تحلیل داده‌های حجیم

مدرس : دکتر ایمان غلامی

[پاییز ۱۴۰۰]

تمرن سری ۴: سوال ۱ نگارنده: امیرمحمد شعبانی

الف

با اینکه سوال تعریفی از $don't know$ نکرده است، اما احتمالا زمانی این اتفاق می‌افتد که تمام k سطری که انتخاب کردیم صفر باشند.

طبق فرض سوال، ابتدا یک جایگشت رندم انجام میشه و سپس k سطر را انتخاب می‌کنیم. این برابر است با k بار انتخاب یک سطر. (چون جایگشت و انتخاب به صورت رندم انجام می‌شود.) بنابراین زمانی در این حالت $don't know$ داریم که تمام سطرهای انتخابی از سطرهای صفر باشد. بنابراین چون $n-m$ سطر غیر صفر داریم، انتخاب یک سطر صفر به احتمال $\frac{n-m}{n}$ است و چون k بار انتخاب میکنیم؛ پس احتمال آن برابر است با $(\frac{n-m}{n})^k$.

ب

زمانی الگوریتم ناموفق که حداقل دو $don't know$ داشته باشیم؛ زیرا اگر یکی داشته‌باشیم، نتیجه می‌شود که با بقیه متفاوت است.

برای محاسبه مقدار k بنظرم چند شرط نیازمندیم که سوال به ما نگفته‌است. در ابتدا این فرض مهمی است که n از مقدار m بسیار بزرگ‌تر باشد. همچنین مقدار k با مقادیر n, m بدست می‌آیند.

$$(\frac{n-m}{n})^k = (1-\frac{m}{n})^k \Rightarrow \frac{m}{n} \to 0, \frac{n}{m} = x \Rightarrow (1-\frac{m}{n})^k = ((1-\frac{1}{x})^x)^{\frac{k}{x}} = (e^{-1})^{\frac{k}{x}} = e^{-10}$$
$$\Rightarrow \frac{k}{x} = 10 \Rightarrow k = 10x \Rightarrow k = \frac{10n}{m}$$