

Note 2: Tools for Probabilistic Analysis

1 Introduction

This note provides some facts and techniques that will be useful for analyzing learning algorithms. The note assumes some familiarity with basic concepts of probability theory. A good introduction to basic concepts and techniques for analysis of algorithms is given in [1] (Chapter 6 in first edition and Appendix C and Chapter 5 in second edition). Any probability text (e.g. [2]) should include the relevant material.

2 Basic Notions from Probability Theory

- Let X be a set of basic events called the sample space. An event is a subset of X . A probability distribution assigns probabilities to events, namely, to elements in $S \subseteq 2^X$.¹
- For example for normal dice we have a basic set of events $\{1, 2, 3, 4, 5, 6\}$ and any union of these is an event. For dice we usually have a uniform probability $Pr[A] = 1/6$ for each of the basic events. Some events are: $E_1 = \{1, 2, 3\}$, $E_2 = \{3, 4\}$ and $E_3 = \{1, 2, 3, 4\}$.
- We can also define distributions over non-countable domains. For example the uniform distribution over $[0, 1]$ assigns a probability of $b - a$ to interval $[a, b]$ where $0 \leq a \leq b \leq 1$. Note that for any single point $x = [x, x]$ the probability is 0. Some events are: $E_4 = [0.2, 0.6]$, $E_5 = [0.7, 0.8]$, and $E_6 = [0.4, 0.65]$.
- A probability distribution must satisfy the axioms of probability.
 1. For any event A , $Pr[A] \geq 0$
 2. $Pr[X] = 1$
 3. For any mutually exclusive events A and B , $Pr[A \cup B] = Pr[A] + Pr[B]$
 4. The same holds for any countable union of events (finite or infinite) $Pr[\cup A_i] = \sum_i Pr[A_i]$
- The following facts follow from the axioms. The complement $A^c = X \setminus A$ of an event A satisfies $Pr[A^c] = 1 - Pr[A]$. As a result we also get $Pr[\emptyset] = 0$.

For two events A and B we have in general

$$Pr[A \cup B] = Pr[A] + Pr[B] - Pr[A \cap B] \leq Pr[A] + Pr[B] \quad (1)$$

with equality if A and B are disjoint (mutually exclusive). The last inequality is called the union bound in our main textbook [4].

¹For discrete distributions we have $S = 2^X$. When X is uncountable some subsets may not be measurable (i.e. not assigned a probability). In general S must be closed under complement, countable union, and countable intersection. We can ignore this issue and assume a countable sample space for much of the course.

- Events A and B are statistically independent if $Pr[A \cap B] = Pr[A]Pr[B]$. This is normally applied to events that are known to be independent (a result of different independent experiments) and are therefore also statistically independent.
- The conditional probability of A given that B happened is $Pr[A|B] = \frac{Pr[A \cap B]}{Pr[B]}$. We therefore get Bayes' law $Pr[A|B] = \frac{Pr[B|A]Pr[A]}{Pr[B]}$.
- Using conditional probabilities we get an alternative definition for statistical independence: Events A and B are statistically independent if $Pr[A|B] = Pr[A]$.
- If E_1, \dots, E_n, \dots form a partition of the basic events (they are disjoint and together cover all events) then $Pr[A] = \sum Pr[E_i]Pr[A|E_i]$.
- Recall the events $E_1 = \{1, 2, 3\}$, $E_2 = \{3, 4\}$ and $E_3 = \{1, 2, 3, 4\}$ in our dice example. We have $Pr[E_2] = Pr[\{3\}] + Pr[\{4\}] = 2/6$ since they are disjoint, but $Pr[E_1 \cup E_2] = Pr[E_3] = 4/6 < Pr[E_1] + Pr[E_2] = 5/6$. Also $Pr[E_1|E_2] = \frac{Pr[\{3\}]}{Pr[E_2]} = \frac{1}{2}$.
- Recall the events $E_4 = [0.2, 0.6]$, $E_5 = [0.7, 0.8]$, $E_6 = [0.4, 0.65]$ for the uniform distribution. The event E_4 is disjoint from E_5 but not from E_6 . Also E_5 is disjoint from E_6 . For these events we have $Pr[E_4 \cup E_5] = Pr[E_4] + Pr[E_5] = 0.5$ and $Pr[E_4 \cup E_6] = Pr[E_4] + Pr[E_6] - Pr[E_4 \cap E_6] = 0.45$.
- Random variables map basic events to real numbers.
- A Bernoulli random variable is just like a coin flip; it maps a basic event to 1 (called success) or 0 (failure). Alternatively, we can define it using the set of basic events mapped to success. A single probability $p = Pr[\text{success}]$ characterizes its behavior. For example we can define a random variable x to be true iff event E_1 happens.
- Notions of independence can be extended to random variables in a natural way.
- The expectation of a continuous random variable x is $E[x] = \int_{-\infty}^{\infty} xPr[x]dx$ and for discrete variables $E[x] = \sum_v vPr[x = v]$. Note that expectation is a linear operator. That is, for constants a, b and random variables x, y , we have $E[ax + by] = aE[x] + bE[y]$. This follows from the definition (and holds even if x, y are not independent).

3 Computing Bounds on Probabilities of Events

We will often take repeated independent samples from the same distribution and measure the probability of an event happening in at least one of the draws. This corresponds to the probability of success in a sequence of independent Bernoulli trials. For example we could compute the probability q that E_1 happens at least once if we roll a die 10 times (each roll is independent of the others): $q = 1 - (1 - Pr[E_1])^{10} = 1 - (\frac{1}{2})^{10} = 0.99902$.

While here we would compute the probability exactly it is often sufficient to obtain a bound for the probability. The inequality

$$(1 - x) < e^{-x} \quad \text{if } 0 < x < 1 \quad (2)$$

provides us with such a bound. Applied to the previous computation it yields $q = 1 - (1 - Pr[E_1])^{10} > 1 - e^{-10Pr[E_1]} = 1 - e^{-5} = 0.99326$. We also have a lower bound for $(1 - x)$

$$(1 - x) > e^{\frac{-x}{1-x}} \geq e^{-2x} \quad \text{if } 0 < x \leq 1/2 \quad (3)$$

that yields $q = 1 - (1 - Pr[E_1])^{10} < 1 - e^{-20Pr[E_1]} = 1 - e^{-10} = 0.99995$. For a proof of Eq. (2) and (3) see [2]. Markov's Inequality provides another upper bound to probabilities:

Theorem 3.1 (*Markov's Inequality*) Let $x \geq 0$ be a positive random variable. Then for any $\alpha > 0$ $Pr[x \geq \alpha] \leq \frac{E[x]}{\alpha}$.

Proof:

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} xPr[x]dx = \int_0^{\infty} xPr[x]dx \\ &= \int_0^{\alpha} xPr[x]dx + \int_{\alpha}^{\infty} xPr[x]dx \\ &\geq \int_0^{\alpha} 0Pr[x]dx + \int_{\alpha}^{\infty} \alpha Pr[x]dx \\ &= \alpha \int_{\alpha}^{\infty} Pr[x]dx = \alpha Pr[x \geq \alpha] \end{aligned}$$

■

For bounded variables a lower bound can be obtained as well:

Theorem 3.2 Let $0 \leq x \leq M$ be a positive bounded random variable such that $E[x] \geq \alpha$. Then $Pr[x \geq \alpha/2] \geq \frac{\alpha}{2M}$.

Proof:

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} xPr[x]dx = \int_0^M xPr[x]dx \\ &= \int_0^{\alpha/2} xPr[x]dx + \int_{\alpha/2}^M xPr[x]dx \\ &\leq \int_0^{\alpha/2} \frac{\alpha}{2} Pr[x]dx + \int_{\alpha/2}^M MPr[x]dx \\ &\leq \frac{\alpha}{2} \cdot 1 + MPr[x \geq \alpha/2] \end{aligned}$$

Therefore

$$Pr[x \geq \alpha/2] \geq \frac{E[x] - \alpha/2}{M} \geq \frac{\alpha}{2M}$$

where the last inequality is true since we assumed that $E[x] \geq \alpha$.

■

Much tighter bounds can be derived for outcomes of independent Bernoulli trials.

Theorem 3.3 (*Chernoff Bound*) Let y_1, \dots, y_n be n independent Bernoulli random variables each with probability of success p and let $\hat{p} = \frac{1}{n} \sum y_i$, then for any $0 < \epsilon, \delta < 1$

(1) $Pr[|\hat{p} - p| \geq \epsilon] \leq 2e^{-2n\epsilon^2}$.

(2) for $n \geq \frac{1}{2\epsilon^2} \ln(\frac{2}{\delta})$, $Pr[|\hat{p} - p| \geq \epsilon] \leq \delta$.

Notice that part (2) follows from (1) by substitution. Our main text [4] states several other similar bounds. The Algorithms text [1] gives some ideas for the proof which can be found in various sources. (See also [3] for one of the original papers.) As part (2) indicates this bound allows us to estimate the probability of success of a Bernoulli variable with high accuracy and confidence by taking a large independent sample and computing \hat{p} . Notice however, that if we want the sample size n to be polynomial, the accuracy ϵ must be inverse polynomial while δ may be exponentially small.

4 Using the Bounds

Example 1: Let x be a Bernoulli variable with probability of success $p = 1/4$ and assume we make $n = 200$ independent trials with \hat{p} defined as above (so $E[\hat{p}] = p$).

Using Markov's Inequality we get $Pr[\hat{p} \geq 1/2] \leq (1/4)/(1/2) = 1/2$. Using Chernoff's bound we get $Pr[\hat{p} \geq 1/2] \leq Pr[|\hat{p} - p| \geq 1/4] \leq 2e^{-2 \cdot 200 \cdot (1/16)} = 2e^{-25} = 2.7 \cdot 10^{-11}$ so we see that the latter gives much sharper bounds. Nevertheless Markov's inequality is useful when we do not have Bernoulli trials. We can also use Theorem 3.2 to get $Pr[\hat{p} \geq 1/8] \leq (1/4)/(2 \cdot 1) = 1/8$. Normally, we will compute parameterized bounds on probabilities as in the translation from (1) to (2) in Chernoff's bound. ■

Example 2: You are given two coins (that you cannot distinguish in advance) and told that one of them has $p = 1/2$ and the other has $p = 1/4$. You are asked to identify the fair coin with probability at least $1 - \delta$ for a small $\delta < 1$ (say $\delta = 0.0001$). Can you design an algorithm to perform that?

The required algorithm picks one of the coins and tosses it n times computing \hat{p} as above. If $\hat{p} \geq 3/8$ it identifies that coin to be the fair coin; otherwise it is the other coin. Using Chernoff's bound we can find a bound on n so that we succeed with high enough probability. Assume that the coin picked for the experiment is the fair coin. Then

$$Pr[\text{coin rejected}] \leq Pr[|\hat{p} - p| > 1/8] \leq 2e^{-2n(1/64)} = 2e^{-n/32} \leq \delta$$

for $n \geq 32 \ln(\frac{2}{\delta})$. If the non-fair coin is chosen, a symmetric argument yields the same result so that either way $Pr[\text{coin rejected}] \leq \delta$. For $\delta = 0.0001$, $n \geq 317$ is sufficient. ■

Example 3: You are given k coins (that you cannot distinguish in advance) and told that one of them has $p \leq \alpha$. Nothing is known on the success probabilities of other coins. You are asked to identify a coin with $p \leq 2\alpha$. This may indeed be the coin you know has a small probability but may be any other coin satisfying this condition. Again you are required to succeed with probability at least $1 - \delta$. Can you design an algorithm to perform that?

The required algorithm tosses each coin n times computing \hat{p} for each coin. It chooses the coin with the smallest \hat{p} . While the details are more involved the idea in the argument is similar to the previous example. Say that a coin is bad if it has $p > 2\alpha$. The algorithm should not choose a bad coin. We can choose n large enough so that by Chernoff's bound (with high probability) we have $|\hat{p} - p| < \epsilon = \alpha/3$. If this holds then we know that the coin with $p \leq \alpha$ will have $\hat{p} \leq 4\alpha/3$. On the other hand, any bad coin will have $\hat{p} \geq 5\alpha/3$. So in this case we are guaranteed that a bad coin will not be chosen. It remains to find a value of n that guarantees that with probability $1 - \delta$ all estimates are within ϵ . We have:

$$Pr[\text{at least one estimate fails}] \leq kPr[\text{one estimate fails}] \leq 2ke^{-2n(\alpha^2/9)} \leq \delta$$

for $n \geq \frac{9}{2\alpha^2} \ln(\frac{2k}{\delta})$. Notice that since k is only included in the $\ln()$, this argument shows that we could estimate the probabilities for exponentially many events. (Of course in the current example we would have had to toss these coins so it is not relevant here, but it can sometimes be used by implicitly making these estimates.) ■

Example 4: You are interested in solving a certain decision problem (e.g. is the input number n a prime or not) and told that there is an algorithm with the following property:

for any prime number n the algorithm always correctly reports that it is prime

for any composite number n , with probability at least $3/4$ the algorithm correctly reports that it is composite.

Now you are given a failure parameter $\delta < 1/4$. Can you design an algorithm with similar properties but with $3/4$ replaced by $1 - \delta$?

Example 5: You are still interested in the decision problem but this time you are supplied with an algorithm with a slightly weaker property:

for any prime number n with probability at least $3/4$ the algorithm correctly reports that it is prime

for any composite number n , with probability at least $3/4$ the algorithm correctly reports that it is composite.

So the error is symmetric. Now you are given a failure parameter $\delta < 1/4$. Can you design an algorithm with similar properties but with $3/4$ replaced by $1 - \delta$?

5 Usage in PAC Learning

Examples in the PAC model are independently sampled from some distribution D over the domain X . The arguments above are therefore very useful in this setting. While there is no single recipe that will fit all problems (and all the tools above will be used in the course) the following are standard arguments:

- We often want to verify that none of several bad events happens. In this case the union bound (Eq. 1) can be used (cf. the use of k in Example 3).
- Sometimes we have a test that we want to perform but it only succeeds with some probability α . By repeating the test n times (independently) we can increase this probability considerably (using Eq. 2). This yields $\Pr[\text{success}] = 1 - (1 - \alpha)^n \geq 1 - e^{-n\alpha}$.
- **Hypothesis testing:** Let c be the target concept being learned, and suppose we somehow computed a hypothesis h and would like to know whether it is a good approximation of c . Let $error(h) = \Pr_D[h(x) \neq c(x)]$. Since for a random example x , $[h(x) \neq c(x)]$ is a Bernoulli variable with success probability $error(h)$, the use of Chernoff's bound in Example 3 demonstrates that we can estimate $error(h)$ up to an inverse polynomial.
- In a similar manner, whenever we have a gap between two probabilities for an event (bad if $> 2\gamma$, good if $< \gamma$) we can test with high probability which of the two is the case. Often we generate situations with such gaps in order to utilize this tool.

References

- [1] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press/McGraw-Hill, 1990.
- [2] William Feller. *An Introduction to Probability and its Applications*, volume 1. John Wiley and Sons, third edition, 1968.
- [3] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.
- [4] M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, 1994.