Notes for COMP 599
Introductory Lecture - Part 2

Prakash Panangaden

2nd September 2020

Prof. Oberman has given a broad outline of the course. In this course we will focus on probability and leave measure theory in the background. The emphasis will be on understanding the theory behind machine learning algorithms. The theory aims to provide asymptotic probabilistic results. The basic framework, due to Leslie Valiant, is called *Probably Approximately Correct* or PAC learning. We will study some basic PAC learning results in the first part of the course. In these notes I will summarize the basic probability background that we expect from you and which I talked about in the first lecture.

1 Probability background

We model an experiment or a process with n possible outcomes. The set of outcomes is $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$; this is called the *sample space*. Each outcome ω_i has a probability p_i with each $p_i \in [0,1]$ and $\sum_i p_i = 1$. A subset $A \subseteq \Omega$ is called an *event* and we associate a probability with each event according to the formula: $P(A) = \sum_{i \in A} p_i$. So $P(\{\omega_i\}) = p_i$; I will just write $P(\omega_i)$. We have the usual axioms: (i) $P(\emptyset) = 0$ and if $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$. I expect you are familiar with these ideas, words and notations.

Suppose $X : \Omega \to S$ where S is some set (often it will be the reals) then we call X a *random variable* taking values in S. We use the following notation which seems bizarre to analysts but is natural for probabilists: if $U \subset S$ we write $(X \in U)$ for the inverse image of U, *i.e.*

$$(X \in U) := \{\omega \in \Omega | X(\omega) \in U\}.$$

we also write $P(X \in U)$ for the probability of the above set. Thus we obtain a probability measure on S. For the special case where U is a singleton, say u, we write $(X = u)$ and $P(X = u)$.

For the most part S will just be $\mathbb{R}$. I will assume this from now on unless I say otherwise. The *expectation value* of a random variable, written $E[X]$ is $\sum_i p_i X(\omega_i)$ or $\sum_i p_i (X = x_i)$ where $X(\omega_i) = x_i$. In the continuous case the sum is replaced by an integral. The following facts are obvious: (i) if X takes a constant value x then $E[X] = x$, (ii) if $\forall i, X(\omega_i) \geq 0$ then $E[X] \geq 0$, (iii) $E[\cdot]$ is a linear function on the space of random variables and (iv) $|E[X]| \leq E[|X|]$.

We say two events A, B are *independent* if $P(A \cap B) = P(A) \cdot P(B)$ and similarly for larger families of events. Note that this is *not an if and only if*. Note also that if you have a set of events it is

possible that every pair of them is independent without the whole set being independent. Here is an example. Let us consider two tosses of a fair coin. Let A_1 be "the first toss is a head," A_2 be "the second toss is a head" and A_3 be the event that "the two tosses give the same result."

Then every pair is independent: $P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$ and the probability of any pair is $\frac{1}{4}$, but the probability of all three is also $\frac{1}{4}$ and not $\frac{1}{8}$.

If $X_1, \ldots, X_k$ are random variables we say that they are independent if for all real $x_1, \ldots, x_k$ the events $(X_1 = x_1), \ldots, (X_k = x_k)$ are independent. The following facts are well known:

1. If $B_1, \ldots, B_k$ are any[1] subsets of $\mathbb{R}$ then $(X_1 \in B_1), \ldots, (X_k \in B_k)$ are independent.
2. Independence does not depend on the order in which the events or random variables are listed.
3. Any subset of a collection of independent random variables is also independent.
4. Functions of independent random variables are independent.
5. If X, Y are *independent* random variables then $E[XY] = E[X] \cdot E[Y]$.

2 Some elementary probability inequalities

The whole subject turns on discovering and using inequalities. Here is the very first one:
Proposition 1 (Markov)**.** Let X be a positive random variable, then for every $a > 0$ we have

$$P(X \geq a) \leq E[X]/a.$$

Proof Let us write x_i for $X(\omega_i)$. Then we have

$$P(X \geq a) = \sum_{i:x_i \geq a} p_i \leq \sum_{i:x_i \geq a} p_i(x_i/a) \leq \sum_i p_i(x_i/a) = E[X]/a.$$

∎

If we use values of a less than $E[X]$ we learn nothing of course. If, however, a is large then this inequality says that it is unlikely that X takes on values much larger than the mean value. We will greatly sharpen this crude estimate as we go on.

Here is an immediate and useful consequence.
Corollary 2 (Tchebyshev-Bienaymé)**.** Let X be a random variable and let $a > 0$ then

$$P(|X - E[X]| \geq a) \leq (\frac{1}{a^2})E[(X - E[X])^2].$$

It follows from Prop. 1 by using $(X - E[X])^2$ as the random variable. The quantity $E[(X - E[X])^2]$ can be rewritten as $E[X^2] - E[X]^2$; this quantity is called the *variance* of X, written $\mathsf{var}(X)$ and its square root, often written $\sigma(X)$ is called its *standard deviation*. Thus Cor. 2 can be written

$$P(|X - E[X]| \geq a) \leq \frac{1}{a^2}\sigma(X)^2.$$

If X, Y are independent random variables then $\mathsf{var}(X + Y) = \mathsf{var}(X) + \mathsf{var}(Y)$.

[1] They must be measurable!

3 Limit theorems

The absolute miracle of randomness is that while a particular outcome of a random process is unpredictable the statistical behaviour of many *independent* experiments with the random process is very well behaved *in the limit*. This is what makes statistics, machine learning and indeed science possible. Gambling and lotteries are based on people failing to understand this.

We consider the following simple scenario. There is an experiment with two outcomes that we call *success* and *failure*. The probability of success is p and of failure is therefore $1-p$ or $\bar{p}$. We perform the same experiment n times and each instance is assumed to be *independent* and *identically distributed*, one often uses the abbreviation iid for this situation. Our probability space is the space of sequences of results. We write S_n for the random variable that gives the number of successes in a sequence of n experiments (or *trials* as they are called in the probability jargon). What is the behaviour of S_n as n tends towards infinity? In general one calls this kind of experiment a *Bernoulli trial*. A random variable X is said to have a Bernoulli distribution with parameter p if it takes on only the values 1 and 0 and $P(X=1)=p$.

The random variable S_n only takes on integer values and we have

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

This is the well known binomial distribution. If one has n iid Bernoulli random variables X_i with parameter p then the sum $\sum_i X_i$ is distributed according to the binomial distribution above. The following proposition follows by direct calculation and using the fact that the expectation and variance of a sum of n iid variables is n times the expectation and variance respectively of one of the variables.

Proposition 3.

$$E[S_n] = np \text{ and } \mathsf{var}(S_n) = np(1-p).$$

If we carry out the experiment n times and measure S_n then we get what is called the *empirical probability of success*: $\frac{S_n}{n}$. The actual (i.e. theoretical) probability of success should roughly be the same as the empirical probability. This is what the first of the famous limit theorems says. For the sample space formed from n trials we write the probability distribution as P_n.

Theorem 4 (Weak law of large numbers)**.** For every $\varepsilon > 0$ we have

$$\lim_{n \longrightarrow \infty} P_n \left(\left| \frac{S_n}{n} - p \right| > \varepsilon \right) = 0.$$

Proof From Prop. 3 we know that the mean of S_n is np and the variance of S_n is $np(1-p)$. So by the Tchebyshev-Bienaymé inequality Prop. 2 we have

$$P_n(|S_n - np| > n\varepsilon) \leq \frac{1}{(n\varepsilon)^2} \mathsf{var}(S_n) = \frac{1}{(n\varepsilon)^2}(np(1-p)) = \frac{1}{n\varepsilon^2} p(1-p) \leq \frac{1}{4n\varepsilon^2}.$$

The result is now immediate[2]. ∎

[2]The last inequality used is $p(1-p) \leq \frac{1}{4}$; you should prove it if it is not obvious.

The estimate we have from the proof above is that the fluctuations are bounded by $p(1-p)/(n\varepsilon)^2$.

This is a taste of a limit theorem. It is a very limited version of the law of large numbers stated for a very particular distribution. The result holds much more generally for sums of iid variables. There is a much stronger theorem called the *strong* law of large numbers which requires measure theory to be proved properly. This can be *greatly* improved; in fact the fluctuations are exponentially suppressed.

4 Some other basic inequalities

Theorem 5. Let $\{x_1, \ldots, x_n\}$ be a set of positive real numbers then

$$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \geq \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

This statement is called the arithmetic-geometric-mean inequality. It is the foundation of many inequalities and it is worth knowing how to prove it.

It is a purely elementary calculation when $n = 2$.
Lemma 6. For any real positive numbers x and y we have $\frac{1}{2} \cdot (x + y) \geq \sqrt{xy}$ with equality when $x = y$.

Proof

$$(\sqrt{x} - \sqrt{y})^2 \geq 0 \text{ hence } (x + y - 2\sqrt{xy}) \geq 0 \text{ or } \frac{1}{2} \cdot (x + y) \geq \sqrt{xy}.$$

If equality holds, it follows easily that $x = y$. ∎

The next proposition is the heart of the matter. Theorem 5 is an immediate consequence of it, since it says that the maximum value of the product is the arithmetic mean to the power n.
Proposition 7. Let $\{x_1, \ldots, x_n\}$ be a set of positive real numbers then with sum s. Then the maximum value of the product $p = \prod x_i$ is attained when the x_i are all equal.

Proof Suppose that not all the numbers are equal, then there must be two of them, say x_i and x_j that are different. We can replace them both by $\frac{1}{2}(x_i + x_j)$ without altering the sum s but by Lemma 6 the product is *strictly* increased. Thus, if we have attained the maximum value of p then all the x_i must be equal to $\frac{s}{n}$. ∎

We should prove that the maximum value of p is actually attained. This is an easy compactness argument. The set of tuples $(x_1, \ldots, x_n)$ satisfying the constraint $\sum x_i = s$ is a compact subset of $\mathbf{R}^n$ and p is a continuous function on this set, so the maximum value must be attained.
Theorem 8 (Bunyakowski-Cauchy-Schwartz)**.** Let V be a real inner product space, *i.e.* a real vector space with an inner product $\langle \cdot, \cdot \rangle$. Let $u, v \in V$ then

$$\langle u, v \rangle \leq \langle u, u \rangle^{\frac{1}{2}} \langle v, v \rangle^{\frac{1}{2}}.$$

Proof Note that the inequality is trivial if either vector is 0; so we will assume that both of them are nonzero. For any u, $\langle u, u\rangle \geq 0$. Now we apply this to $(u - v)$

$$\langle (u - v), (u - v)\rangle \geq 0 \text{ or } \langle u, v\rangle \leq \frac{1}{2}(\langle u, u\rangle + \langle v, v\rangle).$$

Since we have assumed that neither u nor v are zero we can define the unit vectors $\hat{u} = u/\sqrt{\langle u, u\rangle}$ and $\hat{v} = v/\sqrt{\langle v, v\rangle}$ which satisfy $\langle \hat{u}, \hat{u}\rangle = \langle \hat{v}, \hat{v}\rangle = 1$.

Applying the result above for u and v to $\hat{u}$ and $\hat{v}$ we have

$$\langle \hat{u}, \hat{v}\rangle \leq \frac{1}{2}(\langle \hat{u}, \hat{u}\rangle + \langle \hat{v}, \hat{v}\rangle = 1$$

or, substituting in the definition of the unit vectors and simplifying

$$\langle u, v\rangle \leq \sqrt{\langle u, u\rangle} \cdot \sqrt{\langle v, v\rangle}.$$

∎

Proving it for inner product spaces means we can use it equally easily for finite-dimensional vector spaces, for infinite-dimensional vector spaces, for function spaces with the inner product defined by an integral and for inner products defined by weighted sums instead of the standard inner product of $\mathbf{R}^n$.

Can you come up with 3 more different proofs? It's fun!