaggle competitions (`kaggle.com/competitions`). Substantial prize money is awarded to winners of contests, however, students will be competing against anyone in the entire world, so the probability of winning is extremely low. Nevertheless, Kaggle competitions are a valuable learning experience.

1.5 Rankings

Both your performance on the problem sets and competitions will contribute to your club ranking. Rankings will be used by professors at GMU and Dartmouth's Pathology Department to select students for internships. Additionally, Slingshot will automatically verify top participants in our competitions and connect them with start-ups for internships. Top participants may also be viewed favorably in next year's officer selection process and earn free ML Club T-shirts!

1.6 Research

Beyond learning and competitions, Machine Learning Club values applying concepts we learn into real-world applications and research. If you're interested in conducting research relating to ML or computer science in general, we're available to provide support, advice, mentorship, or help with planning and connections to make your project a reality. Our officers have published papers, presented at conferences, and won awards at the Regeneron International Science and Engineering Fair (ISEF) - we can help you do the same!

1.7 Discussions

We also incorporate discussions of modern advances and emerging machine learning technologies into our meetings, we may sometimes set out the first or last few minutes of a meeting to discuss recent developments or have in-depth dives into technologies like self-driving cars and protein folding.

1.8 The Website

Most information is conveyed through the official Machine Learning Club website, `https://tjmachinelearning.com/`. Here, you can find the lectures along with any presentations, notes, rankings, or additional resources.

2 What is Machine Learning?

Machine learning is a subset of Artificial Intelligence that attempts to generate models based off data in a non-explicit way. The goal of a model is to find patterns embedded within data and use what it learns to predict characteristics on unseen data.

2.1 Supervised Learning

Supervised learning algorithms analyze known training data with labels, the characteristics of the data that we want to predict, to predict the labels of unseen data. For example, an e-mail spam filtering algorithm would analyze previously seen e-mails that are already labeled as being spam or non-spam to predict whether new, unseen e-mails are spam or non-spam. A supervised learning problem where the class labels are discrete (i.e. are made up of distinct categories), such as the spam filter example, is called a classification task. Regression is another type of supervised task where the predicted value is continuous (e.g. predicting a student's SAT score based on their GPA).

2.2 Unsupervised Learning

Unsupervised learning algorithms analyze unlabeled training data. One common unsupervised learning task is clustering, which creates different groups for data and categorizes similar data into each group. It seeks to determine how data is organized without labels on each data point. An example might be categorizing visitors of a website into different groups for advertising purposes.

2.3 Reinforcement Learning

Reinforcement learning is a different subset of machine learning where the learning system (agent) can perform different actions and receives rewards or penalties in return. By doing this repeatedly the agent can learn the correct policy, which dictates which action the agent should take in a given situation, to get the most rewards over time.

3 Vocabulary

Here we have listed a few important words that we will use throughout the year.

- **Label** - The thing we're trying to predict. Ex: If we're trying to predict what kind of animal is in a picture, the label would be the type of animal in the picture.
- **Classification** - Taking each instance and assigning it to a particular category. Ex: Determining if tumors are benign or malignant by looking at MRI scans.

- **Regression** - Instead of having discrete classes, like classification, the "class" to be predicted is made up of continuous numerical values. Ex: Predicting house prices based on square footage, number of rooms, etc.
- **Clustering** - For data without pre-labeled classes, clustering is the act of grouping similar data points together. A form of unsupervised learning. Ex: Clustering U.S. households for marketing data.
- **Training Data** - The initial set of data used to discover potentially predictive relationships. It's what your machine learning algorithm "trains" on and learns patterns from.
- **Testing Data** - Set of data used to assess the strength and utility of predictive relationship. Your machine learning algorithm does not see this data during training.
- **Error** - The difference between algorithm's prediction and ground-truth values.
- **Ground Truth** - Data that is known to be correct. A data-set's labels.
- **Features** - The attributes of the data that are used to make a prediction about the labels. Ex: In the house price example in the regression definition, the features would be the square footage, number of rooms, etc.
- **Feature space** - The n-dimensions in which the features live where $n =$ the number of features. Typically, the larger the feature space, the more complex your algorithm will be.
- **Model** - The relationship between features and label. Training means "learning" this relationship based on examples.