CS724: Topics in Algorithms
Probability Review
Slide Set 10

Prof. Dan A. Simovici

Random Variables

Experiments and Random Variables

The result of an experiment whose outcome is determined by chance is described by a *random variable* that gives a numerical value that is the outcome of the experiment.
The *sample space* is the set of all possible outcomes of the experiment.

If the sample space Ω is finite we can assign a probability to each outcome of the experiment.

Example

If we throw a die, the sample space is finite: $\{1, 2, 3, 4, 5, 6\}$ and the probability of each outcome is $\frac{1}{6}$.

In some cases, if the sample space is infinite a probability can be assigned to each outcome.

Example

If the sample space Ω is the set of positive integers at least equal to 1, due to Euler formula, $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$, a probability distribution function $p(i) = \frac{6}{\pi} \frac{1}{i^2}$ can be assigned to each i.

Discrete Probability Distributions

A random variable X on a sample space Ω is a real-valued function on Ω, $X : \Omega \to \mathbb{R}$.
A discrete random variable is a random variable that takes only a finite, or an infinitely countable number of values.

For a discrete random variable X, the notation "$X = a$" includes all elementary events for which the random variable X takes the value a.

Example

If X is the random variable representing the sum of two dice, then the event $X = 5$ corresponds to the elementary events $(1, 4), (2, 3), (3, 2), (4, 1)$ and $P(X = 5) = \frac{4}{36}$.

UMASS BOSTON

Definition

Let Y be a random variable such that

$$Y = \begin{cases} 1 & \text{if an experiment succeeds,} \\ 0 & \text{otherwise.} \end{cases}$$

Y is called a *Bernoulli variable*, or an *indicator variable*.

Definition

The expectation $E[x]$ of a discrete random variable X is

$$E[X] = \sum_i iP(X = i).$$

Example

The expectation of the discrete random variable representing the sum of two dice is:

$$E[X] = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \cdots + \frac{1}{36} \cdot 12 = 7$$

Example

Let X be the discrete random variable that takes the value 2^i with probability $\frac{1}{2^i}$ for $i = 1, 2, \ldots$. The expected value is

$$E[X] = \sum_{i=1}^{\infty} 2^i \cdot \frac{1}{2^i} = \infty.$$

Definition

The random variables X and Y are independent if

$$P((X = x) \cap (Y = y)) = P(X = x)P(Y = y)$$

for all values x and y. This extends to n random variables: $X_1, \ldots, X_n$ are independent if for all $x_1, \ldots, x_n$ we have

$$P((X_1 = x_1) \cap (X_2 = x_2) \cap \cdots \cap (X_n = x_n)) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n$$

Linearity of Expectation

Theorem

For any finite collection of discrete random variables $X_1, \ldots, X_n$ we have

$$E[X_1 + \cdots + X_n] = \sum_{i=1}^{n} E[X_i].$$

For any constant c and random variable X we have $E[cX] = cE[X]$.

Markov and Cebyshev Inequalities

Markov's Inequality

Theorem

Let X be a discrete random variable that takes non-negative values:

$$X : \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix},$$

where $0 \leqslant a_1 < a_2 < \cdots < a_n$ *and* $p_1 + p_2 + \cdots + p_n = 1$.
We have

$$P(X > a) \leqslant \frac{E[X]}{a}.$$

(Markov's Inequality)

Proof

If $a_{j-1} < a \leqslant a_j$, then

$$\begin{aligned} E[X] &= a_1 p_1 + \cdots + a_{j-1} p_{j-1} + a_j p_j + \cdots + a_n p_n \\ &\geqslant a_j p_j + \cdots + a_n p_n \geqslant a(p_j + \cdots + p_n) \\ &= aP(X > a). \end{aligned}$$

Therefore,

$$P(X > a) \leqslant \frac{E[X]}{a}.$$

Variance of a Random Variable

Definition

Let X be a random variable and let $Y = (X - E[X])^2$ be a non-negative random variable that depends on X. The number $E[Y]$ is the variance of X and is denoted by $Var(X)$.

Note that
$Var(X) = E[(X - E[X])^2] = E[X^2 - 2E(X)X + E[X]^2] = E[X^2] - (E[X])^2$.

Cebyshev Inequality

Theorem

Let X be a discrete random variable:

$$X : \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix},$$

where $a_1 < a_2 < \cdots < a_n$ and $p_1 + p_2 + \cdots + p_n = 1$. We have

$$P(|X - E(X)| \geqslant a) \leqslant \frac{Var(X)}{a^2}$$

Proof

Let $Y = (X - E[X])^2$ be a non-negative random variable. By applying Markov's inequality to Y we obtain $P(Y \geqslant a^2) \leqslant \frac{E[Y]}{a^2}$. Note that $E[Y] = var(X)$, and therefore

$$P(|X - E(X)| \geqslant a) = P(Y \geqslant a^2) \leqslant \frac{E[y]}{a^2},$$

which implies

$$P(|X - E(X)| \geqslant a) \leqslant \frac{Var(X)}{a^2}.$$

Definition

The cumulative distribution function of a random variable X is the function $F_X : \mathbb{R} \longrightarrow [0,1]$ defined as

$$F_X(x) = P(X \leqslant x)$$

for every $x \in \mathbb{R}$.

Continuous Probability Distributions

Discrete random variables take a countable number of values.
Continuous random variables have a range that is an interval, or a union of non-overlaping intervals (that can be the entire set $\mathbb{R}$).
The theory of contiuous random variables is similar to the theory of discrete radom variables:

- sums are replaced by integrals;
- probability mass functions are replaced by probability density functions (pdfs).

Example

For the uniform distribution on $[0, 1]$ the probability of any particular value is 0. However, in this case we can define a probability density function $p(x)$, where

$$p(x) = \begin{cases} 1 & \text{if } 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

This implies $P(a < x < b) = \int_a^b p(x)\, dx = \frac{1}{b-a}$.

The Gaussian or Normal Distribution

The normal distribution (or the Gaussian distribution) with mean m and variance σ^2 is defined by the probability density:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

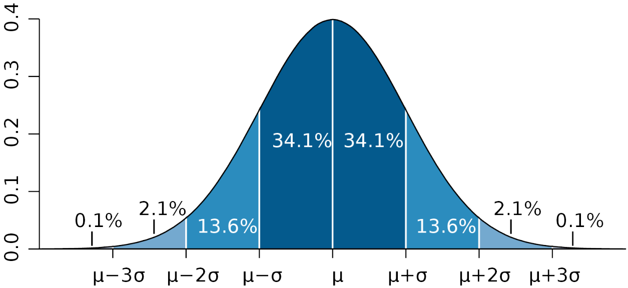
For mean $m = 0$ and $\sigma = 1$ the density becomes

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Standard tables give the values of the integral

$$\int_0^t \phi(x)\, dx = F(t).$$

A one-dimensional Gaussian has its probability mass close to origin.



The Central Limit Theorem

Let $S = X_1 + X_2 + \cdots + X_n$ be a sum of n independent random variables where

$$X_i = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{1}{2}. \end{cases}$$

The expected value $E(X_i)$ of each variable X_i is $1/2$ with variance

$$\sigma_i^2 = (1/2 - 0)^2 1/2 + (1/2 - 1)^2 1/2 = 1/4$$

The expected value of S is $n/2$, and because the variables $X_1, \ldots, X_n$ are independent, the variance of s is the sum of the variances of x_is, so it is $\frac{n}{4}$. Thus, the standard deviation of S is $\frac{\sqrt{n}}{2}$.

The Central Limit Theorem cont'd

Note that

$$\lim_{n\to\infty} \frac{\frac{\sqrt{n}}{2}}{n} = 0.$$

In general, if $X_1, \ldots, X_n$ are independent and identically distributed each with standard deviation σ, then the standard deviation of $X_1 + \cdots + X_n$ is $\sqrt{n}\sigma$, so the random variable $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ has standard deviation σ. A stronger assertion is included next.

Theorem (Central Limit Theorem)

If $X_1, \ldots, X_n$ is a sequence of identically distributed independent random variables each with mean μ and variance σ^2, the distribution of the random variable

$$\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n - n\mu)$$

converges to the Gaussian distribution with mean 0 and variance σ^2.

Theorem

Let $X_1, \ldots, X_n$ be n mutually independent random variables with 0 mean and variance at most σ^2. Suppose that:

$$a \in [0, \sqrt{2} n\sigma^2], s \textit{ is a positive even integer and } s \leqslant \frac{n\sigma^2}{2},$$

and

$$E(X_i^r) \leqslant \sigma^2 r! \textit{ for } 3 \leqslant r \leqslant s.$$

Then,

$$P(|X_1 + \cdots + X_n| \geqslant a) \leqslant \left(\frac{2sn\sigma^2}{a^2}\right)^{\frac{s}{2}}.$$

If, further $s \geqslant \frac{a^2}{4n\sigma^2}$, we also have:

$$P(|X_1 + \cdots + X_n| \geqslant a) \leqslant 3e^{-\frac{a^2}{12n\sigma^2}}.$$

UMASS BOSTON

Proof

We prove first an upper bound on $E(X^r)$ for any positive even r.

$$
\begin{aligned}
(X_1 + \cdots + X_n)^r &= \sum \binom{r}{r_1 \ldots r_n} X_1^{r_1} \cdots X_n^{r_n} \\
&= \sum \frac{r!}{r_1! \cdots r_n!} X_1^{r_1} \cdots X_n^{r_n}.
\end{aligned}
$$

This implies

$$
E(X^r) = \sum \frac{r!}{r_1! \cdots r_n!} E(X_1^{r_1}) \cdots E(X_n^{r_n}).
$$

For those terms where $r_i = 1$, the term is 0 because $E(X_i) = 0$. Thus, we can assume that $(r_1, \ldots, r_n)$ runs over sets of non-zero r_i having the sum r, where each non-zero r_i is at least 2. Thus, there are at most $r/2$ non-zero r_i in each set.

Since $E(X_i^{r_i}) \leqslant \sigma^2 r_i!$ it follows that

$$E(X^r) \leqslant r! \sum_{r_1,\ldots,r_n} \sigma^{2\text{number of non-zero } r_i \text{ in the set}}.$$

Collect terms of the summation with t non-zero r_is for $1 \leqslant t \leqslant \frac{r}{2}$. There are $\binom{n}{t}$ subsets of $\{1, \ldots, n\}$ of cardinality t. Once a subset is fixed as the set of t values of i with non-zero r_i, set each of $r_i \geqslant 2$.

That is, allocate two of each r_i and the alocate the remaining $r - 2t$ to tr_i arbitrarily. The number of such allocations is
$\binom{r-2t+t-1}{t-1} = \binom{r-t+1}{t-1}$.
Let $f(t) = \binom{n}{t}\binom{r-t-1}{t-1}\sigma^{2t}$. We have

$$E(X^r) \leqslant r! \sum_{t=1}^{r/2} f(t).$$

Thus $f(t) \leqslant h(t)$, where

$$h(t) = \frac{(n\sigma^2)^t}{t!} 2^{r-t-1}.$$

Since $r \leqslant r/2 \leqslant \frac{n\sigma^2}{4}$, we have:

$$\frac{h(t)}{h(t-1)} = \frac{n\sigma^2}{2t} \geqslant 2.$$

Thus, we obtain

$$E(x^r) = r! \sum_{t=1}^{r/2} \leqslant r! h(r/2) \left(1 + \frac{1}{2} + \frac{1}{4} + \cdots \right) \leqslant \frac{r!}{(r/2)!} 2^{r/2} (n\sigma^2)^{\frac{r}{2}}.$$

By Markov's Inequality we have

$$P(|X| > a) = P(|X|^r > a^r) \leqslant g(r),$$

where $g(r) = \frac{r!(n\sigma^2)^{r/2} 2^{r/2}}{(r/2)! a^r} \leqslant \left(\frac{2rn\sigma^2}{a^2} \right)^{r/2}$. This holds for $r \leqslant s$, r even, and applying for $r = s$ we get the first inequality of the theorem.

For the second inequality, note that for even r,

$$\frac{g(r)}{g(r-2)} = \frac{4(r-1)n\sigma^2}{a^2} = \frac{r-1}{\frac{a^2}{4n\sigma^2}}.$$

Thus, $g(r)$ decreases as long as $r - 1 \leqslant \frac{a^2}{4n\sigma^2}$. Taking r to be the largest even integer less of equal to $\frac{a^2}{6n\sigma^2}$, the tail probability is at most $e^{r/2}$, which is the most $e \cdot e^{-\frac{a^2}{12n\sigma^2}} \leqslant 3e^{-\frac{a^2}{12n\sigma^2}}$.

As dimension increases, the behavior of d-dimensional Gaussian random variables is changing.

A *d-dimensional spherical Gaussian* variable with 0 mean and variance σ^2 in each coordinate has the density function

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{\|\boldsymbol{x}\|^2}{2\sigma^2}}.$$

The next theorem states that nearly all probability of a spherical Gaussian is concentrated in a thin annulus.

Theorem (Gaussian Annulus Theorem)

For a d-dimensional spherical Gaussian with variance 1 in every direction and for any $\beta \leqslant \sqrt{d}$, there is a fixed positive constant c such that all but at most $3e^{-c\beta^2}$ of the probability lies within the annulus:

$$\sqrt{d} - \beta \leqslant \| \mathbf{x} \| \leqslant \sqrt{d} + \beta.$$

Proof

Note that

$$E(\| \boldsymbol{x} \|^2) = \sum_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d,$$

so the mean square distance of a point to the center is d.
Let $\boldsymbol{x} = (x_1, \ldots, x_d)$ be a point selected from a unit variance Gaussian centered at the origin and let $r = \| \boldsymbol{x} \|$.
The inequality $\sqrt{d} - \beta \leqslant \| \boldsymbol{x} \| \leqslant \sqrt{d} + \beta$ is equivalent to $|r - \sqrt{d}| \geqslant \beta$.
Multiplying with $r + \sqrt{d}$ yields $|r^2 - d| \geqslant \beta(r + \sqrt{d}) \geqslant \beta\sqrt{d}$. To prove the theorem it suffices to bound the probability that $|r^2 - d| \geqslant \beta\sqrt{d}$.

Proof cont'

Note that

$$r^2 - d = (x_1^2 + \cdots + x_d^2) - d = (x_1^2 - 1) + \cdots + (x_d^2 - 1).$$

By introducing new varibles $y_i = x_i^2 - 1$ we can bound the probability that $|y_1 + \cdots + y_d| \geqslant \beta\sqrt{d}$. Note that $E(y_i) = E(x_i^2) - 1 = 0$. To apply the previous theorem we need to bind the s^{th} moment of y_i.

Proof cont'

For $|x_i| \leqslant 1$, we have $|y_i|^s \leqslant 1$ and for $|x_i| \geqslant 1$, $|y_i|^s \leqslant |x_i|^{2s}$. Therefore,

$$\begin{aligned} |E(y_i^s)| &= E(|y_i|^s) \leqslant E(1 + x_i^{2s}) = 1 + E(x_i^{2s}) \\ &= 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-\frac{x^2}{2}}. \end{aligned}$$

Using the substitution $2z = x^2$ we have

$$|E(y_i^s)| = 1 + \frac{1}{\sqrt{\pi}} \int_0^\infty 2^s z^{s-\frac{1}{2}} e^{-z} \, dz \leqslant 2^s s!.$$

Since $E(y_i) = 0$, $Var(y) = E(y_i^2) \leqslant 2^2 \cdot 2 = 8$. We need however, $E(y_i^s) \leqslant 8s!$.

Now we apply a change of variable $y_i = 2w_i$. Then $Var(w_i) \leqslant 2$ and $E(w_i^s) \leqslant 2s!$ and we need to bound the probability that $|w_1 + \cdots + w_d| \geqslant \frac{\beta\sqrt{d}}{2}$. Applying the previous theorem with $\sigma^2 = 2$ and $n = d$ yields the desired result.