

CS5014 Foundations of Data Science & Machine Learning

Quiz 03 — Solutions

April 21, 2021 — 8.30 - 9.30 AM

Instructions

1. This is a **zoom proctored** exam. Please adjust your seating so that **your face, hands, answer book and the mobile phone that you will use to scan the sheets are always in the webcam view**. Do not leave your seat or talk to anyone during the exam.
2. Write your answer on plain paper with your **name and roll number on the first sheet**.
3. This is a **closed book** exam. Do not refer to any books, notes, the Internet or any other person during the exam.
4. You can take **maximum 5 minutes after the exam to scan** the sheets into a **single PDF** file and upload to Moodle. Submissions made after 9:45 AM will be evaluated only if there is a genuine reason for the delay.

Questions

Question 1. Let \mathcal{H} be a *finite* hypothesis class over a domain X , f a binary labelling of X , D a probability distribution over X and $\epsilon, \delta \in (0, 1)$. Show that if a training set S of $n \geq \frac{1}{2\epsilon^2} (\ln |\mathcal{H}| + \ln \frac{2}{\delta})$ elements are sampled independently according to D , then with probability at least $1 - \delta$, every hypothesis $h \in \mathcal{H}$ has an in-sample error ϵ -close to the true error (i.e., $P[\forall h \in \mathcal{H} |E_S(h) - E_D(h)| \leq \epsilon] \geq 1 - \delta$).

Hoeffding Bounds. If X is the average of n $\{0, 1\}$ -random variables X_1, \dots, X_n with $P(X_i = 1) = p$ for each $i \in [n]$, then

$$\begin{aligned} P(X > p + \epsilon) &\leq e^{-2\epsilon^2 n}, \text{ and} \\ P(X < p - \epsilon) &\leq e^{-2\epsilon^2 n}. \end{aligned}$$

Answer 1. Let h be any hypothesis \mathcal{H} . Let $S = \{x_1, \dots, x_n\}$, where each x_i is picked independently according to D . For $i \in [n]$, let X_i denote the Bernoulli random variable

$$X_i = \begin{cases} 0, & h(x_i) = f(x_i) \\ 1, & h(x_i) \neq f(x_i). \end{cases}$$

For each random variable X_i , $p = P[X_i = 1] = E_D(h)$, the true error of h . The in-sample error $E_S(h)$ is $\frac{1}{n} |\{x_i \in S : h(x_i) \neq f(x_i)\}|$ which is equal to the average of $X_1 \dots X_n$. By the Hoeffding bound

$$\begin{aligned} P(|E_S(h) - E_D(h)| > \epsilon) &\leq 2e^{-2\epsilon^2 n} \\ &\leq 2e^{-\ln|\mathcal{H}| - \ln(2/\delta)} & n \geq \frac{1}{2\epsilon^2} \left(\ln |\mathcal{H}| + \ln \frac{2}{\delta} \right) \\ &= \frac{\delta}{|\mathcal{H}|}. \end{aligned}$$

Now by union bound, the probability that there exists some $h \in \mathcal{H}$ such that $|E_S(h) - E_D(h)| > \epsilon$ is at most δ . Hence with probability at least $1 - \delta$, every hypothesis $h \in \mathcal{H}$ has an in-sample error ϵ -close to the true error

Question 2. A directed graph (digraph) is said to be *Eulerian* if the in-degree equals the out-degree for every vertex. Let G be an Eulerian digraph on the vertex set $[n]$ without any self-loops. Consider a random walk on G with transition probabilities

$$p_{i,j} = \begin{cases} \frac{1}{d_i}, & j \in N^+(i), \\ 0, & \text{otherwise} \end{cases}$$

where $N^+(i)$ is the set of out-neighbours of node i and $d_i = |N^+(i)|$ is the out-degree of node i . Find a stationary distribution for this random walk.

Hint. If G is constructed from an undirected graph H by replacing every undirected edge of H with two opposite arcs in G (like we did for Metropolis-Hastings) then G will be Eulerian. You can use this special case to guess the stationary distribution. Do not then forget to prove that your answer is a stationary distribution in the more general case asked in the question.

Answer 2. Let m be the number of arcs in G . Let $\pi = \left(\frac{d_1}{m}, \dots, \frac{d_n}{m}\right)$. π is a probability vector since $m = \sum_{i=1}^n d_i$. The i -th component of $P^T\pi$ is

$$\begin{aligned} (P^T\pi)_i &= \sum_{j=1}^n p_{j,i}\pi_j \\ &= \sum_{j \in N^-(i)} \frac{1}{d_j} \pi_j && (N^-(i) \text{ is the set of in-neighbours of } i) \\ &= \sum_{j \in N^-(i)} \frac{1}{d_j} \frac{d_j}{m} \\ &= \frac{1}{m} |N^-(i)| \\ &= \frac{1}{m} |N^+(i)| && (G \text{ is Eulerian}) \\ &= \frac{1}{m} d_i \\ &= \pi_i. \end{aligned}$$

Hence $P^T\pi = \pi$.