EECS 126 Discussion 2

→ no actual discussion, due to Labor Day; W 2-3 discussion, or the walkthrough, or these notes instead. Also Piazza!

→ thanks for the feedback, will address at the start of next week!

→ I won't be at OH this W 2-4, but other TAs/readers will! I'll cover theirs in future weeks and announce at the start of discussion when that'll happen.

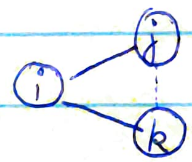
→ If you're reading this on Monday, remember to do self-grades! ☺

1. Clustering Coefficient

WTG: $E[C(i) \mid N(i) \geq 2]$, $C(i) = T(i) / \binom{N(i)}{2}$

$T(i)$: awkward to work with: how do I know how many triangles there are?

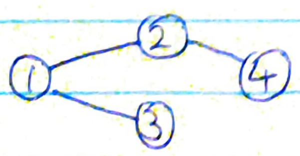
'i' is a vertex, so it has edges to the other two vertices in the triangle.



$T(i) = 1 \iff$ the edge j-k exists.

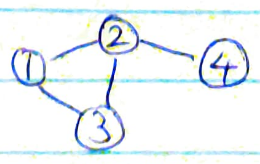
This works for any pair of 'i's neighbours, so $T(i)$ = the number of connections between 'i's neighbours.

Example:



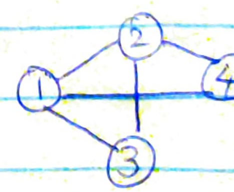
$T(1) = 0$

and #NCs = 0
↑
neighbour connection



$T(1) = 1$ (1-2-3)

and #NCs = 1
(2-3)



$T(1) = 2$ (1-2-3, 1-2-4)

and #NCs = 2
(2-3, 2-4)

Max value that $T(i)$ can be is $\binom{N(i)}{2}$,

so max value that $C(i) = \frac{T(i)}{\binom{N(i)}{2}}$ can be is 1.

Distribution of $C(i)$?

For $N(i)=k$, $C(i)$ is the average proportion of neighbor connections that are made. (So $0 \le C(i) \le 1$ makes sense.)

Every pair is independent of every other one, and the probability of each connection is p. So:

$$E[C(i) \mid N(i)=k] = \frac{(\#\text{ of possible pairs}) \cdot (\text{prob. of each pair})}{(\#\text{ of possible pairs})}$$

$\binom{k}{2}$ as seen above.

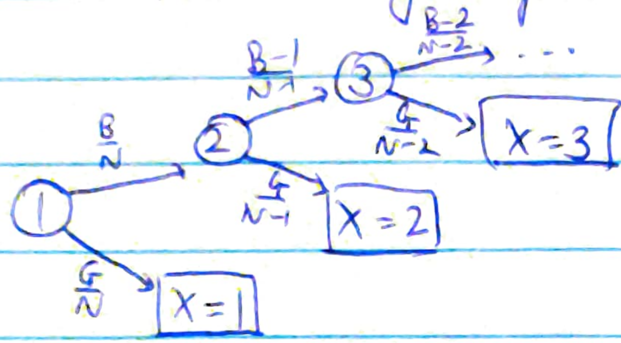
$$= \frac{\binom{k}{2}p}{\binom{k}{2}} = \boxed{p}.$$

This is true for any $k \geq 2$, so our final answer is

$\underline{E[C(i) \mid N(i) \geq 2] = p}$.

2. Sampling without Replacement

I'll draw a diagram first!



$\Rightarrow$

$$P(X=1) = G/N$$
$$P(X=2) = \frac{B}{N} \cdot \frac{G}{N-1}$$
$$P(X=3) = \frac{B}{N} \cdot \frac{B-1}{N-1} \cdot \frac{G}{N-2}$$
$$\vdots$$

Could write out the PMF for $X=1$, $X=2$, ... $X=B$ and compute from definitions of E and var, but that's a lot of work.

Ideas for shortcuts?

- symmetry? → situation doesn't really have it.
- linearity of expectation? → seems good, but how?
- indicator variables? → good choice! It's a sequence of events, so can look at each one separately.

Say X_i : bad item 'i' is before good item 1.

$X = 1 + \sum_{i=1}^{B} X_i$: X has to be at least 1, then one indicator for each bad item you see.

$E[X] = 1 + \sum_{i=1}^{B} E[X_i] = 1 + B\,E[X_1]$

Distribution of all indicators is the same

(think of this as defining a sequence of, e.g., BBGGBBB.....G;

each X_i is just a question about relative order of two items, which one's earlier/later in time than the other Bs doesn't matter).

For $E[X_i]$, consider bad item B_i and the G good items.

$\underline{B_i}$ $\underline{G}$ $\underline{G}$ $\underline{G}$ .... $\underline{G}$

Averaged over all orderings, what's $P(B_i$ is first$)$?

$G+1$ possible choices for first, so $P(B_i$ is first$) = E[X_i] = \frac{1}{G+1}$

$\therefore$ $\boxed{E[X] = 1 + \frac{B}{G+1} = \frac{N+1}{G+1}}$

$\text{var}(X) = \text{var}(1 + \sum_{i=1}^{B} X_i)$. Not as easy: X_i, X_j not independent, so no linearity of variance.

$= \text{var}(\sum_{i=1}^{B} X_i)$

Instead, let's use $\text{var}(X) = E[X^2] - E[X]^2$; we know $E[X]$ so look at $E[X^2]$

$$\text{var}(\sum_{i=1}^{B} X_i) = \underbrace{E[(\sum_{i=1}^{B} X_i)^2]}_{\text{split up.}} - E[X-1]^2$$

$$E[(\sum_{i=1}^{B} X_i)^2] = \sum_{i=1}^{B} E[X_i^2] + \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} E[X_i X_j]$$

All events are identical, so

$$E[(\sum_{i=1}^{B} X_i)^2] = B\,E[X_1^2] + B(B-1)\,E[X_1 X_2]$$

$E[X_1^2] = E[X_1] = \frac{1}{G+1}$ (it's an indicator RV; $0^2 = 0$ and $1^2 = 1$).

$E[X_1 X_2]$ is the same reasoning as $E[X_1]$: get $G+2$ spaces for G good and the 2 bad items, what's $P(B_1, B_2$ in the first two$)$?

$$\underset{\substack{\uparrow \\ G+2 \text{ choices}}}{\underline{B_1}} \quad \underset{\substack{\uparrow \\ G+1 \text{ choices}}}{\underline{B_2}} \quad \underline{G} \quad \underline{G} \quad \ldots \quad \underline{G}$$

$\frac{1}{2}$ for derangement

$$\therefore E[X_1 X_2] = \frac{2}{(G+1)(G+2)}$$

$$\therefore E[(\sum_{i=1}^{B} X_i)^2] = \frac{B}{G+1} + \frac{2B(B-1)}{(G+1)(G+2)}$$

Now, put it all together!

$$\operatorname{var} X = \frac{B}{G+1} + \frac{2B(B-1)}{(G+1)(G+2)} - \left(\frac{B}{G+1}\right)^2$$

For simplicity bring under a common denominator:

$$\operatorname{var} X = \frac{B(G+1)(G+2) + 2B(B-1)(G+1) + B^2(G+2)}{(G+1)^2(G+2)}$$

$$= \boxed{\frac{BG(N+1)}{(G+1)^2(G+2)}}.$$

3. Tricky Markov Bound

(I had this one on HW when I took the class, and it took a long time
- so don't worry too much if you find it hard!)

Markov bound in general: $P(X \geq a) \leq E[X]/a$ if $X \geq 0$

Desired answer: $P(X \geq \alpha) \leq \sigma^2/(\alpha^2 + \sigma^2)$

Markov in general doesn't have a σ factor, so we want to pick an 'a' that depends on σ to factor this in.

We also want to get rid of the $E[X]$, but how? Markov doesn't even work for our X, as it's not nonnegative.

Let's make an RV that is nonnegative and has something to do with σ.

First attempt: X^2.

$E[X^2] = \text{Var } X + E[X]^2 = \text{Var } X = \sigma^2$

Since we're still dealing with the event $X \geq a$, use Markov w/ $X^2 \geq a^2$.

($\alpha > 0$ implies $X \geq \alpha \Leftrightarrow X^2 \geq \alpha^2$)

$P(X^2 \geq \alpha^2) \leq E[X^2]/\alpha^2 = \sigma^2/\alpha^2$.

Good start, but the question has a tighter bound.

Sending $X \rightarrow X^2$ seemed to work well, let's try some other function f that ensures $f(X)$ is nonnegative.

What should f be?

Maybe $f(X) = X^4$? Let's try it:

$$P(X \geq \alpha) \leq \frac{E[X^4]}{\alpha^4}$$

Oops, what's $E[X^4]$? No idea. (also no $\square^4$ in the final answer.)
So we can't go to higher powers. Is there more to do with $\{1, X, X^2\}$?

We also haven't used a general quadratic yet: maybe
$$f(X) = AX^2 + BX + C$$
for some A, B, C.
But remember $f(X) \geq 0$, so we can't use any A, B, C.
Instead, to ensure nonnegativity we'll try

$$\boxed{f(X) = (X+c)^2}$$

c is a free parameter we can vary to give us the tightest bound!

$$P(X \geq \alpha) = P\left((X+c)^2 \geq (\alpha+c)^2\right) \leq \frac{E[(X+c)^2]}{(\alpha+c)^2} = \frac{E[X^2] + 2cE[X] + c^2}{(\alpha+c)^2}$$

$$P(X \geq \alpha) \leq \frac{\sigma^2 + c^2}{(\alpha+c)^2}$$

This looks closer! How do we pick the best c?

Want to minimize $\frac{\sigma^2+c^2}{(\alpha+c)^2}$ wrt c, so take a c derivative:

$$\frac{d}{dc}\left(\frac{\sigma^2+c^2}{(\alpha+c)^2}\right)=0 \Rightarrow \frac{2(\alpha c-\sigma^2)}{(\alpha+c)^3}=0 \Rightarrow \boxed{c=\frac{\sigma^2}{\alpha}}$$

Plug this in to get

$$P(X\geq\alpha) \leq \frac{\sigma^2+\frac{\sigma^4}{\alpha^2}}{\left(\alpha+\frac{\sigma^2}{\alpha}\right)^2} = \frac{\sigma^2\alpha^2+\sigma^4}{(\alpha^2+\sigma^2)^2}$$

$$P(X\geq\alpha) \leq \frac{\sigma^2\alpha^2+\sigma^4}{(\sigma^2+\alpha^2)^2} = \sigma^2\left(\frac{\sigma^2+\alpha^2}{(\sigma^2+\alpha^2)^2}\right)$$

$$\boxed{P(X\geq\alpha) \leq \frac{\sigma^2}{\sigma^2+\alpha^2}}$$