CS-E4600
Basic concepts on discrete probability
slide set 6

Aristides Gionis
Department of Computer Science
Aalto University

reading assignment

- your favorite book on probability, computing, and randomized algorithms, e.g.,
- Randomized algorithms, Motwani and Raghavan (chapters 3 and 4)

  or
- Probability and computing, Mitzenmacher and Upfal (chapters 2, 3 and 4)

Events and probability

- consider a random process
  (e.g., throw a die, pick a card from a deck)
- each possible outcome is a simple event (or sample point)
- the sample space is the set of all possible simple events.
- an event is a set of simple events
  (a subset of the sample space)
- with each simple event E we associate a real number

$$0 \leq \Pr[E] \leq 1$$

which is the probability of E

Probability spaces and probability functions

- sample space Ω: the set of all possible outcomes of the random process
- family of sets $\mathcal{F}$ representing the allowable events:
  each set in $\mathcal{F}$ is a subset of the sample space Ω
- a probability function $\Pr : \mathcal{F} \to \mathbb{R}$ satisfies the following conditions

  1. for any event E, $0 \leq \Pr[E] \leq 1$
  2. $\Pr[\Omega] = 1$
  3. for any finite (or countably infinite) sequence of pairwise mutually disjoint events $E_1, E_2, \ldots$

$$\Pr\left[\bigcup_{i \geq 1} E_i\right] = \sum_{i \geq 1} \Pr[E_i]$$

the union bound

- for any events $E_1, E_2, \ldots, E_n$

$$\Pr\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} \Pr[E_i]$$

conditional probability

- the conditional probability that event E occurs given that event F occurs is

$$\Pr[E \mid F] = \frac{\Pr[E \cap F]}{\Pr[F]}$$

- well-defined only if $\Pr[F] > 0$
- we restrict the sample space to the set F
- thus we are interested in $\Pr[E \cap F]$ "normalized" by $\Pr[F]$

independent events

- two events E and F are independent if and only if

$$\Pr[E \cap F] = \Pr[E]\Pr[F]$$

equivalently if and only if

$$\Pr[E \mid F] = \Pr[E]$$

conditional probability

$$\Pr[E_1 \cap E_2] = \Pr[E_1] \Pr[E_2 \mid E_1]$$

generalization for k events $E_1, E_2, \ldots, E_k$

$$\Pr[\cap_{i=1}^{k} E_i] = \Pr[E_1] \Pr[E_2 \mid E_1] \Pr[E_3 \mid E_1 \cap E_2] \ldots \Pr[E_k \mid \cap_{i=1}^{k-1} E_i]$$

birthday paradox

E_i: the i-th person has a different birthday than all $1, \ldots, i-1$ persons (consider n-day year)

$$\begin{aligned}
\Pr[\cap_{i=1}^{k} E_i] &= \Pr[E_1]\Pr[E_2 \mid E_1] \ldots \Pr[E_k \mid \cap_{i=1}^{k-1} E_i] \\
&\leq \prod_{i=1}^{k} \left(1 - \frac{i-1}{n}\right) \\
&\leq \prod_{i=1}^{k} e^{-(i-1)/n} \\
&= e^{-k(k-1)/2n}
\end{aligned}$$

for k equal to about $\sqrt{2n} + 1$ the probability is at most $1/e$

as k increases the probability drops rapidly

birthday paradox

E_i: the i-th person has a different birthday than all $1, \ldots, i-1$ persons (consider n-day year)

$$
\begin{aligned}
\Pr[\cap_{i=1}^{k} E_i] &= \Pr[E_1] \Pr[E_2 \mid E_1] \ldots \Pr[E_k \mid \cap_{i=1}^{k-1} E_i] \\
&\leq \prod_{i=1}^{k} \left(1 - \frac{i-1}{n}\right) \\
&\leq \prod_{i=1}^{k} e^{-(i-1)/n} \\
&= e^{-k(k-1)/2n}
\end{aligned}
$$

for k equal to about $\sqrt{2n} + 1$ the probability is at most $1/e$

as k increases the probability drops rapidly

random variable

- a random variable X on a sample space Ω is a function $X : \Omega \to \mathbb{R}$
- a discrete random variable takes only a finite (or countably infinite) number of values

random variable — example

- from birthday paradox setting:
- E_i: the i-th person has a different birthday than all $1, \ldots, i-1$ persons
- define the random variable

$$X_i = \begin{cases} 1 & \text{the } i\text{-th person has different birthday} \\ & \text{than all } 1, \ldots, i-1 \text{ persons} \\ 0 & \text{otherwise} \end{cases}$$

expectation and variance of a random variable

- the expectation of a discrete random variable X, denoted by $E[X]$, is given by

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

where the summation is over all values in the range of X

- variance

$$\mathrm{Var}[X] = \sigma_X^2 = E[(X - E[X])^2] = E[(X - \mu_X)^2]$$

linearity of expectation

- for any two random variables X and Y

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

- for a constant c and a random variable X

$$E[cX] = c\,E[X]$$

coupon collector's problem

- n types of coupons
- a collector picks coupons
- in each trial a coupon type is chosen at random
- how many trials are needed, in expectation,
  until the collector gets all the coupon types?

coupon collector's problem — analysis

- let $c_1, c_2, \ldots, c_X$ the sequence of coupons picked
- $c_i \in \{1, \ldots, n\}$
- call c_i success if a new coupon type is picked
- (c_1 and c_X are always successes)
- divide the sequence in epochs: the i-th epoch starts after the i-th success and ends with the $(i+1)$-th success
- define the random variable X_i = length of the i-th epoch
- easy to see that

$$X = \sum_{i=0}^{n-1} X_i$$

coupon collector's problem — analysis (cont'd)

probability of success in the i-th epoch

$$p_i = \frac{n-i}{n}$$

(X_i geometrically distributed with parameter p_i)

$$E[X_i] = \frac{1}{p_i} = \frac{n}{n-i}$$

from linearity of expectation

$$E[X] = E\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} E[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n$$

where H_n is the harmonic number, asymptotically equal to $\ln n$

deviations

- inequalities on tail probabilities
- estimate the probability that
  a random variable deviates from its expectation

Markov inequality

- let X a random variable taking non-negative values
- for all $t > 0$

$$\Pr[X \geq t] \leq \frac{E[X]}{t}$$

or equivalently

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

Markov inequality — proof

- it is $E[f(X)] = \sum_x f(x) \Pr[X = x]$
- define $f(x) = 1$ if $x \geq t$ and 0 otherwise
- then $E[f(X)] = \Pr[X \geq t]$
- notice that $f(x) \leq x/t$ implying that

$$E[f(X)] \leq E\left[\frac{X}{t}\right]$$

- putting everything together

$$\Pr[X \geq t] = E[f(X)] \leq E\left[\frac{X}{t}\right] = \frac{E[X]}{t}$$

Chebyshev inequality

- let X a random variable with expectaction μ_X and standard deviation σ_X

- then for all $t > 0$

$$\Pr[|X - \mu_X| \geq t\sigma_X] \leq \frac{1}{t^2}$$

Chebyshev inequality — proof

- notice that

$$\Pr[|X - \mu_X| \geq t\sigma_X] = \Pr[(X - \mu_X)^2 \geq t^2\sigma_X^2]$$

- the random variable $Y = (X - \mu_X)^2$ has expectation σ_X^2
- apply the Markov inequality on Y

Chernoff bounds

- let $X_1, \ldots, X_n$ independent Poisson trials
- $\Pr[X_i = 1] = p_i$ (and $\Pr[X_i = 0] = 1 - p_i$)
- define $X = \sum_i X_i$, so $\mu = E[X] = \sum_i E[X_i] = \sum_i p_i$
- for any $\delta > 0$

$$\Pr[X > (1+\delta)\mu] \leq e^{-\frac{\delta^2 \mu}{3}}$$

and

$$\Pr[X < (1-\delta)\mu] \leq e^{-\frac{\delta^2 \mu}{2}}$$

Chernoff bound — proof idea

- consider the random variable e^{tX} instead of X
  (where t is a parameter to be chosen later)
- apply the Markov inequality on e^{tX} and work with $E[e^{tX}]$
- $E[e^{tX}]$ turns into $E[\prod_i e^{tX_i}]$, which turns into $\prod_i E[e^{tX_i}]$, due to independence
- calculations, and pick a t that yields the most tight bound

optional homework: study the proof by yourself

Chernoff bound — example

- n coin flips
- $X_i = 1$ if i-th coin flip is H and 0 if T
- $\mu = n/2$
- pick $\delta = \frac{2c\sqrt{n}}{n}$
- then $e^{-\frac{\delta^2 \mu}{2}} = e^{-\frac{4c^2 \cdot n \cdot n}{n^2 \cdot 2 \cdot 2}} = e^{-c^2}$ drops very fast with c
- so

$$\Pr[X < \frac{n}{2} - c\sqrt{n}] = \Pr[X < (1-\delta)\mu] \leq e^{-\frac{\delta^2 \mu}{3}} = e^{-c^2}$$

- and similarly with $e^{-\frac{\delta^2 \mu}{3}} = e^{-2c^2/3}$
- so, the probability that the number of H's falls outside the range $[\frac{n}{2} - c\sqrt{n}, \frac{n}{2} + c\sqrt{n}]$ is very small

Johnson Lindenstrauss lemma

- for any set X of n vectors in $\mathbb{R}^d$, and $\epsilon > 0$
- there exist $k = \Omega(\frac{\log n}{\epsilon^2})$ and linear map $f : \mathbb{R}^d \to \mathbb{R}^k$
- such that for all $x, y \in X$

$$(1-\epsilon)||u-v||^2 \leq ||f(x) - f(y)||^2 \leq (1+\epsilon)||u-v||^2$$

Johnson Lindenstrauss lemma — proof idea

1. consider projecting a fixed unit vector of $\mathbb{R}^d$ in a random subspace of $\mathbb{R}^k$
2. this is equivalent as projecting a random unit vector of $\mathbb{R}^d$ to its k first coordinates
3. expected length is the projection is k/d
4. can show that if $k = \Omega(\frac{\log n}{\epsilon^2})$ the probability of deviating from the expectation by $(1 \pm \epsilon)$ is at most $O(1/n^2)$
5. (why?) Chernoff-like bound: each of the k first coordinates is independently chosen from the normal distribution
6. using the union bound can show that all with high probability the distortion is small for all pairs of vectors