

Exercise for Introduction to Machine Learning,
Advanced materials processing with intelligent systems

As seen in class, one of the most practical definitions of machine learning is the following:

Def := A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. Suppose your email program checks which emails you do mark as spam, and based on that, learns how to filter spam better. What is task T in this setting?

- 1. Classifying emails as spam or not spam*
- 2. Watching you label emails as spam or not spam*
- 3. The number of emails correctly classified as spam*
- 4. None of the above; this is not a machine learning problem*

In this setting, task T would be classifying emails as spam or not spam. This is the task that the email program is trying to improve at with experience E (i.e., watching you label emails as spam or not spam). The performance measure P would be the number of emails correctly classified as spam.

So, the email program is learning from experience E (watching you label emails as spam or not spam) with respect to the task T (classifying emails as spam or not spam) and performance measure P (the number of emails correctly classified as spam). This is an example of a machine learning problem, as the program's performance at the task T improves with experience E.

Classify the following entries in these categories (E, T, P): Topic: playing checkers.

- 1. The experience of playing many games of checkers*
- 2. Playing checkers*
- 3. The probability that the program will win the next game.*

The experience of playing many games of checkers: E (experience)

Playing checkers: T (task)

The probability that the program will win the next game: P (performance measure)

These entries refer to the elements of the definition of machine learning: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." In this case, the task T is playing checkers, the experience E is the experience of playing many games of checkers, and the performance measure P is the probability that the program will win the next game.

You're running a company, and you want to develop algorithms to address each of these problems:

- 1) You have a large inventory of identical items. You want to predict how many of these items will be sold over the next three months*
- 2) You would like to have a software that examines individual customers' accounts, and for each account, decide if it has been compromised. Should these two issues be treated as a classification or regression problems?*

- 1. Treat both as classification problems*

2. *Treat problem 1 as a classification problem, problem 2 as a regression problem*
3. *Treat problem 1 as a regression problem, problem 2 as a classification problem*
4. *Treat both as regression problems*

For these two problems, you should:

Treat problem 1 as a regression problem, and treat problem 2 as a classification problem.

Problem 1 involves predicting a continuous value (i.e., the number of items that will be sold), which is a characteristic of a regression problem.

Problem 2 involves predicting a binary outcome (i.e., whether or not an account has been compromised), which is a characteristic of a classification problem.

Which of the following problems should be addressed using an unsupervised learning algorithm?

1. *Given email labeled as spam/non-spam, learn a spam filter*
2. *Given a set of news articles, group them into sets of articles about the same story*
3. *Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.*

The following problems should be addressed using an unsupervised learning algorithm:

Given a set of news articles, group them into sets of articles about the same story

Unsupervised learning algorithms are used to learn patterns in data without being provided with labeled examples. In this case, the problem involves grouping news articles into sets based on their content, which is a task that can be tackled using an unsupervised learning algorithm.

It's worth noting that the other two problems listed are supervised learning problems, as they involve learning from labeled examples (i.e., email labeled as spam/non-spam, and a dataset of patients with diagnoses). Supervised learning algorithms are used to learn patterns in data by being provided with labeled examples that include both input data and the corresponding output labels.

Given a collection of raw signals from the optical emission from the process zone in a window of 20 ms, and the corresponding labels in terms of defect content. Which of the following algorithms should I use:

1. *Artificial Neural Network*
2. *Logistic regression*
3. *Linear Regression*

Whereas it is difficult to recommend a specific algorithm without knowing more about the specific characteristics of the data and the requirements of the problem, we can provide some general guidance on the potential suitability of different algorithms for this type of problem:

Artificial Neural Network: An artificial neural network (ANN) could be a good choice for this type of problem, as it is a powerful machine learning algorithm that can learn complex patterns in the data.

ANNs are particularly well-suited for tasks that require learning and classifying patterns, such as image or speech recognition.

Logistic regression: Logistic regression is a type of regression algorithm that is used for classification tasks. It is a good choice for problems where the output is binary (e.g., defective or non-defective) or multi-class (e.g., multiple types of defects). However, it may not be as suitable for tasks that require the ability to learn and classify more complex patterns in the data.

Linear regression: Linear regression is a type of regression algorithm that is used for predicting a continuous value (e.g., a numeric output). It is a good choice for problems where the relationship between the input features and the output is approximately linear. However, it may not be as suitable for tasks that require the ability to learn and classify complex patterns in the data.

It's worth noting that these are just a few examples of the many types of machine learning algorithms that are available, and the best choice will depend on the specific characteristics of the data and the requirements of the problem. It may be necessary to try out different algorithms and evaluate their performance to determine the best approach for a particular problem.