Fundamentals of Machine Learning

2021.6.19

Covering:
⋆ Forms of machine learning beyond classification
⋆ Formal evaluation procedures for machine learning models
⋆ Preparing data for deep learning
⋆ Feature engineering
⋆ Tackling over-fitting
⋆ The universal workflow for approaching machine learning problems

This chapter formalize some of your new intuition into a solid conceptual framework for attacking and solving deep-learning problems. We'll consolidate all of these concepts- model evaluation, data preprocessing and feature engineering, and tackling overfitting - into a detailed seven-step workflow for tackling any machine-learning task.

Four branches of machine learning

Supervised learning: Almost all applications of deep-learning that are in the spotlight these days belong in supervised learning, such as optical character recognition, speech recognition, image classification, and language translation. Although supervised learning mostly consists of classification and regression, there are more exotic variants as well, including the following:
Sequence Generation: Given a picture, predict a caption describing it.
Syntax Tree Prediction: Given a sentence, predict its decomposition into a syntax tree.
Object Detection, Image Segmentation.

Unsupervised learning: This branch of machine learning consists of finding transformations of the input data without the help of any targets, for the purpose of data visualization, data compression, or data denoising, or to better understand the correlations present in the data at hand. Unsupervised learning is the bread and butter of data analytics, and it's often a necessary step in better understanding a dataset before attempting to solve a supervised-learning problem. *Dimensionality Reduction* and *clustering* are well-known categories of unsupervised learning.

Self-supervised learning: It is supervised learning without human-annotated labels - you can think of it as supervised learning without any humans in the loop. There are still labels involved (because the learning has to be supervised by something), but they're generated from the input data, typically using a heuristic algorithm.
For instance, *autoencoders* are a well-known instance of self-supervised learning, where the generated targets are the input, unmodified. In the same way, trying to predict the next frame in video, given past frames, or the next word in a text, given previous words, are instances of self-supervised learning.

Reinforcement learning: Won't talk about for now.

4.2 Evaluating machine-learning models

In machine learning, the goal is to achieve models that *generalize* - that perform well on never-before-seen data - and over-fitting is the central obstacle. It's crucial to be able to reliably measure

the generalization power of your model. In this section, we'll focus on how to measure generalization: how to evaluate machine-learning model.

4.2
⋆ *Data representation*: Randomly shuffle data before splitting it into training and test sets.
⋆ *Arrow of time*: If you are trying to predict the future given the past, you should always make sure all data in your test set is *posterior*.
⋆ *Redundancy*: Make sure training set and validation set are disjoint.

4.3 Data preprocessing, feature engineering and learning

Many data-preprocessing and feature-engineering techniques are domain specific. For example, specific to text data or image data.

Data preprocessing for neural network: Data preprocssing aims at making the raw data at hand more amenable to nn. This includes vectorization, normalization, handling missing values, and feature extraction.
Whatever data you need to process, you must first turn into tensors, a step called data vectorization.
Value normalization: take small values, be homogenous. Some optimal but necessary techniques include: normalizae each feature independently to have mean of 0 and standard deviation of 1.

4.4 Overfitting and underfitting

The fundamental issue in machine learning is the tension between *optimization* and *generalization*. Optimization refers to the process of adjusting a model to get the best performance possible on the training data, whereas generalization refers to how well the trained model performs on data it has never seen before.To prevent a model from learning misleading or irrelevant patterns found in the training data, **the best solution is to get more training data**.

⋆ Reducing the network's size.
⋆ Adding weight regularization.
⋆ Adding dropout: the most effective and commonly used method, which can really contribute to a clear improvement over the reference network.

4.5 Universal workflow of machine learning

1. Defining the problem and assembling a dataset.
Just because you've assembled eaxmples of input X and target Y doesn't mean X contains enough information to predict Y.
2. Choosing a measure of success.
3. Deciding on an evaluation protocol.
Maintaining a hold-out validation, Doing k-fold cross-validation, Doing iterated k-fold validation.
4. Preparing data. (data preprocessing and feature engineering)
5. Developing a model that does better than a baseline (a model with statistical power).
6. Scaling up: developing a model that overfits.
7. Regularizing model and tuning hyperparameters.