Lecture 6: “Coupon Collector’s problem”

Sotiris Nikoletseas
Professor

CEID - ETY Course
2017 - 2018

Variance: key features

- Definition:
$$Var(X) = E[(X-\mu)^2] = \sum_x (x-\mu)^2 \Pr\{X=x\}$$
where $\mu = E[X] = \sum_x x \Pr\{X=x\}$

- We call standard deviation of X the $\sigma = \sqrt{Var(X)}$

- Basic Properties:
  - (i) $Var(X) = E[X^2] - E^2[X]$
  - (ii) $Var(cX) = c^2 Var(X)$, where c constant.
  - (iii) $Var(X+c) = Var(X)$, where c constant.

- proof of (i):
$$Var(X) = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] + E[-2\mu X] + E[\mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - \mu^2$$

On the Additivity of Variance

- In general the variance of a sum of random variables is not equal to the sum of their variances
- However, variances do add for independent variables (i.e. mutually independent variables). Actually pairwise independence suffices.

Conditional distributions

- Let X, Y be discrete random variables. Their joint probability density function is
$$f(x,y) = \Pr\{(X=x) \cap (Y=y)\}$$

- Clearly $f_1(x) = \Pr\{X = x\} = \sum_y f(x,y)$
and $f_2(y) = \Pr\{Y = y\} = \sum_x f(x,y)$

- Also, the conditional probability density function is:
$$f(x|y) = \Pr\{X = x | Y = y\} = \frac{\Pr\{(X=x) \cap (Y=y)\}}{\Pr\{Y=y\}} =$$
$$= \frac{f(x,y)}{f_2(y)} = \frac{f(x,y)}{\sum_x f(x,y)}$$

Pairwise independence

- Let random variables $X_1, X_2, \ldots, X_n$. These are called pairwise independent iff for all $i \neq j$ it is
$$\Pr\{(X_i = x)|(X_j = y)\} = \Pr\{X_i = x\}, \forall x, y$$
Equivalently, $\Pr\{(X_i = x) \cap (X_j = y)\} =$
$$= \Pr\{X_i = x\} \cdot \Pr\{X_j = y\}, \forall x, y$$
- Generalizing, the collection is k-wise independent iff, for every subset $I \subseteq \{1, 2, \ldots, n\}$ with $|I| < k$ for every set of values $\{a_i\}, b$ and $j \notin I$, it is
$$\Pr\left\{X_j = b | \bigwedge_{i \in I} X_i = a_i\right\} = \Pr\{X_j = b\}$$

Mutual (or "full") independence

- The random variables $X_1, X_2, \ldots, X_n$ are mutually independent iff for any subset $X_{i_1}, X_{i_2}, \ldots, X_{i_k}, (2 \le k \le n)$ of them, it is $\Pr\{(X_{i_1} = x_1) \cap (X_{i_2} = x_2) \cap \cdots \cap (X_{i_k} = x_k)\} =$ $= \Pr\{X_{i_1} = x_1\} \cdot \Pr\{X_{i_2} = x_2\} \cdots \Pr\{X_{i_k} = x_k\}$

- Example (for $n = 3$). Let A_1, A_2, A_3 3 events. They are mutually independent iff all four equalities hold:

$$\Pr\{A_1 A_2\} = \Pr\{A_1\} \Pr\{A_2\} \tag{1}$$
$$\Pr\{A_2 A_3\} = \Pr\{A_2\} \Pr\{A_3\} \tag{2}$$
$$\Pr\{A_1 A_3\} = \Pr\{A_1\} \Pr\{A_3\} \tag{3}$$
$$\Pr\{A_1 A_2 A_3\} = \Pr\{A_1\} \Pr\{A_2\} \Pr\{A_3\} \tag{4}$$

They are called pairwise independent if (1), (2), (3) hold.

The Coupon Collector's problem

- There are n distinct coupons and at each trial a coupon is chosen uniformly at random, independently of previous trials.
- Let m the number of trials.
- Goal: establish relationships between the number m of trials and the probability of having chosen each one of the n coupons at least once.

Note: the problem is similar to occupancy (number of balls so that no bin is empty).

The expected number of trials needed (I)

- Let X the number of trials (a random variable) needed to collect all coupons at least once each.
- Let $C_1, C_2, \ldots, C_X$ the sequence of trials, where $C_i \in \{1, \ldots, n\}$ denotes the coupon type chosen at trial i. We call the ith trial a success if coupon type chosen at C_i was not drawn in any of the first $i-1$ trials (obviously C_1 and C_X are always successes).
- We divide the sequence of trials into epochs, where epoch i begins with the trial following the ith success and ends with the trial at which the $(i+1)$st success takes place. Let r.v. $X_i (0 \leq i \leq n-1)$ be the number of trials in the ith epoch.

The expected number of trials needed (II)

- Clearly, $X = \sum_{i=0}^{n-1} X_i$
- Let p_i the probability of success at any trial of the ith epoch. This is the probability of choosing one of the $n-i$ remaining coupon types, so:
$$p_i = \frac{n-i}{n}$$
- Clearly, X_i follows a geometric distribution with parameter p_i, so
$E[X_i] = \frac{1}{p_i}$ and $Var(X_i) = \frac{1-p_i}{p_i^2}$
- By linearity of expectation:
$E[X] = E\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} E[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n\sum_{i=1}^{n} \frac{1}{i} =$
$= nH_n$
But $H_n \sim \ln n + \Theta(1) \Rightarrow E[X] \sim n \ln n + \Theta(n)$

The variance of the number of needed trials

- Since the X_i's are independent, we have:

  $$Var(X) = \sum_{i=0}^{n-1} Var(X_i) = \sum_{i=0}^{n-1} \frac{ni}{(n-i)^2} = \sum_{i=1}^{n} \frac{n(n-i)}{i^2} =$$

  $$= n^2 \sum_{i=1}^{n} \frac{1}{i^2} - n \sum_{i=1}^{n} \frac{1}{i}$$

  Since $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} = \frac{\pi^2}{6}$ we get $\qquad Var(X) \sim \frac{\pi^2}{6} n^2$

- Concentration around the expectation

  The Chebyshev inequality does not provide a strong result:

  For $\beta > 1$,

  $\Pr\{X > \beta n \ln n\} = \Pr\{X - n \ln n > (\beta - 1) n \ln n\}$

  $\leq \Pr\{|X - n \ln n| > (\beta - 1) n \ln n\} \leq \frac{Var(X)}{(\beta-1)^2 n^2 \ln^2 n}$

  $\sim \frac{n^2}{n^2 \ln^2 n} = \frac{1}{\ln^2 n}$

Stronger concentration around the expectation

- Let $\mathcal{E}_i^r$ the event: "coupon type i is not collected during the first r trials". Then
$$\Pr\{\mathcal{E}_i^r\} = (1 - \tfrac{1}{n})^r \leq e^{-\frac{r}{n}}$$
For $r = \beta n \ln n$ we get $\quad \Pr\{\mathcal{E}_i^r\} \leq e^{-\frac{\beta n \ln n}{n}} = n^{-\beta}$

- By the union bound we have
$$\Pr\{X > r\} = \Pr\left\{\bigcup_{i=1}^{n} \mathcal{E}_i^r\right\}$$
(i.e. at least one coupon is not selected), so
$$\Pr\{X > r\} \leq \sum_{i=1}^{n} \Pr\{\mathcal{E}_i^r\} \leq n \cdot n^{-\beta} = n^{-(\beta-1)} = n^{-\epsilon},$$
where $\epsilon = \beta - 1 > 0$

Sharper concentration around the mean - a heuristic argument

- Binomial distribution (#successes in n independent trials each one with success probability p)
  $X \sim B(n,p) \Rightarrow \Pr\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}$
  $(k = 0, 1, 2, \dots, n)$
  $E(X) = np, Var(X) = np(1-p)$

- Poisson distribution)
  $X \sim P(\lambda) \Rightarrow \Pr\{X=x\} = e^{-\lambda} \frac{\lambda^x}{x!} \qquad (x = 0, 1, \dots)$
  $E(X) = Var(X) = \lambda$

- Approximation: It is $B(n,p) \xrightarrow{\infty} P(\lambda)$, where $\lambda = np$.
  For large n, the approximation of the binomial by the Poisson is good.

Towards the sharp concentration result

- Let N_i^r = number of times coupon i chosen during the first r trials.
- Then $\mathcal{E}_i^r$ is equivalent to the event $\{N_i^r = 0\}$.
- Clearly $N_i^r \sim B\left(r, \frac{1}{n}\right)$, thus
  $$\Pr\{N_i^r = x\} = \binom{r}{x} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{r-x}$$
- Let λ a positive real number. A r.v. Y is $P(\lambda) \Leftrightarrow$
  $$\Pr\{Y = y\} = e^{-\lambda} \cdot \frac{\lambda^y}{y!}$$
- As said, for suitable small λ and as r approaches ∞, $P\left(\frac{r}{n}\right)$ is a good approximation of $B\left(r, \frac{1}{n}\right)$. Thus
  $$\Pr\{\mathcal{E}_i^r\} = \Pr\{N_i^r = 0\} \simeq e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda} = e^{-\frac{r}{n}} \qquad \text{(fact 1)}$$

An informal argument on independence

We will now claim that the $\mathcal{E}_i^r$ $(1 \le i \le n)$ events are "almost independent", (although it is obvious that there is some dependence between them; but we are anyway heading towards a heuristic).

Claim 1. For $1 \le i \le n$, and any set if indices $\{j_1, \ldots, j_k\}$ not containing i,
$\Pr\left\{\mathcal{E}_i^r \middle| \bigcap_{l=1}^k \mathcal{E}_{j_l}^r\right\} \simeq \Pr\{\mathcal{E}_i^r\}$

Proof: $$\Pr\left\{\mathcal{E}_i^r \middle| \bigcap_{l=1}^k \mathcal{E}_{j_l}^r\right\} = \frac{\Pr\left\{\mathcal{E}_i^r \cap \left(\bigcap_{l=1}^k \mathcal{E}_{j_l}^r\right)\right\}}{\Pr\left\{\bigcap_{l=1}^k \mathcal{E}_{j_l}^r\right\}} = \frac{\left(1 - \frac{k+1}{n}\right)^r}{\left(1 - \frac{k}{n}\right)^r}$$
$$\simeq \frac{e^{-\frac{r(k+1)}{n}}}{e^{-\frac{rk}{n}}} = e^{-\frac{r}{n}} \simeq \Pr\{\mathcal{E}_i^r\}$$

$\square$

An approximation of the probability

- Because of fact 1 and Claim 1, we have:
$$\Pr\left\{\overline{\bigcup_{i=1}^{n} \mathcal{E}_i^m}\right\} = \Pr\left\{\bigcap_{i=1}^{n} \overline{\mathcal{E}_i^m}\right\} \simeq (1 - e^{-\frac{m}{n}})^n \simeq e^{-ne^{-\frac{m}{n}}}$$
For $m = n(\ln n + c) = n \ln n + cn$, for any constant $c \in R$, we then get
$$\Pr\{X > m = n \ln n + cn\} = \Pr\left\{\bigcup_{i=1}^{n} \mathcal{E}_i^m\right\} \simeq \Pr\left\{\bigcap_{i=1}^{n} \overline{\mathcal{E}_i^m}\right\}$$
$$= 1 - e^{-e^{-c}}$$
- The above probability:
  - is close to 0, for large positive c
  - is close to 1, for large negative c

  Thus the probability of having collected all coupons, rapidly changes from nearly 0 to almost 1 in a small interval cantered around $n \ln n$ (!)

The rigorous result

Theorem: Let X the r.v. counting the number of trials for having collected each one of the n coupons at least once. Then, for any constant $c \in R$ and $m = n(\ln n + c)$ it is

$$\lim_{n \to \infty} \Pr\{X > m\} = 1 - e^{-e^{-c}}$$

Note 1. The proof uses the Boole-Bonferroni inequalities for inclusion-exclusion in the probability of a union of events.

Note 2. The power of the Poisson heuristic is that it gives a quick, approximative estimation of probabilities and offers some intuitive insight towards the accurate behaviour of the involved quantities.