

# Machine Learning Theory (CSC 482A/581A) - Lecture 1

Nishant Mehta

## 1 Administrative stuff

All the information about this course can be found on the course webpage:

[http://web.uvic.ca/~nmehta/ml\\_theory\\_spring2018](http://web.uvic.ca/~nmehta/ml_theory_spring2018)

### Who is this course for?

This course is designed for two types of students:

- Students who previously have taken a course in machine learning and would like to gain a deeper understanding of the theory behind machine learning algorithms.
- Students from theoretical computer science who are curious about machine learning and would like an introduction to the theory of machine learning

Please note that this course is primarily for graduate students; however, advanced 4th year undergraduate students may also take the course if they have the requisite background.

### Why is this course?

The goal of this course is to equip you with the ability to understand theoretical machine learning papers and begin doing research in machine learning theory. The emphasis will be on providing provable guarantees and understanding the proofs, and in particular we will predominantly look at the *statistical theory* behind machine learning.

**This is a theory course and will not involve any implementation/coding.**

### Coursework

- Take-home quiz on first day of class - this does not count towards your grade but is required if you wish to register for (or remain registered in) the course
- 4 Problem sets: You will have at least 2 weeks to work on each problem set. You must work on problem sets individually.
- Take-home final exam: There will be a take-home final exam taking place near the end of the course and possibly overlapping with the examination period.
- Project (graduate students only)

## Marks breakdown

Graduate students		Undergraduate students	
4 Problem sets	15% each	4 Problem sets	20% each
Take-home final exam	20%	Take-home final exam	20%
Project	20%		

## Project

All graduate students will work on a project, starting around the midpoint of the course (at which time I also will release more information about the project and example topics). You may work either individually or in groups of 2. You will submit a brief project report and give a pedagogical presentation to the class near the end of the course or in the exam period. Because there is much to learn, all undergraduate and graduate students must attend the course project presentations.

## 2 Machine learning

The goal of machine learning is to devise algorithms that learn from experience. The “experience” typically is encapsulated by data consisting of *labeled examples*  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ , where

- The *input*  $x$  belongs to an *input space*  $\mathcal{X}$ , such as the set of all  $100 \times 100$  pixel grayscale images;
- The *label*  $y$  belongs to a *label space*  $\mathcal{Y}$ , such as  $\{0, 1\}$  in binary classification.

Given a collection of examples, a learning algorithm seeks to output a *hypothesis*  $\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}$  that maps each input  $x \in \mathcal{X}$  to a predicted label  $\hat{y} \in \mathcal{Y}$ . The hypothesis  $\hat{f}$  output by the learning algorithm is often restricted to belong to some set of hypotheses  $\mathcal{F}$ . Given the true label  $y$  and the predicted label  $\hat{y}$ , the way we charge for errors is determined by some notion of discrepancy between  $y$  and  $\hat{y} = \hat{f}(x)$ :

**Definition 1.** A *loss function* is defined as a mapping  $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ , where the first argument is the actual label and the second argument is the predicted label.

The two most widely used examples of loss functions are shown in the table below.

	$\ell(y, \hat{y})$	$\mathcal{Y}$	Problem
<b>0-1 loss</b>	$\mathbf{1}[\hat{y} \neq y]$	$\{0, 1\}$	binary classification
<b>Squared loss</b>	$(\hat{y} - y)^2$	$\mathbb{R}$	regression

## 3 Statistical learning vs Online learning

With the basic setup in place, we can now introduce the two fundamental types of learning problems that are the focus of this course: statistical learning and online learning. We will often describe examples as coming from some (potentially benevolent) opponent Nature and call the learning algorithm Learner.

### 3.1 Statistical learning

In a statistical learning problem, each labeled example is an independent and identically distributed (i.i.d.) draw from some fixed but unknown probability distribution  $P$  over  $\mathcal{X} \times \mathcal{Y}$ . Learner first observes a batch of  $n$  labeled examples drawn from  $P$  and is charged for errors according to how well it predicts the label of new examples drawn from the same distribution  $P$ . The statistical learning protocol is shown below.

**Protocol:**

1. Nature draws  $n$  labeled examples  $(x_1, y_1), \dots, (x_n, y_n)$  i.i.d. from  $P$  and reveals them to Learner.
2. Learner plays hypothesis  $\hat{f} \in \mathcal{F}$ .
3. Learner suffers error  $\mathbb{E}_{(X,Y) \sim P}[\ell(Y, \hat{f}(X))]$ .

In the above protocol, the  $n$  labeled examples often are called “training examples”, and the collection thereof is referred to as the “training set”<sup>1</sup> or the “training data”. In the last step of the protocol, the learner suffers some error based on its expected performance in making predictions on new examples drawn from the same distribution  $P$ . This expected future loss is called the risk:

**Definition 2.** Given a probability distribution  $P$  and a loss function  $\ell$ , the *risk* of a hypothesis  $f$  is defined as

$$R(f) = \mathbb{E}_{(X,Y) \sim P} [\ell(Y, f(X))].$$

There is a clear connection between the training examples and the new, random example appearing in the definition of the risk, since they all are drawn from the same distribution  $P$ . Consider instead the following hopeless learning problem: let  $\mathcal{Y} = \{-1, +1\}$ , and let the training examples be drawn from  $P$  but the expectation in the risk instead taken with respect to a different distribution  $Q$ . Moreover, perhaps under distribution  $P$  we have that  $Y = f^*(X)$  with probability 1 for some function  $f^*$  while, on the other hand, under distribution  $Q$  we instead have  $Y = -f^*(X)$ . Obtaining low risk in new problem is impossible due to the complete mismatch between a good hypothesis for the risk under  $P$  versus a good hypothesis for the risk under  $Q$ .

**Example 1.** Spam classification is a classical<sup>2</sup> example of binary classification. The training examples consist of a corpus of emails that are labeled spam or not spam. An email is represented as a binary vector that indicates whether not each word from a vocabulary of size  $d$  occurs in the email. The label  $+1$  corresponds to “spam” and the label  $-1$  corresponds to “not spam”. After learning some hypothesis  $\hat{f}$  based on the training set, Learner is assessed according to their misclassification probability on new examples, i.e., their expected 0-1 loss:  $\mathbb{E} [\mathbf{1}[\hat{f}(X) \neq Y]] = \Pr(\hat{f}(X) \neq Y)$ .

### 3.2 Online learning

Whereas in statistical learning, Learner observes a single batch of training examples and plays a single action (the hypothesis  $\hat{f}$ ), in online learning the examples are revealed sequentially, and prior to observing the label of a given example Learner must first predict the label and suffer the corresponding loss. The online learning protocol is formalized below.

---

<sup>1</sup>Note, however, that they need not constitute a set and should instead be thought of as a sequence.

<sup>2</sup>Since the 2000’s, anyways

**Protocol:**

For round  $t = 1 \rightarrow T$ :

1. Nature selects example  $(x_t, y_t)$  and reveals  $x_t$  to Learner.
2. Learner plays hypothesis  $f_t$ .
3. Nature reveals true label  $y_t$  and Learner suffers loss  $\ell(y_t, f_t(x_t))$ .

In the online learning setting, we in general make no assumptions about Nature's choice for the sequence  $(x_1, y_1), \dots, (x_T, y_T)$ . That is, not only is it possible that the examples are not drawn i.i.d. from some probability distribution; it might even be the case that Nature is *adversarial* and devises a sequence which reacts to the previous predictions of Learner in order to maximize the amount of loss suffered by Learner.

Consider for a moment the implications of a deterministic Learner playing against a fully adversarial Nature in the case of 0-1 loss. Since Learner's algorithm is deterministic, for any input  $x_t$ , Nature can always ensure that Learner makes a mistake by playing  $y_t = -f_t(x_t)$ . Thus, Learner always suffers the maximum possible cumulative loss of  $T$  against such an adversary. There are at least two potential resolutions to this trivial outcome:

1. Place restrictions on adversary, so that there is some underlying pattern that ties together the labels to the inputs. This is the restriction that we will look at for the time being.
2. Instead of seeking to minimize the cumulative loss suffered by Learner, instead seek to minimize the difference between Learner's cumulative loss and the cumulative loss of some benchmark online prediction strategy. We will explore this in more detail when we get to the online learning portion of this course.