Generalized Birthday Paradox

Karl Stratos

Consider any discrete set $\mathcal{X}$ and any distribution P over $\mathcal{X}$. For any subset $X \subseteq \mathcal{X}$ and any iid samples $S \sim P^N$, write $\text{Pure}_X(S)$ to denote the event that S contains no duplicates of elements from X.

Impurity statement. Suppose $\min_{x \in \mathcal{X}} P(x) \geq 1/M$. Then given any $\delta \in (0, 1)$, if

$$N \geq \sqrt{2M \ln(1/\delta)} + 1$$

we have $\text{Pr}_{S \sim P^N} (\neg \text{Pure}_{\mathcal{X}}(S)) \geq 1 - \delta$.

Application. If people's birthdays are uniformly random on $M = 365$ days, then there is a birthday collision among $N = 24$ random people with $\geq 50\%$ chance.

Purity statement. Suppose $\max_{x \in X} P(x) \leq 1/M$. Then given any $\delta \in (0, 1)$, if

$$N \leq \min \left\{ 0.01M, 1.4\sqrt{M \ln(1/(1-\delta))} \right\}$$

we have $\text{Pr}_{S \sim P^N} (\text{Pure}_X(S)) \geq 1 - \delta$.

Remark. Note the first requirement forces that $N \ll M$ and the statement is not very useful when M is too small (e.g., in the birthday problem above, we can only say weak statements like: there is no birthday collision among 3 random people with $\geq 50\%$ chance). The requirements on N can be equivalently written as a requirement on M:

$$M \geq \max \left\{ 100N, \frac{0.505}{\ln(1/(1-\delta))} N^2 \right\}$$

Application. Once we sort the elements of $\mathcal{X}$ in decreasing probabilities so that

$$P(x_1) \geq P(x_2) \geq \cdots$$

then the largest possible value for $P(x_M)$ is $1/M$, thus we have $P(x_i) \leq 1/M$ for all $i > M$. This means in N samples with probability at least $1-\delta$ we have no duplicates of x_i where $i > \max\left\{100N, 0.505/\ln(1/(1-\delta))N^2\right\}$.

Related lemma (outlier risk). In any $N \geq 2$ iid samples, with probability at least $1/4$ we fail to observe a phenomenon which occurs with probability $1/N$.

Application. For any $F : \mathcal{X} \to [0, F_{\max}]$, an estimate of $\mathbf{E}_{x \sim Q}\left[e^{F(x)}\right]$ based on $N \geq 2$ samples can never guarantee that it is less than $(1/N)e^{F_{\max}}$ with high confidence, since with probability at least $1/4$ there exists $x \in \mathcal{X}$ such that $Q(x) = 1/N$ and $F(x) = F_{\max}$.

A Proofs

By the independence of samples,

$$\begin{aligned}\Pr_{S\sim P^N}(\text{Pure}_X(S)) &= \prod_{i=2}^{N} \Pr\left(\forall j = 1\ldots i-1:\; x_i \notin X \vee x_j \notin X \vee x_i \neq x_j\right)\\ &= \prod_{i=2}^{N}\left(1 - \sum_{j=1}^{i-1} \Pr\left(x_i, x_j \in X \wedge x_i = x_j\right)\right)\end{aligned}$$

Proof of the impurity statement. Follows by solving for N in

$$\Pr_{S\sim P^N}(\text{Pure}_{\mathcal{X}}(S)) = \prod_{i=2}^{N}\left(1 - \sum_{j=1}^{i-1} P(x_j)\right) \leq \prod_{i=2}^{N}\left(1 - \frac{i-1}{M}\right) \leq \exp\left(-\frac{N(N-1)}{2M}\right) \leq \delta$$

Proof of the purity statement. First note that

$$\Pr\left(x_i, x_j \in X \wedge x_i = x_j\right) \leq \Pr\left(x_i = x_j \middle| x_i, x_j \in X\right) = \Pr\left(x_j \middle| x_j \in X\right) \leq \frac{1}{M}$$

Using the fact that $1 - x \geq e^{-1.01x}$ for $x \in [0, 0.01]$,

$$\Pr_{S\sim P^N}(\text{Pure}_{\mathcal{X}}(S)) \geq \prod_{i=2}^{N}\left(1 - \frac{i-1}{M}\right) \geq \exp\left(-\frac{0.505N^2}{M}\right)$$

Solving this for $1 - \delta$ yields the result.

Proof of the outlier risk lemma. This probability is $(1 - 1/N)^N$ which is at least $1/4$ for all $N \geq 2$.