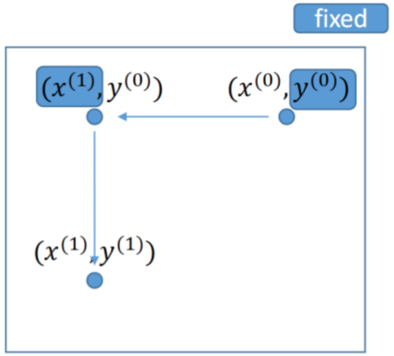
SDSC3001 Midterm

2024.11.28

Q1 (15 points) Prove that the stationary distribution of Gibbs sampling (page 15, Lec_3_2) is $p(x)$ by showing that Gibbs sampling satisfies the reversibility condition (page 14, Lec_3).

Gibbs Sampling

- Sample from a multivariate distribution $p(\mathbf{x})$, $\mathbf{x} = (x_1, x_2, \ldots, x_d)$
- Marginal distribution $p(x_i \mid \mathbf{x}_{\neg i})$, $\mathbf{x}_{\neg i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$
- Given the current $\mathbf{x}$
  - Randomly choose a coordinate i
  - Sample y_i based on $p(x_i \mid \mathbf{x}_{\neg i})$
  - Set next sample $\mathbf{y}$ as $(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_d)$



- A sufficient (but not necessary) condition: reversibility/detailed balance condition

$$\forall (i, j), \pi_i P_{ij} = \pi_j P_{ji}$$

Q1 (15 points) Prove that the stationary distribution of Gibbs sampling (page 15, Lec_3_2) is $p(x)$ by showing that Gibbs sampling satisfies the reversibility condition (page 14, Lec_3).

Suppose we have two states $\mathbf{x}$ and $\mathbf{x}'$.

We need to prove the reversibility condition：

$$T(\mathbf{x} \to \mathbf{x}')p(\mathbf{x}) = T(\mathbf{x}' \to \mathbf{x})p(\mathbf{x}')$$

- If $\mathbf{x}$ and $\mathbf{x}'$ are not "neighbors", then $T(\mathbf{x} \to \mathbf{x}') = T(\mathbf{x}' \to \mathbf{x}) = 0$
- Neighbor: Two vectors differ only in one dimension.

The transition probabilities are given by:

$$T(\mathbf{x} \to \mathbf{x}') = \pi_i p(x_i'|\mathbf{x}_{\setminus i})$$

where π_i is the probability of choosing to update the ith variable

Then we have:

$$T(\mathbf{x} \to \mathbf{x}')p(\mathbf{x}) = \pi_i p(x_i'|\mathbf{x}_{\setminus i}) \underbrace{p(x_i|\mathbf{x}_{\setminus i})p(\mathbf{x}_{\setminus i})}_{p(\mathbf{x})}$$

and

$$T(\mathbf{x}' \to \mathbf{x})p(\mathbf{x}') = \pi_i p(x_i|\mathbf{x}'_{\setminus i}) \underbrace{p(x_i'|\mathbf{x}'_{\setminus i})p(\mathbf{x}'_{\setminus i})}_{p(\mathbf{x}')}$$

But $\mathbf{x}'_{\setminus i} = \mathbf{x}_{\setminus i}$ so detailed balance holds.

Q2 (25 points) Given are the following eight transactions on items $\mathcal{V} = \{A, B, C, D, E, F\}$.

Transaction id	Transaction (set of items)
1	ABCD
2	BCD
3	CEF
4	BC
5	CDF
6	ABCDE
7	ABD
8	AF

The support of an itemset/pattern $S \subseteq \mathcal{V}$ is the number of transactions containing all items of S. Let the minimum support be $minSup$ =3. Find all frequent itemsets/patterns whose supports are at least $minSup$. List all frequent itemsets and their corresponding supports.

- Size < 4
- If a subset of a set is infrequent, then the set itself is also infrequent.

A 4
B 5
C 6
D 5
E 2
F 3

AB 3	CF 2
AC 2	DF 1
AD 3	
AF 1	
BC 4	
BD 4	
BF 0	
CD 4	

ABC	ADF	CDE
ABD	AEF	CDF
ABE	BCD	DEF
ABF	BCE	
ACD	BCF	
ACE	BDE	
ACF	BDF	
ADE	BEF	

➡

ABD 3		
	BCD 3	

Q2 (25 points) Given are the following eight transactions on items $\mathcal{V} = \{A, B, C, D, E, F\}$.

Transaction id	Transaction (set of items)
1	ABCD
2	BCD
3	CEF
4	BC
5	CDF
6	ABCDE
7	ABD
8	AF

The support of an itemset/pattern $S \subseteq \mathcal{V}$ is the number of transactions containing all items of S. Let the minimum support be $minSup$ =3. Find all frequent itemsets/patterns whose supports are at least $minSup$. List all frequent itemsets and their corresponding supports.

{A:4, B:5, C:6, D:5, F:3}

{AB: 3, AD:3, BC:4, BD:4, CD:4}

{ABD: 3, BCD: 3}

Q3 (10 points) Suppose we have two fair dice. When we roll each die, we will obtain a number from the set $\{1, 2, 3, 4, 5, 6\}$, each with an equal probability of $1/6$. We throw the two dice independently and consider the following event

Event 1: the first die gets an even number E1

Event 2: the second die gets an even number E2

Event 3: the sum of the two dice is an even number E3

(1) (7 points) Show that the above three events are pairwise independent.
(2) (3 points) Are these events mutually independent? Please explain your answer.

$P(E1|E2) = 0.5$
$P(E2|E1) = 0.5$
$P(E1|E3) = 0.5$
$P(E3|E1) = 0.5$
$P(E2|E3) = 0.5$
$P(E3|E2) = 0.5$

$P(E3|E2, E1) = 1$

(If event 1 and event 2 occur, then event 3 must occur.)

Q4 (20 points) Let H be a hypothesis set of binary classifiers. Suppose the underlying (but unknown) distribution of samples is $p(x)$ and every sample x has a label $y_x \in \{0,1\}$. A classifier $h \in H$ is a function where $h(x) \in \{0,1\}$. The expected error of a classifier $h \in H$ is

$$f(h) = \int p(x) I(y_x \neq h(x)) dx,$$

Error on all data

where $I(\cdot)$ is an indicator function. $I\big(y_x \neq h(x)\big) = 1$ if $y_x \neq h(x)$ and $I\big(y_x \neq h(x)\big) = 0$ otherwise. If we have a collection of i.i.d. training samples $S = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, the empirical error of $h \in H$ is defined as

$$f(h; S) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq h(x_i)).$$

Error on the training data

The **Uniform Convergence** condition on S is characterized as

$$\Pr\{\exists h \in H, |f(h) - f(h; S)| \geq \epsilon\} \leq \delta.$$

Suppose we apply the **Empirical Risk Minimization** learning scheme on the training dataset S to learn the optimal classifier from H. We obtain the following classifier

$$h^\circ = \arg\min_{h \in H} f(h; S).$$

Optimal model on the training data

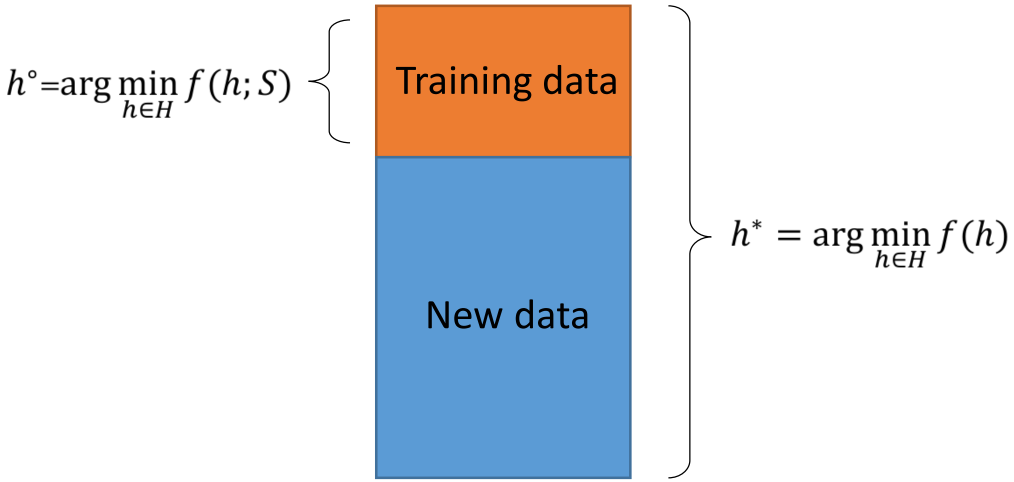
Let $h^* = \arg\min_{h \in H} f(h)$ be the real optimal classifier in H that has the lowest expected error. Prove the following statements.

Optimal model on all data

(1) (10 points) If the Uniform Convergence condition holds for S, then with probability at least $1 - \delta$ we have f° as a nearly optimal classifier such that $f(h^\circ) \leq f(h^*) + 2\epsilon$.

(2) (10 points) If S contains at least $\frac{3}{\epsilon^2} \ln \frac{2|H|}{\delta}$ i.i.d. samples, then the Uniform Convergence condition holds. (Hint: use the absolute error Chernoff bound in page 26 of Lec 2)

- This is a machine learning theory question.
- We cannot obtain all the data for training because some data is unavailable and p(x) is unknown.
- We hope that the classifier learned from the limited training set S can generalize to the entire dataset.



- Train data is used for training the model f.
- New data can be infinite, as it represents the data the model may encounter in real-world applications.
- f can be either a traditional model or a deep learning model.

Example: Predict whether a user is interested in a particular item. If they are, predict 1; otherwise, predict 0.

User ID	Age	Gender	Purchase History	Rating	Predicted Interest
1	25	Male	Yes	4.5	1
2	30	Female	No	2.0	0
3	22	Female	Yes	5.0	1
4	28	Male	Yes	3.0	1
5	35	Female	No	1.0	0
6	27	Male	Yes	4.0	1

- We deploy this model online so that when a new user arrives, we can predict their interests.
- Such new users could be infinitely many, and we don't know their feature distribution p(x), but we hope the learned model can still accurately predict their interests.

NO.

DATE

Q4:

(1) if the uniform convergence condition holds for S.

then with prob. $1-\delta$. $\forall h \in H$.

$$f(h)-\epsilon \le f(h;S) \le f(h)+\epsilon$$ → This inequality holds for any h, including h° and h^*

so for both h°, $h^* \in H$.

$$\begin{cases} f(h^\circ)-\epsilon \le f(h^\circ;S) \le f(h^\circ)+\epsilon \\ f(h^*)-\epsilon \le f(h^*;S) \le f(h^*)+\epsilon \end{cases}$$

According to the ERM rule:

$$h^\circ = \arg\min_{h\in H} f(h;S)$$

$$\Downarrow$$

$\therefore$ $f(h^\circ;S) \le f(h^*;S)$ → Empirical error of h° is less than h^*

$\therefore$ $f(h^\circ)-\epsilon \le f(h^\circ;S) \le f(h^*;S) \le f(h^*)+\epsilon$

$\therefore$ $f(h^\circ)-\epsilon \le f(h^*)+\epsilon$

$\therefore$ with prob. at leat $1-\delta$, we have $f(h^\circ) \le f(h^*)+2\epsilon$.

(2) According to absolute error bound.

$$Pr(|\bar{X}-\mu| \ge \epsilon) \le 2e^{-\frac{\epsilon^2 n}{3\mu}}$$

$\because \begin{cases} f(h;S) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \ne h(x_i)) \\ f(h) = \int p(x) I(y_x \ne h(x))\,dx \end{cases}$ $\Rightarrow$ $\because$ S is sampled from $p(x)$. $\therefore E(f(h;S)) = f(h)$.

$\therefore \Downarrow$

$$Pr(|f(h;S)-f(h)| \ge \epsilon) \le 2e^{-\frac{\epsilon^2 n}{3\mu}}$$

when $n \ge \frac{3}{\epsilon^2}\ln\frac{2|H|}{\delta}$ $\therefore Pr \le 2e^{-\frac{\ln\frac{2|H|}{\delta}}{f(h)}}$.

$\because f(h) = \int p(x) I(y_x \ne h(x))\,dx$ $\because I \in \{0,1\}$ $\therefore f(h) \in [0,1]$.

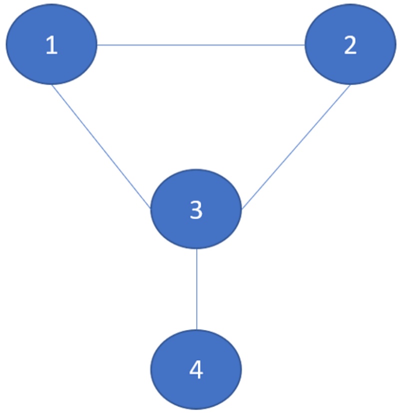
$$\therefore Pr \le 2e^{-\frac{\ln\frac{2|H|}{\delta}}{f(h)}} \le \frac{1}{2e^{\frac{\ln 2|H|/\delta}{f(h)}}} \le \frac{1}{\frac{|H|}{\delta}} = \frac{\delta}{|H|}$$

set $\delta' = \frac{\delta}{|H|}$ so that it statifies uniform convergenc

- Absolute error Lec2, Page 26
  - $\Pr(|\bar{X}-\mu| \ge \epsilon) \le 2e^{-\frac{\epsilon^2 n}{3\mu}}$, $\mu = E[\bar{X}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i]$

$f(h;S) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \ne h(x_i))$ $f(h) = \int p(x) I(y_x \ne h(x))dx$

Q5 (15 points) We have an undirected graph as shown in the figure.



(1) (10 points) Suppose the PageRank values of the nodes in the graph are x_1, x_2, x_3, x_4 for node 1, 2, 3, 4, respectively. List 4 linear equations indicating the relationships among x_1, x_2, x_3, and x_4. The probability of jumping to a random node at each step is assumed to be α.

(2) (5 points) Prove that if $0 < \alpha < 1$, node 3 has the greatest PageRank value, that is, $x_3 = \max\{x_1, x_2, x_3, x_4\}$.

- Matrix form

$$\pi^\top = (1-\alpha)\pi^\top \mathbf{P} + \frac{\alpha \mathbf{1}^\top}{n}$$

- PageRank as an eigenvector ($\mathbf{E}$ is a matrix with all 1 entries)

$$\pi = \left((1-\alpha)\mathbf{P}^\top + \frac{\alpha}{n}\mathbf{E}\right)\pi$$

$$x_i = \frac{\alpha}{N} + (1-\alpha) \sum_{j \in \text{Neighbors}(i)} \frac{x_j}{\deg(j)}$$

$$x_1 = \frac{\alpha}{4} + (1-\alpha)\left(\frac{x_2}{2} + \frac{x_3}{3}\right)$$

$$x_2 = \frac{\alpha}{4} + (1-\alpha)\left(\frac{x_1}{2} + \frac{x_3}{3}\right)$$

$$x_3 = \frac{\alpha}{4} + (1-\alpha)\left(\frac{x_1}{2} + \frac{x_2}{2} + \frac{x_4}{1}\right)$$

$$x_4 = \frac{\alpha}{4} + (1-\alpha) \cdot \frac{x_3}{3}$$

$$\Downarrow$$

$$x_3 = \frac{3}{4} + \frac{3}{\alpha^2+\alpha-8}$$

$$x_4 = \frac{1}{4} + \frac{1-\alpha}{\alpha^2+\alpha-8}$$

$$x_1 = x_2 = \frac{\alpha - 4}{2(\alpha^2+\alpha-8)}$$

(symmetrical)

$$x_3 - x_2 = \frac{3\alpha^2 + \alpha - 4}{4(\alpha^2+\alpha-8)}$$

Q6 (15 points) Suppose we have an undirected graph $G =< V, E >$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Suppose that G is connected, which means for any pair of nodes $u, v \in V$, u and v can reach each other by traveling edges in E. Denote by d_v the degree of the node v, which is the number of neighbor nodes of v in the graph. Let $n_v = \frac{d_v}{D}$ be the normalized degree of v, where $D = \sum_{v \in V} d_v$ is the sum of the degrees of all nodes. We simulate a random walk of M steps as the following: (1) the starting point is randomly selected (that is, each node is selected as the starting point with probability $\frac{1}{|V|}$), and (2) at each step, we randomly jump to a neighbor node of the current node. Let $\mathbf{p} = (p_1, p_2, \dots, p_{|V|})$ be the long-term average probability distribution, where p_v denotes the probability that we randomly choose a step k from the M steps and v is visited at step k. Similarly, let $\mathbf{n} = (n_1, n_2, \dots, n_{|V|})$ be the normalized degree distribution.

(1) (8 points) Prove that for any connected undirected graph G, $\mathbf{p}$ converges to $\mathbf{n}$ as M increases, that is, $\lim_{M \to \infty} \mathbf{p} = \mathbf{n}$. (Hint: use the reversibility condition)

(2) (7 points) Suppose the graph G is a social network where each node v is a social network user and we know the average degree $\bar{d} = \frac{\sum_{v \in V} d_v}{|V|}$. Each user v has a label $x_v \in \{0,1\}$ indicating her/his opinion to a new product. Therefore, $\frac{1}{|V|} \sum_{v \in V} x_v$ is the average opinion which may reflect how popular the new product will be. Suppose we can only use the random walk described above to collect users' opinions and we have a collection of samples $\{(v_1, x_{v_1}), (v_2, x_{v_2}), \dots . (v_M, x_{v_M})\}$. Design an aggregate function over the M samples to estimate $\frac{1}{|V|} \sum_{v \in V} x_v$. (Hint: for each user v_i collected in the samples, we know her/his degree and you may want to use this information)

Q6:

(1) the ~~Detailed~~ Detailed Balance Condition / Reversibily holds when

$\pi_i P_{ij} = \pi_j \cdot P_{ji}$

at the stage of $\vec{n}$ $\boxed{\text{if } i,j \text{ are not neighbors, both prob. are 0 also satify Reversibility}}$

for node i, the prob. transfor to j is $\frac{1}{d_i}$, d_i is the degree of i in random walk

$\therefore \pi_i \cdot P_{ij} = \frac{d_i}{D} \cdot \frac{1}{d_i} = \frac{1}{D}$ if i,j are neighbours.

similarly for node j. $\pi_j \cdot P_{ji} = \frac{d_j}{D} \cdot \frac{1}{d_j} = \frac{1}{D} = \pi_i P_{ij}$

$\therefore$ at the stage of $\vec{n}$ it could satisfy Detailed Balance Condition

it could converge to stationary state. then we will demonstrate $\vec{n}$ is the stationary state.

$$n^T P = \left(\frac{d_1}{D}, \frac{d_2}{D}, \cdots \frac{d_n}{D}\right) \begin{pmatrix} P_{11} & P_{12} & \cdots \\ & & \vdots \\ & & P_{nn} \end{pmatrix} = (X_1, X_2 \cdots X_n).$$

$\because X_1 = \sum_{i=1}^{n} \frac{d_i}{D} \cdot P_{i1} = \frac{1}{D} \sum_{i=1}^{n} d_i P_{i1} = \cancel{\frac{1}{D} \sum_{i=1}^{n} d_i}$

$\because$ in Random walk $P_{ij} = \frac{1}{d_i}$ if i,j neighbours, else 0

$\therefore X_1 = \frac{1}{D} \sum_{i=1}^{n} d_i \cdot \frac{1}{d_i} \cdot I(1, i \text{ neighbours})$

$\because$ node1 has d_1 neighbours $\therefore X_1 = \frac{d_1}{D}$

for other entries, ~~all of~~ the equation holds

$\therefore n^T P = n^T$

$\therefore \vec{p}$ converges to $\vec{n}$ when M increases. $\lim_{M\to\infty} \vec{p} = \vec{n}$

(2) In the whole graph. $\text{sign}(\sum_{v \in V} X_v)$ represents overall opinion

Basing on Question 1, the $\vec{p}$, which is the long-term average prob. distribution, will converge to the $\vec{n}$.

which is $(n_1, n_2 \cdots n_n)$.

so if there are enough collections of samples

$C_1 = \{(V_1, X_{v_1}) \cdots (V_m, X_{v_m})\}$ $C_2 = \{(V_1, X_{v_1}) \cdots (V_m, X_{v_m})\}$

$\therefore E[\text{sign}(\sum_{i=1}^{M} C_i)]$, $C_i = X_{v_1} + X_2 + \cdots X_{v_m}$, $X_v \in C_i$.

$\because$ according to Q1.

$E[\text{sign}(\sum_{i=1}^{M} C_i)] = \text{sign}\left[\sum_{v \in V} X_v \cdot \frac{d_{x_v}}{D}\right]$

$\therefore E[\text{sign}(\sum_{i=1}^{M} X_i / d_i)] = \text{sign}(\sum_{v \in V} X_v)$.

$\therefore$ the aggregate function over the M samples are.

$$\text{sign} \sum_{i=1}^{M} \frac{X_{v_i}}{d_{v_i}}$$