Machine Learning	**Spring semester, 1997**

Lecture 1: Feb. 19th, 1997

Lecturer: Yishay Mansour — *Scribe: Sigal Sahar*

1.1 Introduction and Overview

1.1.1 Goals of Machine Learning

The goals of Machine Learning are to specify and find automatic processes aimed at organizing or classifying data in order to be able to automatically draw conclusions from the it.

Given a data set, we can have various objectives:

Automatic Inference of a Value

Example # 1: credit. A bank has a large data set of its customers, including information such as the customer's age, income, marital status, financial reputation, etc. When a customer requests a line of credit, the bank must decide whether or not to approve such a request. What is needed is a program that can receive as input the data on the hundred of thousands of the bank's customers and output a set of decision rules on the approval of requests for loans.

It is interesting to note that human decision makers are not capable of flawless decision making in this matter.

Example # 2: medical. In this case we are provided with medical data regarding a particular disease, say, cancer. We would like to construct a hypothesis that given a patient's medical record, will determine whether or not the patient has cancer.

Note that in this case, in contrast to the previous example, the answer is clear and deterministic.

Distribution Estimation

For example, when provided with information regarding the transmissions on a network (when a packet was received, its length, etc.), our aim is to generate a distribution that will simulate the communication on the network.

Automatic Specification — Clustering

Example # 1: Handwriting recognition. When provided with letters written by different people we would like to be able to automatically subdivide the letters into different groups. Preferably, each group will describe a different letter of the alphabet.

Example # 2: Customer Information. Given a data set of potential customers and the profile of their purchasing habits, we would like to subdivide the customers into characteristic shopper groups (the splurger, the frugal shopper, etc.).

Notice that in this case we first make the group subdivision and only then analyze (and label) the groups.

Approaches to the Problems

Each one of the three problems cited above motivates research in the field of Machine Learning. There are two main approaches to the solution of the problems:

Expert Systems. Here the basic assumption is that human experts can solve the problem at hand. With the help of these experts, one can then specify the parameters of the problem and the desired solutions. The problems with this approach are:

- The system cannot perform better than the experts (although it can be based on the combined efforts of various experts).
- The solution is specific to a certain problem and can rarely be applied to a different one.

Machine Learning. Here one aims to develop generic tools that can successfully handle a wide range of problems. The application of a tool of this kind should involve a minimal setting of parameters.

In this course we will concentrate on the second approach, dealing with both theoretical issues and practical algorithmic methods. The purpose in the course will be to combine the two outlooks, theoretical and practical.

In this lecture an overview of the course will be provided; we will see how different methods try to provide solutions to the problems we have seen in this section. The lecture will end with a brief statistical review.

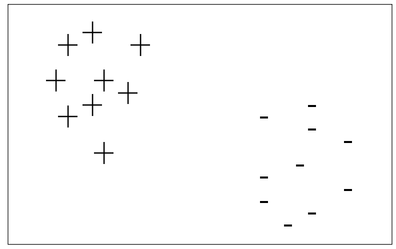

Figure 1.1: Classified Data Set

1.1.2 Complete Information

In a two dimensional space (two coordinates) we are given a data set that has been classified into two groups: '+' and '−' as in Figure (1.1). We would like to be able to classify a new point, (x, y).

In this scenario, we assume we are provided with **Complete Information**: the distribution that generates the '+'s, $\mathcal{D}_+$, the distribution that generates the '−'s, $\mathcal{D}_-$, and that $\mathcal{D}(z) = \frac{1}{2}\mathcal{D}_+(z) + \frac{1}{2}\mathcal{D}_-(z)$. We need to calculate $\mathbf{Pr}[+|(x,y)]$ and $\mathbf{Pr}[-|(x,y)]$ (note that their sum is one).

$$\mathbf{Pr}[+|(x,y)] = \frac{\mathbf{Pr}[(x,y)|+] \cdot \mathbf{Pr}[+]}{\mathbf{Pr}[(x,y)]} = \frac{\mathcal{D}_+(x,y) \cdot \frac{1}{2}}{\frac{1}{2}\mathcal{D}_+(x,y) + \frac{1}{2}\mathcal{D}_-(x,y)} = \alpha \tag{1.1}$$

We are now interested in classifying the point (x, y). Which classification will we choose? That depends on the "**loss**" we will experience for every one of our classification guesses. We, of course, want to minimize the loss. Let us look at various losses.

Number of errors

Here, if we commit an error, we lose 1. Otherwise, we lose nothing.

We compare $\mathbf{Pr}[+|(x,y)] = \alpha$ versus $\mathbf{Pr}[-|(x,y)] = 1 - \alpha$ and make our guess according to α:

- If $\alpha > \frac{1}{2}$, we guess '+', and,
- If $\alpha < \frac{1}{2}$, we guess '−'.

In this case our loss is $min\{\alpha, 1-\alpha\}$. If we chose to make our decision according to an arbitrary q, the loss is $q \cdot \alpha + (1-q)(1-\alpha)$. If $q \neq 0$ or $q \neq 1$, this is a value between α and $1 - \alpha$. Therefore, our choice of α minimize our loss.

Logarithmic Loss

For a guess q '+' we will be penalized as follows:

- $-\log q$ for an erroneous '+', and,
- $-\log(1-q)$ for an erroneous '−'.

Our loss will be $-\alpha \log q - (1-\alpha)\log(1-q)$. For a value of $\alpha = q$ we will achieve minimal loss (one can see this by equating the first derivative to zero). Therefore, one can think of q as the estimate for α. Notice that in order to arrive at minimal loss we need to guess the "right probability".

Quadratic Loss

In this case the loss is given by $(\alpha - q)^2$. Once again minimal loss can be gotten for $\alpha = q$. Notice how here one must know α in order to approximate the loss.

1.1.3 Classifiers

In most cases we do not have complete information on the distributions. In these cases we can take one of two approaches:

Parametric Approach. This approach assumes the distribution is of a certain form (a combination of Gaussians), and the relevant parameters are approximated.

Aparametric Approach. This approach is not based on a particular distribution and should work for any distribution.

Bayes Classifier

We start out with a family of hypotheses that describe the distribution of the '+'s and the '−'s, $\mathcal{D}_+$ and $\mathcal{D}_-$, which we assume to be normal. According to a preliminary distribution of hypotheses, we will choose the hypotheses for the samples, for example, $\frac{1}{2}\mathcal{N}_+ + \frac{1}{2}\mathcal{N}_-$.

Given a sample set $\langle (x_i, y_i), b_i \rangle$ where $b_i \in \{0, 1\}$, we want to construct the distribution of the hypotheses after the sampling:

$$\mathbf{Pr}[h|\mathcal{D}] = \frac{\mathbf{Pr}[\mathcal{D}|h] \cdot \mathbf{Pr}[h]}{\mathbf{Pr}[\mathcal{D}]}.$$

Notice that we need the sample distribution here.

Now we have the distribution $\mathbf{Pr}[h|\mathcal{D}]$ that defines the probability of each hypothesis given the sample. Given a new point we can calculate the probability it will be classified '+' or '−'. Notice that this method is not practical since one needs to go over all the the hypotheses and for each one calculate its distribution.

K-NN: K Nearest Neighbors

We receive a sample set $\langle (x_i, y_i), b_i \rangle$ where $b_i \in \{0,1\}$ is the classification of the respective (x_i, y_i). Given a new point (x, y), we can find the point, (x_i, y_i), that is closest to it and guess b_i. This method is called **1-NN**. **K-NN** involves finding the K closest neighbors of (x, y) and guessing according to the result of a majority vote of the respective b_is.

PAC: Probably Approximately Correct

We assume some "simple" hypotheses class that will serve to classify the new points, (x, y)s as '+' or '−'. For example, in the 2-dimensional case, we can choose a straight line as in Figure (1.2).

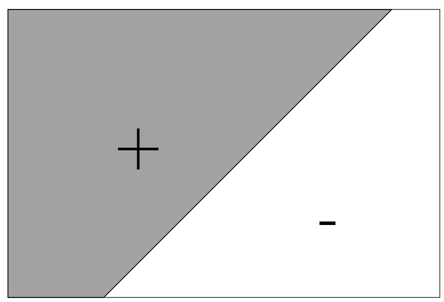

Figure 1.2: Straight line hypothesis

Given a sample set, we choose a hypothesis that will minimize the error on the sample. Each new point will then be classified according to the hypothesis.

We will be extremely interested in the connection between:

- the "complexity" of the hypotheses class,
- the sample set size, and,
- the accuracy of the hypothesis.

The provided sample set and the accuracy level we want to achieve will determine how complex a class we can learn.

Decision Trees

This provides us with an algorithm for finding a "simple" hypothesis with low error. We provide a 2-dimensional example of a decision tree in Figure (1.3).

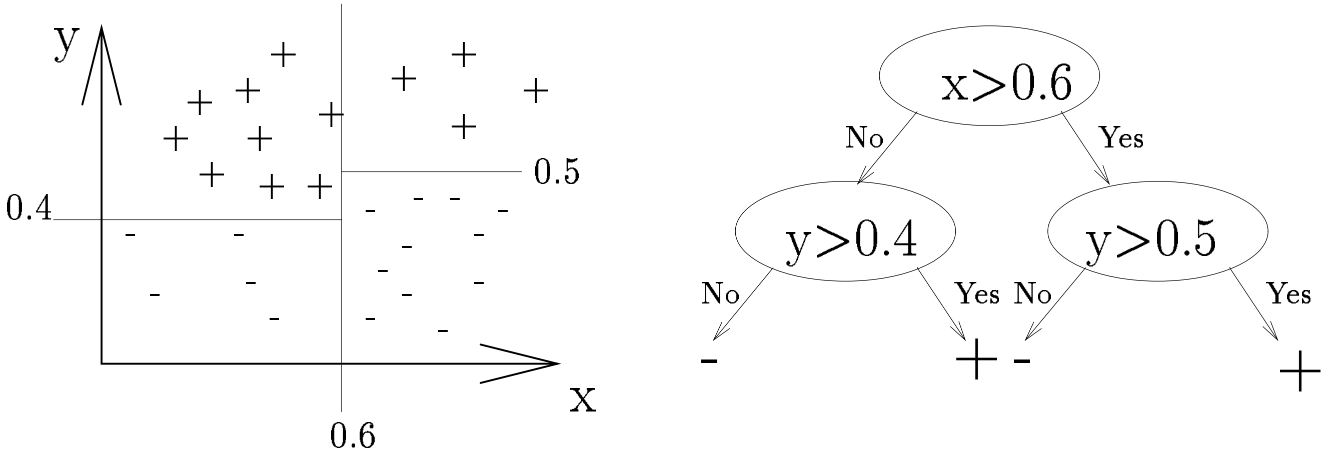

Figure 1.3: An example with a decision tree

The question is, given a sample set, how does one construct the decision tree? We need:

- A criterion for splitting the space (the questions at the nodes). The criterion is used recursively for the tree construction.
- A criterion to determine how much we should allow the tree to grow (the class complexity).

1.1.4 Choosing the Model

Suppose our hypotheses family consists of all finite groups of segments (that it classifies as '+'). In the example of Figure (1.4), we have two hypotheses choices:

1. The one marked with a dotted line, composed of a single segment.
2. The one marked with a dashed line, composed of three smaller segments.

The first method is to be preferred for its simplicity. The second method is to be preferred since it has no error. However, the second method might have fallen in the pit of **over fitting**; if there is noise in the sample, we may be fitting according to the noise. For example, if we believe the rightmost '+' and the two leftmost '−'s are errors, the first hypothesis is a better one. (We will show that even without explicit noise, this phenomenon can still occur).

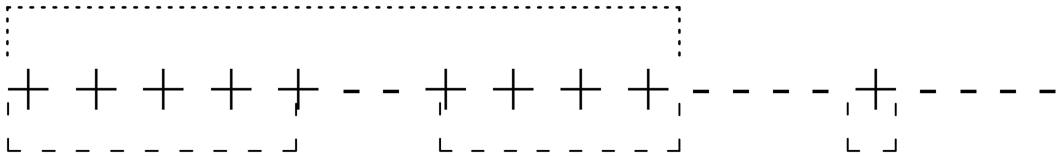

Figure 1.4: Two hypotheses for the sample sample set

Many times it is more advantageous to prefer the simpler (less segments in our example) hypothesis over the more complex one, even at the price of raising the error over the training set.

Clustering

In this case we do not have the classification of the sample set: we must determine it ourselves. The idea is to find a "natural" division of the points in the sample set, as exemplified in Figure (1.5). Ideally, the subdivision will form a group for every possible classification.

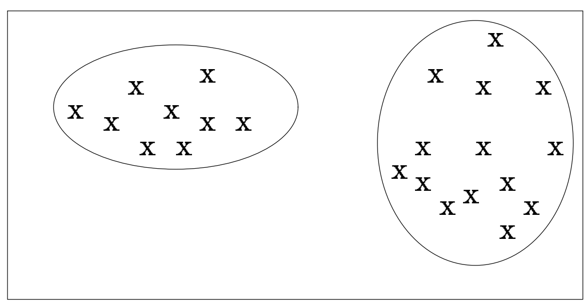

Figure 1.5: Clustering

Let us look at an example of an algorithm that can be used for clustering:

EM: Expectation Maximization

We assume that the hypothesis is of the form:

$$h = \frac{1}{2}\mathcal{N}_1(\vec{x_1}) + \frac{1}{2}\mathcal{N}_2(\vec{x_2}),$$

where $\mathcal{N}_1$ and $\mathcal{N}_2$ are normal distributions with centers $\vec{x_1}$ and $\vec{x_2}$ respectively.

If for each point we knew whether it was generated from $\mathcal{N}_1$ or $\mathcal{N}_2$, then we could compute $\vec{x_1}$ and $\vec{x_2}$ as follows: we would define S_1 to be the set of points $\vec{z}$ constructed from $\mathcal{N}_1$ (and, similarly, we define S_2). Then,

$$\vec{x_1} = average_{z \in S_1}(z).$$

In the same way, if we knew $\mathcal{N}_1$ and $\mathcal{N}_2$ we could compute for each z the probability that $z \in S_1$ or $z \in S_2$ according to:

$$\mathbf{Pr}[z \in S_1] = \frac{\mathcal{N}_1(z)}{\mathcal{N}_1(z) + \mathcal{N}_2(z)}.$$

However, since we do not have this information, we use an iterative algorithm that reconstructs $\vec{x_1}$ and $\vec{x_2}$ and the classification of each point as follows:

1. We start with a guess for $\vec{x_1}$ and $\vec{x_2}$.
2. We calculate S_1 and S_2. If $\mathcal{N}_1(z) > \mathcal{N}_2(z)$, then $z \in S_1$. Otherwise, $z \in S_2$.
3. We re-calculate $\vec{x_1}$ and $\vec{x_2}$ according to

   $$\vec{x_1} = average_{z \in S_1}(z).$$

   and similarly for $\vec{x_2}$.
4. Go to step # 0.

1.1.5 Basic Concepts in Probability

Suppose we have a hypothesis, h_1 that approximates the function f and that someone offers us the hypothesis h_2. How can we determine whether or not we are better off switching to h_2?

Suppose there is a distribution $\mathcal{D}$ on the samples, $\langle x, f(x) \rangle$. We define the error of hypothesis h_i to be:

$$\mathbf{error}(h_i) = \mathbf{Pr}_{\mathcal{D}}[f \neq h].$$

We want to determine if $error(h_1) > error(h_2)$. We randomly choose a sample of size m according to distribution $\mathcal{D}$: $\langle x_i, f(x_i) \rangle$. We will calculate the error on the sample:

$$\hat{error}(h_i) = \frac{k_i}{m},$$

where k_i is the number of errors made over h_i, i.e., the number of is for which $f(x_i) \neq h(x_i)$, for h_1 and h_2. In order to make the switch, $\hat{error}(h_1) < \hat{error}(h_2)$, but is that sufficient?

Let us suppose that both h_1 and h_2 have the same error: $(p_1 + p_2)/2$, where p_i is the error of h_i in the sample set. We want to find out what is the probability that the real error of the hypothesis is $(p_1 + p_2)/2$ and that we will spot an error of p_i in the sample. Let y_i mark the event where $h_1(x_i) \neq f(x_i)$. We will need the following inequalities:

Markov Inequality. For a non-negative random variable, y,

$$\mathbf{Pr}[y \geq \lambda] \leq \frac{E[y]}{\lambda}$$

Chebyshev Inequality:

$$\mathbf{Pr}[|y - E[y]| \geq \lambda] \leq \frac{Var(x)}{\lambda^2}$$

Chernoff Inequality. For i.i.d. (independent identically distributed) $y_i \in \{0, 1\}$,

$$\mathbf{Pr}[|\frac{1}{m}\Sigma y_i - E[Y]| \geq \lambda] \leq e^{-\frac{\lambda^2 m}{2}}$$

In our case, $E[Y] = (p_1 + p_2)/2$ and $\lambda = (p_1 + p_2)/2 - p_1 = (p_2 - p_1)/2$. Therefore, the probability is bounded by

$$e^{-\frac{(p_2-p_1)^2 m}{8}}.$$

That is, the critical (switching) value is $p_1 - p_2 \sim \frac{c}{\sqrt{m}}$.

A more refined test would have us look at *four* parameters defined in Table (1.1.5).

	h_2 is correct	h_2 is mistaken
h_1 is correct	p_{11}	p_{10}
h_1 is mistaken	p_{01}	p_{00}

Table 1.1: Definition of p_{11}, p_{10}, p_{01} and p_{00}.

We are interested in comparing h_1 and h_2. Therefore we are interested only in p_{10} and p_{01}, the points of "difference". Let m' be the number of points (out of m) on which h_1 and h_2 disagree. Much in the same way as before, we can calculate the critical value to be $p_{10} - p_{01} \sim \frac{c}{\sqrt{m'}}$ since the probability is bounded by $e^{-\frac{(p_2-p_1)^2 m}{8}}$.

We can see the second test is more refined, since the first one could not distinguish between two sets of hypotheses with the following errors:

$$\begin{bmatrix} 0 & 40 \\ 60 & 0 \end{bmatrix} \text{and} \begin{bmatrix} 40 & 0 \\ 20 & 40 \end{bmatrix},$$

while the second one can.