COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Amit Koren

Lecture #2
February 6, 2003

1 Probability Review

1.1 Basic Definitions

Event - A probabilistic outcome. Ex. A coin flip resulting in Heads; a die roll resulting in 3; a voter voting for bush
Probability - The likelihood that an event will occur. To get this you need to run an experiment many times and get the frequency of the event. Ex. The probability that a coin flip results in Heads is 1/2. Written as Pr[H]=1/2.
Random Variable - A variable that probabilistically takes on values from a given domain. Ex. Random Variable (RV) X = 1 w/ probability p ϵ [0, 1] and X = 0 w/ probability 1-p. This is known as a Bernoulli RV.
Distribution - The set of probabilities for all possible values of a RV.
$0 \leq$ Pr[X=x] ≤ 1. For discrete cases:

$$\sum_x Pr[X = x] \quad = 1$$

1.2 Expectation

Definition
The expectation of a Random Variable is defined as:

$$E[X] = \sum_x Pr[X = x] \cdot x$$

Ex. For a Bernoulli random variable:

$$E[X] = p \cdot 1 + (1 - p) \cdot 0 = p$$

Ex. For a die:

$$E[x] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

Properties
Expectations of Functions of Random Variables

$$E[f(X)] = \sum_x Pr[X = x] \cdot f(x)$$

Constants and Expectations

$$E[c] = c \qquad\qquad E[c \cdot X] = c \cdot E[X]$$

Linearity of Expectations

$$E[X + Y] = E[X] + E[Y]$$

Example - 1000 coin tosses.
X_i is the result of the i^{th} toss. $X_i = 1$ w/ probability $\frac{1}{2}$ and $X_i = 0$ w/ probability $\frac{1}{2}$.
We flip the coin 1000 times and set a new RV S equal to the sum of the coin flips.

$$S = X_1 + X_2 + \ldots + X_{1000}$$

$$E[S] = E[X_1] + E[X_2] + \ldots + E[X_{1000}] = 500$$

1.3 Conditional Probability

Given two events, A and B, the conditional probability Pr[A|B] equals the probability of A happening given that B has already happened.

$$Pr[A|B] = \frac{Pr[A \wedge B]}{Pr[B]}$$

Ex. 1 - A = voter voted for bush, B = voter is Republican
Pr[A|B] = the fraction of Republican voters who voted for bush
Ex. 2

$$Pr[X \ is \ odd \mid X \geq 3] = \frac{Pr[X \ \epsilon \ \{3, \ 5\}]}{Pr[X \geq 3]} = \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{1}{2}$$

1.4 Independence

Two events, A and B, are independent if Pr[A|B] = Pr[A]

$$\Leftrightarrow \qquad Pr[A] = \frac{Pr[A \wedge B]}{Pr[B]}$$

$$\Leftrightarrow \qquad Pr[A \wedge B] = Pr[A] \cdot Pr[B]$$

Two Random Variables, X and Y, are independent if:

$$\forall \ x, y \quad Pr[(X = x) \ \wedge \ (Y = y)] \ = \ Pr[X = x] \cdot Pr[Y = y]$$

If this is the case:

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

1.5 Union Bound

Given two events A and B:

$$Pr[A \ \vee \ B] \leq Pr[A] + Pr[B]$$

This can be visualized with a Venn diagram. The area of the union of any two regions is always less than the sum of the areas of the two.

1.6 Markov's Inequality

Given a nonnegative RV X and a nonnegative constant k:

$$Pr[X \geq k \cdot E[X]] \leq \frac{1}{k}$$

proof
let $\delta = k \cdot E[X]$

$$E[X] = \sum_{x} Pr[X = x] \cdot x$$

$$= \sum_{all\ x\ \geq\ \delta} Pr[X = x] \cdot x + \sum_{all\ x\ <\ \delta} Pr[X = x] \cdot x$$

$$\geq \sum_{all\ x\ \geq\ \delta} Pr[X = x] \cdot \delta$$

$$= Pr[X \geq \delta] \cdot \delta$$

$$\Rightarrow Pr[X \geq \delta] \leq \frac{E[X]}{k \cdot E[X]} = \frac{1}{k}$$

2 The PAC Learning Model

How does the consistency model relate to machine learning? Finding an exact rule is difficult, so instead we aim for the high accuracy prediction rule, which we obtain by "learning" from a training set.

Assumption 1 - Examples are random. However, we make no assumptions about how they are distributed, so that the results will be distribution-free.

Assumption 2 - All the training and test examples are taken from the same distribution. They are independent and identically distributed (i.i.d.).

Assumption 3 - The labels come from some unknown concept c from some known concept class $\mathcal{C}$

The goal of the learning algorithm is to find a hypothesis h, approximate to c. In other words, the goal is to minimize the error of h, where err[h] = $Pr_{X \sim D}$[h(x) $\neq$ c(x)] (the true error). If err[h] is small, h is "approximately correct." Since the sample is random, it is always possible that a bad sample will be chosen that prevents us from finding an h that is approximately correct. Therefore, we allow the learning algorithm to fail with some small, controllable probability. Thus, we want h to be "probably approximately correct" or PAC. Formally, h is PAC if

$$Pr[err(h)\ \leq\ \epsilon] \geq 1 - \delta.$$

A hypothesis h is ϵ-bad if its error err[h] is bigger than ϵ. Otherwise it is ϵ-good.
$\mathcal{C}$ is PAC-learnable by $\mathcal{H}$ if:
$\exists$ an algorithm A such that
$\forall$ c ϵ $\mathcal{C}$ and
$\forall$ $\epsilon > 0$ and

$\forall\ \delta > 0$ and

$\forall$ distributions D on the examples...

A takes m examples $\langle (x_1, c(x_1)), (x_2, c(x_2)), \ldots (x_m, c(x_m)) \rangle$ and outputs $h \in \mathcal{H}$ such that Pr[h is ϵ-bad] $\leq \delta$ (ie, Pr[h is ϵ-good] $\geq$ 1-δ) where m is a polynomial function of $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.
A is efficient if it runs in polynomial time in m

2.1 Example - The real line

The domain X=$\Re$
The concept class $\mathcal{C}$={all positive half lines}

The concept c is a real number. Any number greater than c is defined as a positive example; any number less than c is defined as a negative example. The hypothesis h is also a real number, and lies somewhere between the least positive example and the greatest negative example.

Define a region to the right of c that covers ϵ of the distribution as R_+ (the rightmost point in this region is labeled b_+). Similary, define R_- and b_- to the left of c.

Define the region between c and h as e_+ if h >c and e_- if h < c.
The error of h is equal to the probability that some element x lies in e_- or e_+.

The ϵ-bad cases are when h < b_- or h > b_+. In other words, h lies outside of R_+ and R_-.

Look at the positive side: h > b_+ only if all the positive training examples are greater than b_+:

$$\Rightarrow x_1 \notin R_+ \wedge\ x_2 \notin R_+ \wedge\ \ldots\ \wedge\ x_m \notin R_+$$

$$Pr[h > b_+] \leq Pr[x_1 \notin R_+ \wedge\ x_2 \notin R_+ \wedge\ \ldots\ \wedge\ x_m \notin R_+]$$

$$= Pr[x_1 \notin R_+] \cdot Pr[x_2 \notin R_+] \cdot \ldots\ \cdot Pr[x_m \notin R_+]$$

$$= (1-\epsilon) \cdot (1-\epsilon) \cdot \ldots \cdot (1-\epsilon) = (1-\epsilon)^m$$

The argument is symmetric for the negative side.

$$Pr[h\ is\ \epsilon - bad] = Pr[h > b_+ \vee h < b_-]$$

$$\leq Pr[h > b_+] + Pr[h < b_-]$$

$$= 2(1-\epsilon)^m$$