تحلیل داده‌های حجیم

مدرس : دکتر ایمان غلامی

[پاییز ۱۴۰۰]

تمرن سری ۴: سوال ۲ نگارنده: امیرمحمد شعبانی

الف

برای حل مسئله ابتدا به بدست‌آوردن چند احتمال ساده نیازمندیم و سعی می‌کنیم گام به گام به احتمالی که می‌خواهیم برسیم.

می‌دانیم این تابع زمانی می‌تواند دو اثر انگشت را در یک باکت قرار دهد که هردو در هر ۳ خانه $minutiae$ داشته باشند. حال احتمال اینکه یک اثر انگشت در یک خانه $minutiae$ داشته باشد ۰.۲ است و احتمال اینکه اثر انگشت مشابه آن نیز در همان خانه $minutiae$ داشته باشد با فرض اینکه اثر انگشت در آن خانه $minutiae$ دارد، ۰.۸ است. بنابراین احتمال اینکه هردو در یک باکت قرار گیرند برابر است با $0.2^3 \times 0.8^3 = 0.004096$.

بنابراین احتمال اینکه توی یک تابع این دو در یک باکت قرار نگیرند برابر است با $1 - 0.004096 = 0.995904$. بنابراین احتمال اینکه در همه تابع‌ها در یک باکت قرار نگیرند برابر است با $0.995904^{2024} = 0.000246702$. بنابراین احتمال اینکه حداقل در یکی از توابع در یک باکت قرار بگیرند برابر است با $1 - 0.000246702 = 0.999753298$.

حال $false\ negative$ برابر است با احتمال اینکه دوتا عکس که مشابه‌اند اشتباهی متفاوت تشخیص دهیم که برابر است با احتمال اینکه در هیچ تابعی در یک باکت قرار نگیرند که طبق بالا برابر است با 0.000246702.

حال احتمال اینکه دو اثر انگشت متفاوت را باهم یکسان تشخیص دهد برابر است با احتمال اینکه هردو اثر انگشت در هر ۳ خانه دارای $minutiae$ باشند؛ بنابراین برابر است با $0.2^3 \times 0.2^3 = 0.000064$. احتمال اینکه یکی از توابع بگه شبیه نیستند برابر است با $1 - 0.000064 = 0.999936$ و احتمال اینکه در همه توابع شبیه نباشند برابر است با $0.999936^{2024} = 0.87849932$ و احتمال اینکه حداقل در یکی از توابع در یک باکت قرار گیرند برابر است با $1 - 0.87849932 = 0.12150068$ که دقیقا برابر است با احتمال اینکه به اشتباه یک تشابه غلط بدست آید که همان $false\ positive$ است.

برای حالتی که به دوگروه تقسیم می‌شوند به صورت مشابه می‌توان استدلال کرد. احتمال اینکه دو اثر انگشت مشابه در یک باکت بیافتند را داریم. حال چون and است، احتمال اینکه همه توابع یک گروه مشابه تشخیص دهند برابر است با $0.004096^{1024} \simeq 0$ و احتمال اینکه حداقل یکی از آن‌ها مشابه تشخیص ندهد برابر است با $1 - 0 = 1$. تا اینجا احتمال این را بدست آوردیم که در یک گروه دو اثر انگشت یکسان مشابه تشخیص داده نشوند. حال احتمال اینکه مشابه تشخیص داده نشوند برابر است با احتمال اینکه در هردو گروه مشابه تشخیص داده نشوند که برابر است با $1 \times 1 = 1$ که برابر $false\ negative$ است. (اینجا بود که بنظرم رسید شاید سوال جای «و» و «یا» را اشتباه داده‌است.)

همچنین احتمال در یک باکت افتادن دو اثر انگشت متفاوت را داریم. برای اینکه دو اثر انگشت متفاوت مشابه تشخیص داده شوند باید یکی از دو گروه مشابه تشخیص دهد. بنابراین اگر احتمال اینکه مشابه تشخیص داده نشوند را بدست آوریم می‌توانیم با یک منهای آن به جواب برسیم. احتمال اینکه در یک گروه مشابه تشخیص داده نشوند برابر است با یک منهای احتمال اینکه مشابه تشخیص داده شوند. احتمال اینکه همه توابع بگویند این دو اثر انگشت مشابه‌اند برابر است با $0.000064^{1024} \simeq 0$ و احتمال اینکه یکی از آن‌ها مشابه تشخیص ندهد برابر است با $1 - 0 = 1$ و احتمال اینکه هردو گروه این دو اثر انگشت را مشابه تشخیص ندهد برابر است با $1 \times 1 = 1$ و احتمال اینکه حداقل یکی از این دو گروه این دو اثر انگشت را مشابه تشخیص دهد برابر است با $1 - 1 = 0$ که برابر با $false\ positive$ است.

از آنجایی که حس می‌کنم سوال «و» و «یا» را برعکس گفته، برای حالتی که در هر گروه یا شود و سپس دو گروه و شود نیز حل می‌کنم. احتمال اینکه یک تابع دو اثر انگشت مشابه را اشتباه تشخیص دهد برابر است با $1 - 0.004096 = 0.995904$ و احتمال اینکه همه توابع یک گروه مشابه تشخیص ندهند برابر است با $0.995904^{1024} = 0.014951892$ و احتمال اینکه حداقل یکی مشابه تشخیص دهد برابر است با $1 - 0.014951892 = 0.985048108$ و احتمال اینکه در هردو گروه مشابه تشخیص داده شود برابر است با $0.985048108^2 = 0.970319775$ و احتمال اینکه حداقل یکی از این دو گروه مشابه تشخیص ندهد که باعث می شود ما نیز مشابه تشخیص ندهیم برابر است با $1 - 0.970319775 = 0.029680225$ که برابر است با $false\ negative$.

احتمال اینکه دو اثر انگشت متفاوت نیز مشابه تشخیص داده شوند را نیز داریم. احتمال اینکه همه تابع‌های در یک گروه دو اثر انگشت را متفاوت تشخیص دهند برابر است با $(1 - 0.000064)^{1024} = 0.936563366$ و احتمال اینکه حداقل یکی آن‌ها را مشابه تشخیص دهد برابر است با $1 - 0.936563366 = 0.063436634$ و احتمال اینکه هردو گروه تشخیص دهند این‌دو مشابه‌اند برابر است با $0.063436634 \times 0.063436634 = 0.004024207$ که همان $false\ positive$ است.

ب

باتوجه به توضیحاتی که بالا داده‌شده برای حالتی که تابع‌ها را گروه‌بندی نکردیم داریم:

$$fn : false\ negative = (1 - 0.004096)^n$$
$$fp : false\ positive = 1 - (1 - 0.000064)^n$$
$$\min fn + fp = 1 - (1 - 0.000064)^n + (1 - 0.004096)^n \Rightarrow n = \frac{\log(\frac{\log(\frac{15625}{15561})}{\log(\frac{15625}{15624})})}{\log\frac{248}{247}}$$

همچنین برای حالتی که تابع‌ها را به دوگروه تقسیم کردیم و در هرگروه «و» انجام می‌دهیم، داریم:

$$fn : false\ negative = (1 - 0.004096^{\frac{n}{2}})^2$$
$$fp : false\ positive = 1 - (1 - 0.000064^{\frac{n}{2}})^2$$
$$\min fn + fp = 1 - (1 - 0.000064^{\frac{n}{2}})^2 + (1 - 0.004096^{\frac{n}{2}})^2 \Rightarrow n \simeq 0.392388$$

همچنین برای حالتی که تابع‌ها را به دوگروه تقسیم کردیم و در هرگروه «یا» انجام می‌دهیم، داریم:

$$fn : false\ negative = 1 - (1 - (1 - 0.004096)^{\frac{n}{2}})^2$$

$fp : false\ positive = (1 - (1 - 0.000064)^{\frac{n}{2}})^2$

$\min fn + fp = 1 - (1 - (1 - 0.004096)^{\frac{n}{2}})^2 + (1 - (1 - 0.000064)^{\frac{n}{2}})^2 \Rightarrow n \to \infty$