Graphs in Machine Learning

Michal Valko

DeepMind Paris and Inria Lille

TA: Omar Darwiche Domingues with the help of Pierre Perrault

Partially based on material by: Andreas Krause, Branislav Kveton, Michael Kearns
Piazza for Q&A’s

Purpose

▶ registration for the class
▶ register with your school email and full name
▶ online course discussions and announcements
▶ questions and answers about the material and logistics
▶ students encouraged to answer each others’ questions
▶ homework assignments
▶ virtual machine link and instructions
▶ draft of the slides before the class

https://piazza.com/ens_cachan/fall2019/mvagraphsml NO EMAILS!

class code given during the class
Scribes 2019/2020

Details

▶ number of people that can volunteer is limited (to 2x number of lectures)
▶ first-come-first serve by writing your name below (also in the resource section on piazza)
▶ use https://www.overleaf.com/articles/tml-scribe-108/wqhhzhgdprbdasthetemplate
▶ grade bonus: -0.5-2 points depending on the quality
▶ important: you work in pairs to proofread what the other is typing
▶ example:
  http://imagine.enpc.fr/~oboizinsg/teaching/mva_gm/fall2017/
▶ deadline: 1 month after the lecture

https://piazza.com/ens_cachan/fall2019/mvagraphsml live now
Graphs from social networks

▶ people and their interactions

▶ directed (Twitter) and undirected (Facebook)

▶ structure is rather a phenomena

▶ typical ML tasks
  ▶ advertising
  ▶ product placement
  ▶ link prediction (PYNK)
Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{ A\, \text{Alice}, B\, \text{Bob}, C\, \text{Charlie}, D\, \text{Dorothy}, E\, \text{Eric}, F\, \text{Fiona}\} \]

\[ F(S) = \text{Expected number of people influenced when targeting } S \subseteq V \text{ under some propagation model - e.g., cascades} \]

How would you choose the target customers?

highest degree, close to the center, ...

Maximizing the Spread of Influence through a Social Network

Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{\text{Alice, Bob, Charlie, Dorothy, Eric, Fiona}\} \]

Maximizing the Spread of Influence through a Social Network
Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{ \text{Alice, Bob, Charlie, Dorothy, Eric, Fiona} \} \]
Success story #1 Product placement - problem

Who should get free cell phones?
\[ V = \{Alice, Bob, Charlie, Dorothy, Eric, Fiona\} \]

Maximizing the Spread of Influence through a Social Network
Who should get free cell phones?

\[ V = \{ \text{Alice, Bob, Charlie, Dorothy, Eric, Fiona} \} \]
Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{\text{Alice, Bob, Charlie, Dorothy, Eric, Fiona}\} \]

Maximizing the Spread of Influence through a Social Network
Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{ \text{Alice, Bob, Charlie, Dorothy, Eric, Fiona} \} \]

Maximizing the Spread of Influence through a Social Network

Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{A\text{lice}, B\text{ob}, C\text{harlie}, D\text{orothy}, E\text{ric}, F\text{iona}\} \]

Maximizing the Spread of Influence through a Social Network
Success story #1  Product placement - problem

Who should get free cell phones?

\[ V = \{\text{Alice}, \text{Bob}, \text{Charlie}, \text{Dorothy}, \text{Eric}, \text{Fiona}\} \]
Success story #1 Product placement - problem

Who should get free cell phones?

\[ V = \{Alice, Bob, Charlie, Dorothy, Eric, Fiona\} \]

\[ F(S) = \text{Expected number of people influenced when targeting} \]
\[ S \subseteq V \text{ under some propagation model - e.g., cascades} \]

How would you choose the target customers?

highest degree, close to the center, ...

Maximizing the Spread of Influence through a Social Network
Submodularity: Definition

A set function on a discrete set $A$ is **submodular** if for any $S \subseteq T \subseteq A$ and for any $e \in A \setminus T$

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

Example: $S = \{\text{stuff}\} = \{\text{bread, apple, tomato, ...}\}$

$f(V) =$ cost of getting products $V$

- $f(\{\text{bread}\}) = c(\text{bakery}) + c(\text{bread})$
- $f(\{\text{bread, apple}\}) = c(\text{bakery}) + c(\text{bread}) + c(\text{market}) + c(\text{apple})$
- $f(\{\text{bread, tomato}\}) = c(\text{bakery}) + c(\text{bread}) + c(\text{market}) + c(\text{tomato})$
- $f(\{\text{bread, tomato, apple}\}) = c(\text{bakery}) + c(\text{bread}) + c(\text{market}) + c(\text{tomato}) + c(\text{apple})$

Adding an apple to the smaller set costs more!

$$\{\text{bread}\} \subseteq \{\text{bread, tomato}\}$$

$$f(\{\text{bread, apple}\}) - f(\{\text{bread}\}) > f(\{\text{bread, tomato, apple}\}) - f(\{\text{tomato, bread}\})$$

Diminishing returns: Buying in bulk is cheaper!
Submodularity: Application

**Objective:** Find \( \text{arg max}_{S \subseteq A, |S| \leq k} f(S) \)

**Property:** NP-hard in general

**Special case:** \( f \) is also nonnegative and monotone.

**Other examples:** information, graph cuts, covering, ...

---

**Link to our product placement problem on a social network graph?**

- submodular?, nonnegative?, monotone?, \( k \)?

---


Let \( S^* = \text{arg max}_{S \subseteq A, |S| \leq k} f(S) \) where \( f \) is monotonic and submodular set function and let \( S_{\text{Greedy}} \) be a greedy solution.

Then \( f(S_{\text{Greedy}}) \geq (1 - \frac{1}{e}) \cdot f(S^*). \)
Submodularity: Greedy algorithm

1: **Input:**
2: \( k \): the maximum allowed cardinality of the output
3: \( V \): a ground set
4: \( f \): a monotone, non-negative, and submodular function
5: **Run:**
6: \( S_0 = \emptyset \)
7: for \( i = 1 \) to \( k \) do
8: \( S_i \leftarrow S_{i-1} \cup \{ \text{arg max}_{a \in V \setminus S_{i-1}} [f(\{a\} \cup S_{i-1}) - f(S_{i-1})] \} \)
9: end for
10: **Output:**
11: Return \( S_{\text{Greedy}} = S_k \)

Let \( S^* = \arg \max_{S \subseteq A, |S| \leq k} f(S) \) where \( f \) is monotonic and submodular set function and let \( S_{\text{Greedy}} \) be a **greedy solution**. Then \( f(S_{\text{Greedy}}) \geq (1 - \frac{1}{e}) \cdot f(S^*) \).
Submodularity: Approximation guarantee of Greedy

Let $S_i$ be the $i$-th set selected by Greedy, $S_{\text{Greedy}} = S_k$. We show

$$f(S^*) - f(S_i) \leq (1 - \frac{1}{k})^i \cdot f(S^*).$$

Difference from the optimum before the $i$-th step ...

$$f(S^*) - f(S_{i-1}) \leq f(S^* \cup S_{i-1}) - f(S_{i-1})$$
$$\leq \sum_{a \in S^* \setminus S_{i-1}} (f(\{a\} \cup S_{i-1}) - f(S_{i-1}))$$
$$\leq \sum_{a \in S^* \setminus S_{i-1}} (f(S_i) - f(S_{i-1}))$$
$$\leq k (f(S_i) - f(S_{i-1}))$$

Difference from the optimum after the $i$-th step ...

$$f(S^*) - f(S_i) = f(S^*) - f(S_{i-1}) - (f(S_i) - f(S_{i-1}))$$
$$\leq f(S^*) - f(S_{i-1}) - \frac{f(S^*) - f(S_{i-1})}{k}$$
Submodularity: Graph-related examples

- Influence maximization on networks (current example)
- Maximum-weight spanning trees
- Graph cuts
- Structure learning in graphical models (PGM course)

back to the influence-maximization example …
Success story #1 Product placement - solution

Key idea: Flip coins $c$ in advance $\rightarrow$ “live” edges

MIIA: http://hanj.cs.illinois.edu/pdf/dmkd12_cwang.pdf/
Course: Jeff Billmes at UW
Success story #1 Product placement - solution

Key idea: Flip coins \( c \) in advance \( \rightarrow \) “live” edges

MIIA: http://hanj.cs.illinois.edu/pdf/dmkd12_cwang.pdf/
Course: Jeff Bilmes at UW
**Success story #1 Product placement - solution**

**Key idea:** Flip coins \( c \) in advance \( \rightarrow \) “live” edges

\[
F_c(V) = \text{People influenced under outcome } c \text{ (set cover!)}
\]


Course: Jeff Bilmes at UW
**Key idea:** Flip coins $c$ in advance $\rightarrow$ “live” edges

$f_c(V) = \text{People influenced under outcome } c$ (set cover!)

$f(V) = \sum_c p(c)f_c(V)$ is submodular as well!


Course: Jeff Bilmes at UW
Key idea: Flip coins $c$ in advance → “live” edges

$F_c(V) = \text{People influenced under outcome } c \text{ (set cover!)}$

$F(V) = \sum_c P(c)F_c(V)$ is submodular as well!

Computational issues?

MIIA: http://hanj.cs.illinois.edu/pdf/dmkd12_cwang.pdf/
Course: Jeff Billmes at UW
Success story #1 Product placement - comparison

influence on the ArXiv/Physics co-authorship graph

greedy approximation does better than the centrality measures
Graphs from utility and technology networks

- link services
- power grids, roads, transportation networks, Internet, sensor networks, water distribution networks
- structure is either hand designed or not
- typical ML tasks
  - best routing under unknown or variable costs
  - identify the node of interest

Berkeley’s Floating Sensor Network
Graphs from information networks

- web
- blogs
- wikipedia

- typical ML tasks
  - find influential sources
  - search (PageRank)

Blog cascades (ETH) - submodularity
Objective: **Rank** all web pages (nodes on the graph) by how many other pages link to them and how **important** they are.

basic PageRank is independent of query and the page content

Internet $\rightarrow$ graph $\rightarrow$ matrix $\rightarrow$ stochastic matrix $\mathbf{M} \left( \sum_j M_{ij} = 1 \right)$
Objective: Rank all web pages (nodes on the graph) by how many other pages link to them and how important they are.

Basic PageRank is independent of query and the page content.

Internet $\rightarrow$ graph $\rightarrow$ matrix $\mathbf{M}$ ($\sum_j M_{ij} = 1$)
Objective: Rank all web pages (nodes on the graph) by how many other pages link to them and how important they are.

Basic PageRank is independent of query and the page content.

Internet \rightarrow \text{graph} \rightarrow \text{matrix} \rightarrow \text{stochastic matrix} \ M \ (\sum_j M_{ij} = 1)

Random Surfer Process

What is wrong with it?

- Dangling pages act like sinks.
Success story #2 Google PageRank

http://infolab.stanford.edu/~backrub/google.html:

PageRank can be thought of as a model of user behavior. We assume there is a “random surfer” who is given a web page at random and keeps clicking on links, never hitting “back” but eventually gets bored and starts on another random page.

▶ page is important if important pages link to it
▶ circular definition
▶ importance of a page is distributed evenly
▶ probability of being bored is 15%
Success story #2 Google PageRank

Google matrix: \( G = (1 - p)M + p \cdot \frac{1}{N} \mathbf{1}_{N \times N} \), where \( p = 0.15 \)
Success story #2 Google PageRank

Google matrix: \( G = (1 - p)M + p \cdot \frac{1}{N}1_{N \times N} \), where \( p = 0.15 \)

\( G \) is stochastic \( \text{why?} \) What is \( G_a \) for any \( a \)? We look for \( Gv = 1 \times v \), steady-state vector, a right eigenvector with eigenvalue 1. \( \text{why?} \)

Perron’s theorem: Such \( v \) exists and it is unique if the entries of \( G \) are positive.

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]
Success story #2 Google PageRank

Google matrix: \( G = (1 - p)M + p \cdot \frac{1}{N} \mathbf{1}_{N \times N} \), where \( p = 0.15 \)

\( G \) is stochastic  \( \textbf{why?} \)  What is \( G_a \) for any \( a \)? We look for \( Gv = 1 \times v \), steady-state vector, a right eigenvector with eigenvalue 1. \( \textbf{why?} \)

**Perron’s theorem:** Such \( v \) exists and it is **unique** if the entries of \( G \) are positive.

\[
\begin{align*}
A & : 0.64 \\
B & : 0.38 \\
C & : 0.31 \\
D & : 0.54 \\
E & : 0.38 \\
F & : 0.64
\end{align*}
\]
Success story #2 Google PageRank

Google matrix: $G = (1 - p)M + p \cdot \frac{1}{N} \mathbf{1}_{N \times N}$, where $p = 0.15$

$G$ is stochastic \(\text{why?}\) What is $G_a$ for any $a$? We look for $Gv = 1 \times v$, steady-state vector, a right eigenvector with eigenvalue 1. \(\text{why?}\)

Perron's theorem: Such $v$ exists and it is unique if the entries of $G$ are positive.
Success story #2 Google PageRank

History: [Desikan, 2006]
- The anatomy of a large-scale hypertextual web search engine [Brin & Page 1998]
- US patent for PageRank granted in 2001
- Google indexes 10’s of billions of web pages ($1 \text{ billion} = 10^9$)
- Google serves $\geq 200$ million queries per day
- Each query processed by $\geq 1000$ machines
- All search engines combined process more than 500 million queries per day
Success story #2 Google PageRank

Problem: Find an eigenvector of a stochastic matrix.

- $n = 10^9$ !!!
- luckily: sparse (average outdegree: 7)
- better than a simple centrality measure (e.g., degree)
- power method

$$v_0 = (1_A \ 0_B \ 0_C \ 0_D \ 0_E \ 0_F)^T$$

$$v_1 = Gv_0$$

$$v_{t+1} = Gv_t = G^{t+1}v$$

$v_{t+1} = v_t \implies Gv_t = v_t$ and we found the steady vector.

But wait, $M$ is sparse, but $G$ is dense! What to do?
Graphs from biological networks

- protein-protein interactions
- gene regulatory networks
- typical ML tasks
  - discover unexplored interactions
  - learn or reconstruct the structure

Diffuse large B-cell lymphomas - Dittrich et al. (2008)
Graphs from similarity networks

graph is not naturally given
Graphs from similarity networks

but we can construct it
Graphs from similarity networks

and use it as an abstraction
Graphs from similarity networks

- vision
- audio
- text

- typical ML tasks
  - semi-supervised learning
  - spectral clustering
  - manifold learning

movie similarity
Two sources of graphs in ML

Graph as models for networks

- given as an input
- discover interesting properties of the structure
- represent useful information (viral marketing)
- be the object of study (anomaly detection)

Graph as nonparametric basis

- we create (learn) the structure
- flat vectorial data $\rightarrow$ similarity graph
- nonparametric regularizer
- encode structural properties: smoothness, independence, …
Random Graph Models

**Erdős-Rényi**
- independent edges

**Barabási-Albert**
- preferential attachment

**Stochastic Blocks**
- modeling communities

Watts-Strogatz, Chung-Lu, Fiedler, ....
Erdős number project

- http://www.oakland.edu/enp/ try it!
- an example of a real-world graph
- 401,000 authors, 676,000 edges ($\ll 401000^2 \rightarrow$ sparse)
- average degree 3.36
- average distance for the largest component: 7.64
- 6 degrees of separation [Travers & Milgram, 1967]
- heavy tail
Spanish flu in San Francisco 1918–1919

Small-world phenomenon and diseases

http://rsif.royalsocietypublishing.org/content/4/12/155

Small world: Obvious?
Black death!
Black death: spread

source: catholic.org

https://www.youtube.com/watch?v=EEK6c9Bh5CQ
Links to the other courses

- **Introduction to statistical learning**
  - links to the learning theory on graphs: label propagation, learnability, generalization

- **Reinforcement learning**
  - link to the online learning (bandit) lecture at the end of the semester

- **Advanced learning for text and graph data**
  - data-mining graph course on the topics not covered in this course
  - details on the next slide
Statistical Machine Learning in Paris!

https://sites.google.com/site/smileinparis/home

Speakers: ML researches and Ph.D. students - former MVA students

Topic: Francis Bach: Double descent

Date: Thursday October 17th

Time: 15:00 - 16:30

Place: Inria Paris https://forms.gle/H7VTjNfnLGFKQpYf7
Parisian Deep and Sequential Seminar - New in 2019

Organizers: Brain, DeepMind, Criteo, FB, P6, Dauphine, P11
Next session: soon be to announced
Link above: soon be to active
Time: Tuesday afternoons, next week at 13:30
Place: ENS Cachan, next week at Salle Condorcet

7 lectures + 3 recitations (TDs)

Validation: grades from TDs (40%) + class project (60%)

Research: contact me for internships, Ph.D. theses, projects, etc.

Course website:
http://researchers.lille.inria.fr/~valko/hp/mva-ml-graphs

Contact, online class discussions, and announcements:
https://piazza.com/ens_cachan/fall2019/mvagraphsml
class code given during the class
Graph theory refresher
Graph theory refresher
Graph theory refresher

- 250 years of graph theory
- Seven Bridges of Königsberg (Leonhard Euler, 1735)
- necessary for Eulerian circuit: 0 or 2 nodes of odd degree
- after bombing and rebuilding there are now 5 bridges in Kaliningrad for the nodes with degrees $[2, 2, 3, 3]$
- the original problem is solved but not practical

http://people.engr.ncsu.edu/mfms/SevenBridges/
Similarity Graphs

Input: $x_1, x_2, x_3, \ldots, x_N$

- raw data
- flat data
- vectorial data
Similarity Graphs

Similarity graph: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ — (un)weighted

Task 1: For each pair $i, j$: define a similarity function $s_{ij}$

Task 2: Decide which edges to include

$\varepsilon$-neighborhood graphs – connect the points with the distances smaller than $\varepsilon$

$k$-NN neighborhood graphs – take $k$ nearest neighbors

fully connected graphs - consider everything

This is art (not much theory exists).

http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/publications/Luxburg07_tutorial.pdf
Similarity Graphs: \( \varepsilon \)-neighborhood graphs

- Edges connect the points with the distances smaller than \( \varepsilon \).
- Distances are roughly on the same scale (\( \varepsilon \)).
- Weights may not bring additional info \( \rightarrow \) unweighted.
- Equivalent to: similarity function is at least \( \varepsilon \).
- Theory [Penrose, 1999]: \( \varepsilon = \left( \frac{\log N}{N} \right)^{1/d} \) to guarantee connectivity.
  - \( N \) nodes, \( d \) dimension, https://projecteuclid.org/euclid.aop/1022677261

Practice: choose \( \varepsilon \) as the length of the longest edge in the MST - minimum spanning tree.

What could be the problem with this MST approach?
Similarity Graphs: \( k \)-nearest neighbors graphs

Edges connect each node to its \( k \)-nearest neighbors.

- asymmetric (or directed graph)
  - option OR: ignore the direction
  - option AND: include if we have both direction (mutual \( k \)-NN)
- how to choose \( k \)?
  - \( k \approx \log N \) - suggested by asymptotics (practice: up to \( \sqrt{N} \))
  - for mutual \( k \)-NN we need to take larger \( k \)
  - mutual \( k \)-NN does not connect regions with different density
- why don’t we take \( k = N - 1 ? \)
Similarity Graphs: Fully connected graphs

Edges connect everything.

- choose a “meaningful” similarity function $s$
- default choice:

$$s_{ij} = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)$$

- why the exponential decay with the distance?
- $\sigma$ controls the width of the neighborhoods
  - similar role as $\varepsilon$
  - a practical rule of thumb: 10% of the average empirical std
  - possibility: learn $\sigma_i$ for each feature independently
- metric learning (a whole field of ML)
Similarity Graphs: Important considerations

- calculate all $s_{ij}$ and threshold has its limits ($N \approx 10000$)
- graph construction step can be a huge bottleneck
- want to go higher? (we often have to)
  - down-sample
  - approximate NN
    - LSH - Locally Sensitive Hashing
    - CoverTrees
    - Spectral sparsifiers
- sometime we may not need the graph (just the final results)
- yet another story: when we start with a large graph and want to make it sparse (later in the course)
- these rules have little theoretical underpinning
- similarity is very data-dependent
Next class on Tuesday, October 8th at 13:30!
Michal Valko
contact via Piazza