Foundations of Data Science

Lecture 14: Summarization

MING GAO

DaSE@ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

Jan. 3, 2017

Outline

Schedule

Background

DS overview

Probability and Statistics

- Random variable, distribution and expectation
- Estimation, hypothesis testing, regression
- Markov chain, Bayesian network, and Markov random field

Schedule

Algebra

- Vector, Matrix, and operation
- Eigenvalues and eigenvectors
- Matrix factorization

Graph

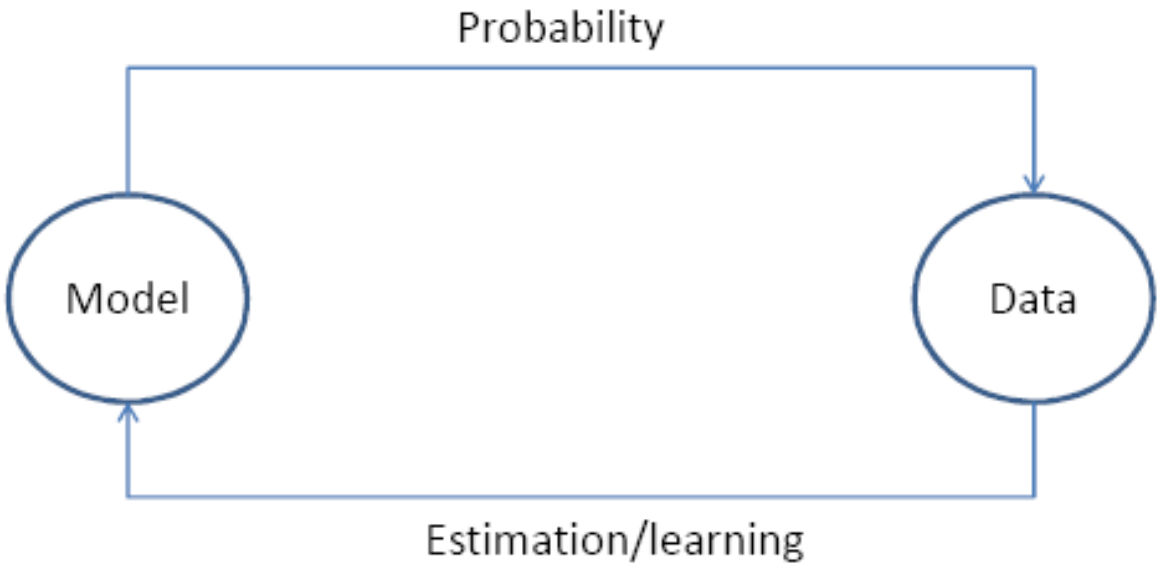
- Graph and patterns
- Centrality
- Proximity
- Information cascade

Schedule

Streaming Data

- Streaming overview
- Count-min sketch, frequent item mining, and moment estimation
- Hashing: bloom filter and LSH

Big picture of probability and statistics



Joint probability distribution

Given a set of random variables $X_1, X_2, \cdots, X_n$

- How to calculate probability $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$
  - Chain rule is always true

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$
$$= P(X_1 = x_1)\Pi_{i=2}^{n} P(X_i = x_i | X_{i-1} = x_{i-1}, \cdots, X_1 = x_1)$$

- If X_i and X_j are independent,
  $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \Pi_{i=1}^{n} P(X_i = x_i)$.
- In directed graph, Bayesian network
  $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \Pi_i P(X_i = x_i | pa_i)$;
- In undirected graph, $P(X)$ can be represented as

$$P(X) = \frac{1}{Z}\Pi_{c \in \mathcal{C}} \psi_c(x_c)$$

where $\mathcal{C}$ is the set of all cliques, and $Z = \sum_x \Pi_{c \in \mathcal{C}} \psi_c(x_c)$ is the "partition function";

Markov chain

A stochastic processes $\{X_t | t \in T\}$ is a collection of random variables. The index t is often called time, as the process represents the value of a random variable changing over time. Let Ω be the set of values assumed by the random variables X_t. We call each element of Ω a state, as X_t represents the state of the process at time t.

Definition of Markov property

A process $X_0, X_1, \cdots$ satisfies the Markov property if

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, \cdots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

for all n and all $x_i \in \Omega$.

Definition of Markov chain

A stochastic process $X_0, X_1, \cdots$ of discrete time and discrete space is a Markov chain if it satisfies the Markov property

A random walk on a graph can be modeled as a Markov chain

Transition probability and convergence

Definition

Let a Markov chain have $P_{x,y}^{(t+1)} = P[X_{t+1} = y | X_t = x]$, and the finite state space be $\Omega = [n]$. This gives us a transition matrix $P^{(t+1)}$ at time t. The transition matrix is an $N \times N$ matrix of nonnegative entries such that the sum over each row of $P^{(t)}$ is 1, since $\forall n$ and $\forall x_i \in \Omega$

$$\sum_y P_{x,y}^{(t+1)} = \sum_y P[X_{t+1} = y | X_t = x] = 1$$

- Let $\pi^{(t)}$ be the state distribution of the chain at time t, that $\pi_x^{(t)} = P[X_t = x]$.
- Let $X_0, X_1, \cdots$, be an irreducible aperiodic Markov chain with finite state space Ω, transition matrix P, and arbitrary initial distribution $\pi^{(0)}$. Then, there exists a unique stationary distribution π such that $\pi P = \pi$, and $\lim_{t\to\infty} \pi^{(0)} P^t = \pi$.

Bayes rule

- We know that $P(gender = 0) = 0.6$
- If we also know that the length of students hair is h_0, then how this affects our belief about her/his gender?

$$P(gener = 0|hair = h_0) = \frac{P(gender = 0)P(hair = h_0|gender = 0)}{P(hair = h_0)}$$

- A simple naive Bayes model for antispam, we have collected a vocabulary, denoted as $X_1, X_2, \cdots, X_n$

$$\begin{aligned} & P(S|X_1, X_2, \cdots, X_n) \\ &= \frac{P(X_1, X_2, \cdots, X_n|S)P(S)}{P(X_1, X_2, \cdots, X_n|S)P(S) + P(X_1, X_2, \cdots, X_n|\overline{S})P(\overline{S})} \end{aligned}$$

- where $P(X_1, X_2, \cdots, X_n|S) = \Pi_{i=1}^{n} P(X_i|S)$ and $P(X_1, X_2, \cdots, X_n|\overline{S}) = \Pi_{i=1}^{n} P(X_i|\overline{S})$

Maximum likelihood estimation

Finding the parameter values that maximizes the likelihood

- Suppose there is a sample $x_1, x_2, \cdots, x_n$ of n *i.i.d.* observations coming from pdf. $f(\cdot)$ (unknown).
- We first specifies the joint density function for all observations as $\mathcal{L}(\theta|x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$.
- MLE: $\widehat{\theta} = \arg\max_{\theta\in\Theta} \ln \mathcal{L}(\theta|x_1, x_2, \cdots, x_n)$ (more convenient).

For example

For a sample observing from a normal distribution

$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- $\ln \mathcal{L}(\mu, \sigma) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$
- $\frac{\partial}{\partial\mu} \ln \mathcal{L}(\mu, \sigma) = 0$ and $\frac{\partial}{\partial\sigma} \ln \mathcal{L}(\mu, \sigma) = 0$
- $\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu})^2$

Expectation maximization algorithm

EM algorithm

The EM algorithm is used to find (locally) MLEs of a model when the equations cannot be solved directly, such as missing value, likelihood contain latent variables.

E-step Cont.

$$Q(\theta|\theta^k) = E_{Z|X,\theta^k} \ln \mathcal{L}(\theta; x, z) = \sum_{i=1}^{n} E_{Z|X,\theta^k} \ln \mathcal{L}(\theta; x_i, z_i)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2} P(Z_i = j|x_i; \theta^{(k)}) \ln \mathcal{L}(\theta_j; x_i, z_i) = \sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^k \ln \mathcal{L}(\theta_j; x_i, z_i)$$

M-step (for j = 1, 2)

$$\tau_j^{k+1} : \frac{\sum_i T_{j,i}^k}{\sum_i \sum_j T_{j,i}^k}, \mu_j^{k+1} : \frac{\sum_i T_{j,i}^k x_i}{\sum_i T_{j,i}^k}, \Sigma_j^{k+1} : \frac{\sum_i T_{j,i}^k (x_i - \mu_j^{k+1})(x_i - \mu_j^{k+1})^T}{\sum_i T_{j,i}^k}$$

Exponential family

An exponential family is a set of distributions

$$p(x;\theta) = \frac{1}{Z(\theta)} Exp(\theta^T \phi(x))$$
$$= Exp(\theta^T \phi(x) - A(\theta))$$

parameterized by $\theta \in \Theta \subset \mathbb{R}^d$, $Z(\theta) = \sum_x Exp(\theta^T \phi(x))$ and $A(\theta) = \log Z(\theta)$ is the "log-partition function". We care because: (1) Many interesting properties; (2) Undirected models are an exponential family

Examples

- Bernoulli, Gaussian, Binomial, Poisson, Exponential, Weibull, Laplace, Gamma, Beta, Multinomial, Wishart distributions are all exponential families.
- $\frac{\partial}{\partial \theta_i} A(\theta) = E(\phi(x)(i))$

Probabilistic inequality

Markov inequality

If X is any r.v. and $0 < a < +\infty$, then $P(X > a) \leq \frac{E(X)}{a}$.

Chebyshevs inequality

If r.v. X has mean and variance $\mu = E(X)$ and $\sigma^2 = E[(X - \mu)^2]$, then

$$P(|X - \mu| > a) \leq \frac{\sigma^2}{a^2} \text{ or } P(|X - \mu| > aE(X)) \leq \frac{\sigma^2}{a^2 E(X)^2}$$

Chernoff bound

Let X_i be a sequence of independent r.v.s with $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. r.v. $X = \sum_{i=1}^{n} X_i$ and $\mu = \sum_{i=1}^{n} p_i$.

- $P(X < (1 - \epsilon)\mu) < \exp(-\mu\epsilon^2/2)$;
- $P(X > (1 + \epsilon)\mu) < \exp(-\mu\epsilon^2/4)$;
- $P(|X - \mu| > \epsilon\mu) < \exp(-\mu\epsilon^2/3)$.

Application

Morris algorithm [Morris 1978]

1: initialize $X \leftarrow 0$;
2: for each update, increment X with probability $\frac{1}{2^X}$;
3: for a query, output $\widehat{n} = 2^X - 1$.

- $E2^{X_N} = N + 1$, and $E2^{2X_N} = \frac{3}{2}N^2 + \frac{3}{2}N + 1$.
- Tug of War–boosting success probability:
  - we obtain independent estimators $\widehat{n}_1, \cdots, \widehat{n}_s$ from independent instantiations of Morris' algorithm, denoted as Morris+.
  - We run t instantiations of Morris+, each with failure probability $\frac{1}{3}$. We then output the median estimate from all the t Morris+.
  - Define $Y_i = \begin{cases} 1, & \text{if the ith Morris+ instantiation succeeds;} \\ 0, & \text{otherwise.} \end{cases}$
  - Note that $\mu = E\sum_i Y_i = \frac{2t}{3}$. Then by the Chernoff bound, $P(\sum_i Y_i \leq \frac{t}{2}) \leq P(|\sum_i Y_i - \mu| \geq \frac{1}{4}\mu) \leq 2\exp(-2t/3(1/4)^2/3) < 2\exp(-t/3) < \delta$,

Vector and matrix

Vector, matrix and operations

- Vector, such as entity, set, distribution, and latent vector, etc.
- Matrix, such as interactions between entities, a set of entities, etc.
- Operations: such as addition, manipulation, determinant, inverse, trace, derivatives, etc.

Derivatives

Type	scalar	vector	matrix
scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial Y}{\partial x}$
vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
matrix	$\frac{\partial y}{\partial X}$		

Matrix derivatives

Derivatives by scalar

Assume that $x, y, a \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^{n\times 1}$, $X, Y, A \in \mathbb{R}^{n\times m}$, and $\mathbf{y}, \mathbf{a} \in \mathbb{R}^{m\times 1}$.
Let a, $\mathbf{a}$ and A be constant scalar, vector and matrix.

- $\frac{\partial y}{\partial x}$ and $\frac{\partial a}{\partial x} = 0$
- $\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \cdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$ and $\frac{\partial \mathbf{a}}{\partial x} = \mathbf{0}$ (vector)
- $\frac{\partial Y}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1m}}{\partial x} \\ \cdots & \cdots & \cdots \\ \frac{\partial y_{n1}}{\partial x} & \cdots & \frac{\partial y_{nm}}{\partial x} \end{bmatrix}$ and $\frac{\partial A}{\partial x} = \mathbf{0}$ (matrix)

Matrix derivatives

Derivatives by vector

- $\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \cdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}^T$ and $\frac{\partial a}{\partial \mathbf{x}} = \mathbf{0}^T$ (vector)
- $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$, $\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0}$ (matrix) and $\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I}$ (matrix)

Derivatives by matrix

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{n1}} \\ \cdots & \cdots & \cdots \\ \frac{\partial y}{\partial x_{1m}} & \cdots & \frac{\partial y}{\partial x_{nm}} \end{bmatrix} \text{ and } \frac{\partial a}{\partial X} = \mathbf{0}^T \text{ (matrix)}$$

Common properties of matrix derivatives

Properties

c1 $\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}^T$

c2 $\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T$

c3 $\frac{\partial (\mathbf{x}^T\mathbf{a})^2}{\partial \mathbf{x}} = 2\mathbf{x}^T\mathbf{a}\mathbf{a}^T$

c4 $\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} = A$ and $\frac{\partial \mathbf{x}^T A}{\partial \mathbf{x}} = A^T$

c5 $\frac{\partial \mathbf{x}^T A\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(A + A^T)$

Proof

c2 Let $s = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^n x_i^2$. Then, $\frac{\partial s}{\partial x_i} = 2x_i$. So, $\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T$.

c3 Let $s = \mathbf{x}^T\mathbf{a}$. Then $\frac{\partial s^2}{\partial x_i} = 2s\frac{\partial s}{\partial x_i} = 2sa_i$. Thus, $\frac{\partial (\mathbf{x}^T\mathbf{a})^2}{\partial \mathbf{x}} = 2\mathbf{x}^T\mathbf{a}\mathbf{a}^T$.

Linear regression

Problem

Given a set of n points (x_i, y_i) on a scatterplot, find the relationship between x and y: $\widehat{y_i} = \beta_0 + \beta_1^T x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

- We can write linear regression in matrix form:

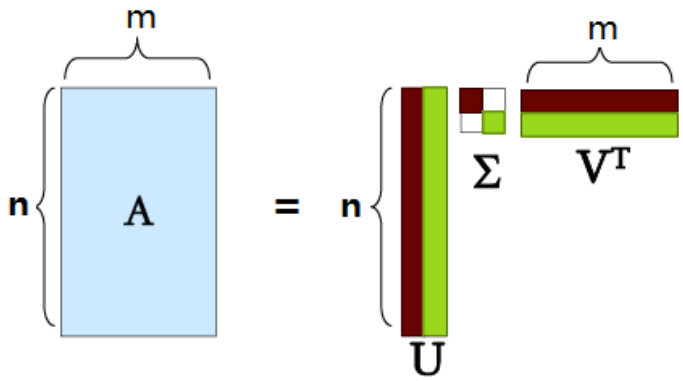
$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \epsilon_1 \\ \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \cdots \\ \beta_0 + \beta_1 x_n + \epsilon_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}$$

- Rewrite as $Y_{n\times 1} = X_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1}$, thus residuals are $\epsilon = Y - X\beta$. We would like to minimize sum of squared residuals $\epsilon^T\epsilon = (Y - X\beta)^T(Y - X\beta)$
- $\frac{\partial}{\partial\beta}(Y - X\beta)^T(Y - X\beta) = -2X^T(Y - X\beta) = 0$, thus $\beta = (X^TX)^{-1}XY$

SVD

Definition

Any real $m \times n$ matrix A can be decomposed uniquely as $A_{[n\times m]} \sim U_{[n\times r]} D_{[r\times r]} V^T_{r\times m}$ where U, V are orthogonal matrix, D is a diagonal matrix (non-negative real values called singular values).

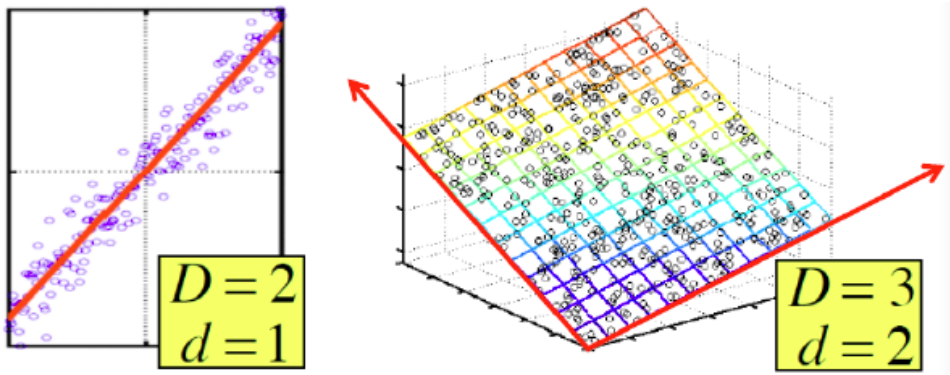


- A: an input $n \times m$ data matrix, e.g., n users and m items.
- U: left singular vectors in a $n \times r$ matrix, e.g., n users and r interests.
- D: a $r \times r$ diagonal matrix, e.g., strength of each interest.
- V: right singular vectors in a $r \times m$ matrix, e.g., r interests and m

SVD cont.

Properties

- It is always possible to decompose a real matrix A into $A = UDV^T$
  - U, D, V are unique, and D is a diagonal matrix.
  - U, V are column orthogonal (i.e., $U^T U = I$ and $V^T V = I$)
  - Entries of D are positive and sorted in decreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.
- Axes of this subspace are effective representation of the data.



Assumptions

- Data lies on or near a low $d-$dimensional subspace.
- Axes of this subspace are effective representation of the data.

Matrix factorization

$$J = \min_{q^*, p^*} \frac{1}{2} \sum_{(u,i) \in \mathcal{K}} (r_{ui} - p_u^T q_i)^2,$$

where r_{ui} is the known rating of user u for item i, $\widehat{u}_{ui} = p_u^T q_i$ is the predicted rating given by u for i, and $\mathcal{K} = \{(u,i) | r_{ui} \text{ is observed}\}$.

- The error between the predicted rating and the real rating: $e_{ui}^2 = (r_{ui} - \widehat{r}_{ui})^2 = (r_{ui} - \sum_{k=1}^{d} p_{uk} q_{ki})^2$.
- Batch gradient descent algorithm and stochastic gradient descent algorithm.
- To avoid overfitting, regularization is a common approach to address the problem.

$$\min_{q^*, p^*} J = \frac{1}{2} \Big[\sum_{(u,i) \in \mathcal{K}} (r_{ui} - q_i^T p_u)^2 + \lambda (\|Q\|^2 + \|P\|^2) \Big],$$

Representation of graph

Representation

Given a finite graph $G = (V, E)$, an adjacency matrix A is a $|V| \times |V|$ matrix, whose elements indicate whether pairs of vertices are adjacent or not in the graph.

- Let $D = diag(d_1, d_2, \cdots, d_n)$ be a diagonal matrix, and $P = D^{-1}A$ is the transition probability matrix of the discrete-time Markov chain X is the symmetric random walk on G.
- Combinatorial Laplacian of G is $L = D - A$.
- Normailized Laplacian of G is $\mathcal{L} = D^{-1/2}LD^{-1/2}$.

Important properties

- P has an eigenvalue $1 - \lambda_i$, where λ_i is an eigenvalue of $\mathcal{L}$.
- The regularization of graph G: $F^T\mathcal{L}F = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}}\right)^2$.

Centrality

- Centrality
  - Degree
  - Eigenvector
  - Closeness
  - Betweenness
  - Clustering coefficient
- Weighted graph centrality
- PageRank and HITS
  - PageRank
  - HITS
  - Co-HITS: iterative framework, regularization framework, and Bayesian network.
- Graph robustness

Proximity and information diffusion

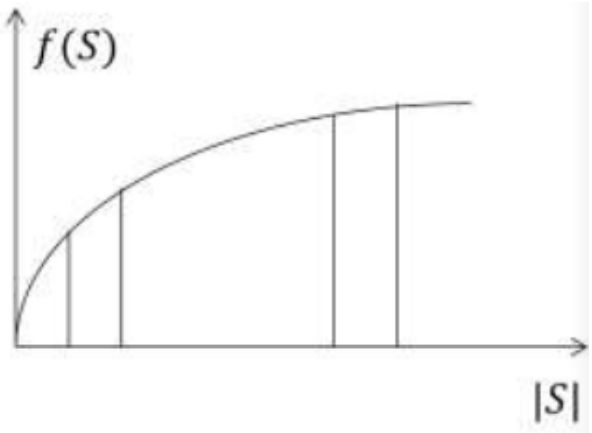
Proximity

- Community detection
- Node proximity
  - Simple approaches
  - Graph-theoretic approaches
  - SimRank
  - Random walk based approaches
- Network proximity
  - Known nodes
  - Unknown nodes, such as isomorphism, and graph kernel (symmetry and positive semi-definite).

Information diffusion

- Information cascade: IC and LT models
- Solutions: submodular.

Submodular



Definition

A functions $f : 2^V \to R$ is a submodular if for all $S, T \subseteq V$, we have that

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T).$$

We have two equivalent definitions:

- Diminishing marginal return: for all $S \subseteq T \subseteq V$, all $v \in V \setminus T$,

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T).$$

- Group diminishing returns: for all $S \subseteq T \subseteq V$, and $C \subseteq V \setminus T$,

$$f(S \cup C) - f(S) \geq f(T \cup C) - f(T).$$

Proof of equivalence

Proof

$f(S\cup T)+f(S\cap T)\leq f(S)+f(T) \Leftrightarrow f(S\cup\{v\})-f(S)\geq f(T\cup\{v\})-f(T).$

- $\Rightarrow$: let $S\subset T$, consider two sets $S\cup\{v\}$ and T, if $v\notin T$, then $f(S\cup\{v\}\cup T)+f((S\cup\{v\})\cap T)\leq f(S\cup\{v\})+f(T)$.
  Note that $f(S\cup\{v\}\cup T)=f(T\cup\{v\})$ and $f((S\cup\{v\})\cap T)=f(S)$.
  Thus, we have $f(S\cup\{v\})-f(S)\geq f(T\cup\{v\})-f(T)$.
- $\Leftarrow$: let $T\backslash S=\{v_1,v_2,\cdots,v_k\}$, $T_j=\{v_1,v_2,\cdots,v_j\}$, $A_j=(S\cap T)\cup T_j$, and $B_j=S\cup T_j$, then we have $f(A_j\cup\{v_{j+1}\})-f(A_j)\geq f(B_j\cup\{v_{j+1}\})-f(B_j)$ for $j=0,1,2,\cdots,k-1$.
  Summing up all these equations, we have $f(S\cup T)+f(S\cap T)\leq f(S)+f(T)$.

Hill-climbing algorithm

1: initialize $S = \emptyset$;
2: for $i = 1$ to k do
3: select

$$u = \arg \max_{v \in V \setminus S} [f(S \cup \{v\}) - f(S)];$$

4: $S = S \cup \{u\}$;
5: end for
6: output S;

Theorem

If the set function f is monotone and submodular with $f(\emptyset) = 0$, then the greedy algorithm achieves $(1 - \frac{1}{e})$ approximation ratio, that is, the solution S found by the algorithm satisfies:

$$f(S) \geq (1 - \frac{1}{e}) \max_{S' \subseteq V, |S'| = k} f(S').$$

where f is monotonicity if $f(S) \leq f(T)$ for all $S \subseteq T \subseteq V$.

Streaming data algorithm

Problems

- Item frequency: sampling, count sketch, and count min sketch
- Distinct element estimation
- Frequency moment estimation

Algorithms

We shall typically seek to compute only an approximation of the true value of $\phi(\sigma)$ (provably not be computed exactly using sublinear space).

- Definition 1: let $\mathcal{A}(\sigma)$ denote the output of a randomized streaming algorithm $\mathcal{A}$ on input σ and ϕ be the function that $\mathcal{A}$ is supposed to compute. We say that the algorithm $(\epsilon, \delta)-$approximation ϕ if we have $P\left[|\frac{\mathcal{A}(\sigma)}{\phi(\sigma)} - 1| > \epsilon\right] \leq \delta$.
- Definition 2: In the above setup, the algorithm $(\epsilon, \delta)-$additively approximation ϕ if we have $P[|\mathcal{A}(\sigma) - \phi(\sigma)| > \epsilon] \leq \delta$.

Boosting success probability

Strategies

- Tug of war
  - $O(\frac{1}{\epsilon^2})$ independent copies of the algorithm and average their outputs (Chebyshev's inequality).
  - We then output the median estimate from all the $O(\log 1/\delta)$ instantiations (Chernoff's inequality).
  - For example, Morris $++$, count-sketch, min-hash, etc.
- Min-skecth
  - $O(\frac{1}{\epsilon})$ independent copies of the algorithm and average their outputs (Markov's inequality).
  - We then output the minimum estimate from all the $O(\log 1/\delta)$ independent instantiations (Min-sketch).
  - For example, count min sketch, etc.

We can boost the success probabilities of many algorithms in these manners.

Take-home messages

- Overview
- Probability and statistics theory
- Algebraic foundations
- Graph mining
- Streaming data mining