Foundations of Machine Learning (CS725)

Instructor: Saketh

Contents

Chapter 1

Basic Concepts and Definitions

Machine learning aims at developing algorithms that mimic the ability in humans to *learn* i.e., improve their "performance" with experience. By performance, we mean their various cognitive abilities. A short note about this is presented below. It is easy to observe that machine learning algorithms will have far reaching consequences in all aspects of living and hence are worth developing.

1.1 Human Learning

Humans seem to have various cognitive abilities. Some of which[1] are: i) finding similarities or patterns or groupings in data, ii) categorizing objects not seen before as novel iii) sequence completion iv) recognition abilities including speech and object recognition v) Summarizing or abstracting text/images/multi-media vi) Decision making vii) Problem solving etc. A little thought will convince that all these abilities improve[2] with increasing experience.

Above discussion convinces that associated with learning there is always a target/unknown concept/ability. Hence-forth, we will call this as the **unknown-concept**. For e.g., in the phenomenon of learning to group data, the unknown concept is the relationship between data/objects and group-ids. In the phenomenon of speech recognition, it is the relationship between speech utterances and English transcriptions etc.

Needless to say, learning happens through experience. Hence **experience** is an important aspect of learning. It is easy to see that **experience** is simply a finite-sized realization (sampling) of the truth in the **unknown-concept**. Depending on

[1]Examples are picked based on the settings that machine researchers have formally studied.
[2]Not necessarily strictly monotonically improving.

the need/application humans express[3] the **unknown-concept** through some **input-output** pairs. For e.g., in the grouping/clustering of data example, the input is the datum and the output is the group-id etc. This clarifies what we mean by **input** and **output**.

Also, how well or fast will a person learn will definitely depend on his current state/capacity/bias/background. We will refer to this as **background** hence-forth. Given this terminology, one could say that the objective in human learning is to determine the **unknown-concept**, but it seems enough to say that the goal is to be able to provide the "right" output for *any* input (because this is how usually humans express their ability/unknown-concept that has been learnt). Summarizing, here are the five important aspects in human learning:

1. **Input** and **Output**
2. **Unknown-concept**
3. **Experience**
4. **Background**
5. **Objective of learning**

1.2 Machine Learning

Though humans possess very many abilities, they are currently far from understanding how they learn/acquire/improve these abilities. So the idea in machine learning is to develop mathematical models and algorithms that mimic human learning rather than understanding the phenomenon of human learning and replicating it.

The formal study of machine learning begins by restricting oneself to certain limited aspects in human learning and postponing the mimicing of human learning in entirety to a later stage. Accordingly we study the follow basic types of learning, which are categorized based on the type of supervisor:

Supervised Learning: This concerns the cases of human learning where the supervisor/supervision is specific/specialized to the goal at hand.

Passive version: mimics learning that happens in babies when taught by *maata-pita*. i.e., learning through specific examples (and non-examples):

[3]If humans could express the unknown-concept directly, rather than in terms of input-output pairs, then perhaps, humans would also understand how they learn!

Classification: E.g., the parents show the baby pictures of various objects while clarifying the name of each object. After being shown a few, the baby starts identifying those objects.

Regression: E.g., everyday (based on few environmental clues) the parents make a prediction about the chance that it will rain (based on which they may advice their kids not to go out to play :). After few days, the kids themselves start making these predictions.

Active version: This concerns the learning that happens in a *shishya* when taught by an *aachaarya*. E.g., After some passive supervised learning, the *shishya* asks questions about the most confusing examples to be clarified by the *aachaarya*. Note that here the choice of example is with the learner rather than the supervisor. This is more popularly known as "Active Learning".

Un-supervised Learning: This concerns the cases of human learning where the supervisor/supervision is not specific/specialized to the goal at hand.

Passive version: E.g., Based on various sentences (spoken in mother tongue) that various people speak to a kid, the kid forms a idea of his language. Then he can predict whether a word, which was never heard by him earlier, to belong to his language or not. Note that here the supervisor is not specific and more importantly, the supervision (the sentences spoken) is not explicitly intended to teach the kid what is his language's formal definition/grammar. Other examples are problems of support/mean/density estimation.

Active version: E.g., Based on actually touching hot and cold water multiple times (and perhaps getting hsi fingers burnt sometimes), the baby figures out the right temperature range of water that is "safe" for him. This kind of learning is popularly known as "Reinforcement Learning". This will not be covered in this course[4].

Now we shall write down the mathematical concepts by which we represent each of the five aspects in human learning mentioned earlier.

1. **Input** by $x \in \mathcal{X}$ ($\subset \mathbb{R}^n$, usually) and **Output** by $y \in \mathcal{Y}$ ($\subset \mathbb{R}$, usually). $\mathcal{X}, \mathcal{Y}$ are called as input space and output space respectively.

2. **Unknown-concept** by a Probability distribution [Ross, 2002] over the inputs (hence-forth denoted by F_X^U) or over the input-output pairs (hence-forth

[4]There are many other learning settings that are formally studied by ML researchers and will not be studied in this course.

denoted by F_{XY}^U). For e.g., in case of support/mean/density estimation or sequence filling problems the probability distribution is over the inputs alone and in case of object recognition, language identification applications described above, the probability distribution is over the input-output pairs.

3. **Experience** by:

   (a) set of input-output pairs. This is the case for e.g., in **supervised learning**.

   (b) set of inputs. This is the case for e.g., in **unsupervised learning**.

   In either case, we call this set as the **training set**. Most importantly, the training set and the unknown distribution have a relation: we assume that the training set is an instantiation of a set of iid random samples from the unknown distribution. We denote the set of iid random variables generating the training set as $D = \{X_1, \ldots, X_m\}$ in un-supervised case and as $D = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ in the supervised case. The training set is an instantiation of D (hence-forth denoted by $\mathcal{D}$).

4. **Background** by (mathematical) Model. We will give examples later.

5. **Objective of learning** by some mathematical statement. For e.g. contruct $f : \mathcal{X} \mapsto \mathcal{Y}$ such that f satisfies certain mathematical condition(s). Examples later.

Chapter 2

Examples of Machine Learning Models/Algorithms

We then began by giving examples of some machine learning models/algorithms. We began with the simplest, and yet perhaps the most powerful and generic, nearest neighbour classifier, which perhaps models the way in which humans identify objects.

2.1 Nearest Neighbour Classifier

Please refer section 13.3 in Hastie et al. [2009] gives a nice overview of this method.

Here the model is extremely simple: the idea is to remember/store the entire training data and when a (new) input is given, search for the nearest input in the training data and assign the label of the (new) input as that of this nearest input.

We began by analyzing this model/classifier formally. The formal analysis is due to Cover and Hart [1967][1].

[1] Available at `http://web.stanford.edu/~montanar/TEACHING/Stat319/papers/cover_nn.pdf`

Bibliography

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2 edition, 2009.

S. M. Ross. *A First Course in Probability.* Pearson Education, 6 edition, 2002.