1. ЛЕКЦИЯ - Введение в графовые представления

1.1. Информация о графах

Граф G – математический объект, состоящий из двух множеств. Одно из них – любое конечное множество, его элементы называются вершинами графа V. Другое множество состоит из пар вершин, эти пары называются ребрами графа E. Другими словами, граф G – структура данных, моделирующих множество объектов и их связей между собой (рис. 1). Каждый объект отождествляется с одной из вершин $v \in V$, где V – их множество, связи между некоторыми двумя объектами описываются с помощью ребер $e = (v_1, v_2) \in E$, где E – их множество. Таким образом, можно записать $G = (V, E)$. Такие графы называются обычными [1].

Графы возникают во многих задачах анализа, особенно, когда есть множество объектов и очевидное множество отношений на нём. Формализм графов заключается в том, что он фокусируется на отношениях между точками (а не на свойствах отдельных точек). Этот формализм графа можно использовать для представления социальных сетей, взаимодействий между лекарствами и белками, взаимодействий между атомами в молекуле или соединений между терминалами в телекоммуникационной сети — и это лишь несколько примеров.

Если ребра – упорядоченные пары, то такой граф называется ориентированным (сокращенно орграф), если же неупорядоченные, то неориентированным. Основное отличие простого орграфа от обычного графа является то, что каждое ребро орграфа имеет направление.

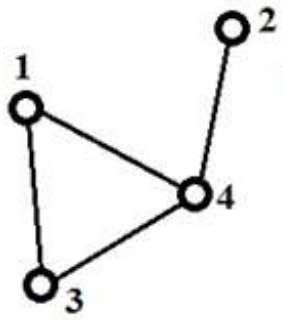
Ребро типа (v, v), т.е. соединяющее вершину с ней же самой, называют петлей.

Пока нас будут интересовать только обычные графы, где одно ребро между каждой парой узлов, нет петель, и где ребра не ориентированы, т.е. $(u, v) \in E \leftrightarrow (v, u) \in E$.

Каждому ребру $e = (v_1, v_2)$ можно приписать некоторый вес w_{ij}, и такой граф называют взвешенным. В случае если граф имеет равновзвешенные ребра, то весами можно пренебречь и такой граф называют невзвешенным [2].

Удобный способ представления графов — через матрицу смежности $A \in R\ |V| \times |V|$. Пусть – граф с вершинами, пронумерованными числами от 1 до n. Матрица смежности – это таблица с n строками и n столбцами, в которой элемент, стоящий на пересечении строки с номером i и столбца с номером j , равен 1, если вершины с номерами i и j смежны, и 0, если они не смежны. Таким образом, можно представить наличие ребер в виде записей в матрице: $A[u, v] = 1$, если $(u, v) \in E$, и $A[u, v] = 0$ в противном случае. Представим матрицу смежности для

рис.1.



$V = \{1, 2, 3, 4\}$

$E = \{(1, 3), (1, 4), (3, 4), (4, 2)\}$

$$\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

Рис. 1. Пример графа.

На главной диагонали матрицы смежности обыкновенного графа всегда стоят нули (нет петель) и эта матрица симметрична относительно главной диагонали (граф неориентированный). При этом, если граф содержит только неориентированные ребра, то A будет симметричной матрицей, но если граф направленный то A не обязательно будет симметричным.

Также графы могут иметь взвешенные ребра, где элементы смежности матрицы являются произвольными вещественными значениями, а не только $\{0, 1\}$. Например, взвешенное ребро на графе взаимодействия может указывать на силу связи между двумя белками.

Помимо различия между неориентированными, направленными и взвешенными ребрами, также необходимо рассматривать графы, которые имеют разные типы ребер. Например, на графиках, представляющих взаимодействие лекарств, мы могли бы захотеть, чтобы разные ребра соответствовали разным побочным эффектам, которые могут возникнуть, когда принимается несколько лекарств одновременно. В этих случаях необходимо расширить обозначение ребра, включив в него тип ребра или отношения, например, $(u; \tau; v) \in E$, тогда можно определить матрицу смежности A для каждого типа ребра. Такие графы будем называть мультиреляционными, и весь граф можно суммировать тензором смежности $A \in \mathbb{R}^{|V|\times|R|\times|V|}$, где R — множество отношений. Существуют два вида мультиреляционных графов: гетерогенные и мультиплексные.

В гетерогенных графах узлы определены типами, что означает, что можно разбить множество узлов на непересекающиеся множества $V = V_1 \cup V_2 \cup \ldots \cup V_k$, где $V_i \cap V_j = \emptyset \ \forall i \neq j$. Ребра в таких разнородных графах обычно удовлетворяют ограничениям в соответствии с типами узлов, чаще всего это ограничения, в которых ребра соединяют только узлы определенных типов, т. е. $(u; \tau_i; v) \in E \rightarrow u \in V_j, v \in V_k$. Например, в гетерогенном биомедицинском графе может быть один тип узла, представляющий белки, один тип, представляющий лекарства, и один тип, представляющий болезни. Ребра, представляющие «лечение», будут встречаться

только между узлами лекарств и узлами болезни. Аналогично, ребра, представляющие «побочные эффекты», могут возникать только между двумя узлами лекарств. Многодольные графы — это хорошо известный частный случай гетерогенных графов, где ребра могут соединять только узлы разных типов, т. е. $(u; \tau_i; v) \in E \rightarrow u \in V_j, v \in V_k \wedge j \neq k$

В мультиплексных графах предполагается, что граф можно разложить на набор из k слоев. Предполагается, что каждый узел принадлежит каждому слою, и каждый слой соответствует уникальному отношению, представляющему тип ребра внутри слоя для этого слоя. Предполагается, что могут существовать типы ребер между слоями, которые соединяют один и тот же узел. Мультиплексные графы лучше всего понять на примерах. Например, в мультиплексной транспортной сети каждый узел может представлять город, а каждый уровень может представлять определенный вид транспорта (например, на самолете или на поезде). Тогда внутриуровневые ребра будут представлять города, связанные различными видами транспорта, а межуровневые ребра представляют возможность переключения видов транспорта в конкретном городе.

Наконец, во многих случаях может существовать атрибутивная или функциональная информация, связанная с графом (например, изображение профиля, связанное с пользователем в социальной сети). Чаще всего это атрибуты уровня узла, которые представляются с помощью матрицы $X \in \mathbb{R}^{|V| \times m}$ с действительными значениями, где предполагается, что порядок узлов согласуется с порядком в матрице смежности. Таким образом, в гетерогенных графах обычно предполагается, что каждый отдельный тип узла имеет свой собственный тип атрибутов. В редких случаях рассматриваются графы, которые имеют реберные признаки с действительными значениями в дополнение к дискретным типам ребер.

1.2. Машинное обучение на графах

Машинное обучение по своей сути является проблемно-ориентированная дисциплина, которая стремится создавать модели, которые могут учиться на данных для решения конкретных задач. Машинное обучение с графами ничем не отличается, однако обычные категории контролируемого (с учителем) и неконтролируемого обучения (без учителя) не обязательно являются наиболее информативными или полезными, когда речь идет о графах.

Приведем краткий обзор наиболее важных и хорошо изученных задач машинного обучения на графовых данных. Рассмотрим, задачи с контролируемым обучением, которые популярны для графических данных. Однако задачи машин-

ного обучения на графах часто стирают границы между традиционными категориями машинного обучения.

Классификация узлов

Предположим, есть большой набор данных социальной сети с миллионами пользователей, но мы знаем, что значительное число этих пользователей на самом деле являются ботами. Идентификация этих ботов может быть важна по многим причинам: компания может не захотеть рекламировать ботов, или боты могут фактически нарушать условия обслуживания социальной сети. Вручную проверять каждого пользователя, чтобы определить, является ли он ботом, было бы непомерно дорого, поэтому в идеале мы хотели бы иметь модель, которая могла бы классифицировать пользователей как ботов (или нет), учитывая лишь небольшое количество примеров, помеченных вручную [3].

Это классический пример классификации узлов, цель которого состоит в том, чтобы предсказать метку y_u, которая может быть типом, категорией или атрибутом, связанным со всеми узлами $u \in V$, когда нам даны истинные метки на вершине представляющие обучающий набор узлов $V_{train} \subset V$. Классификация узлов – пожалуй, самая популярная задача машинного обучения на графовых данных, особенно в последние годы.

На первый взгляд, классификация узлов кажется прямым вариантом стандартной контролируемой классификации, но на самом деле есть важные отличия. Самое важное отличие состоит в том, что узлы в графе не являются независимыми и одинаково распределёнными. Обычно, когда мы строим модели машинного обучения с учителем, мы предполагаем, что каждая точка данных статистически независима от всех других точек данных; в противном случае нам может понадобиться смоделировать зависимости между всеми нашими входными точками. Мы также предполагаем, что точки данных одинаково распределены; в противном случае у нас нет возможности гарантировать, что наша модель будет обобщать новые точки данных. Классификация узлов полностью нарушает это предположение о независимости и одинаковой распределенности. Вместо моделирования набора независимых и одинаковых распределенности точек данных, мы моделируем взаимосвязанный набор узлов. Например, люди склонны заводить дружеские отношения с другими людьми, которые разделяют те же интересы или демографические данные. Основываясь на понятии гомофилии (это тенденция индивидов ассоциироваться и образовывать связи с другими подобными, как в пословице «рыбак рыбака видит издалека».), мы можем создавать модели машинного обучения, которые пытаются назначать одинаковые метки соседним узлам на графе. Помимо гомофилии существуют также такие понятия, как структурная эк-

вивалентность, которая заключается в том, что узлы с похожей структурой локального соседства будут иметь одинаковые метки, а также гетерофилия (тенденция индивидов собираться в разнообразные группы; противоположность гомофилии.), предполагающая, что узлы будут преимущественно связаны с узлами. с разными этикетками. Когда мы строим модели классификации узлов, мы хотим использовать эти концепции и моделировать отношения между узлами, а не просто рассматривать узлы как независимые точки данных [4].

Под наблюдением или под наблюдением? Из-за нетипичного характера классификации узлов исследователи часто называют ее полууправляемой (полуконтролируемое обучение). Эта терминология используется потому, что при обучении моделей классификации узлов мы обычно имеем доступ к полному графу, включая все непомеченные (например, тестовые) узлы. Единственное, чего нам не хватает, так это меток тестовых узлов. Однако мы все еще можем использовать информацию о тестовых узлах (например, знание их соседства на графе), чтобы улучшить нашу модель во время обучения. Это отличается от обычной контролируемой обстановки, в которой немаркированные точки данных совершенно не наблюдаются во время обучения.

Прогноз отношений

Классификация узлов полезна для получения информации на основе его связи с другими узлами в графе. Но иногда может не хватать информации об отношениях (связях). Встает вопрос можно использовать машинное обучение для определения ребер между узлами в графе? У этой задачи много названий, таких как предсказание связей, завершение графа и реляционный вывод, в зависимости от приложения к конкретной предметной области.

Для понимания будем просто называть это прогноз отношений. Наряду с классификацией узлов, это одна из самых популярных задач машинного обучения с графическими данными, имеющая бесчисленное количество реальных применений: рекомендации контента пользователям на социальных платформах, прогнозирование побочных эффектов лекарств или вывод новых фактов в реляционных базах данных. Все эти задачи можно рассматривать как частные случаи прогноза отношений.

Стандартная установка для прогноза отношений состоит в том, что нам дан набор узлов V и неполный набор ребер между этими узлами $E_{train} \subset E$. Сложность этой задачи сильно зависит от типа исследуемых графических данных. Например, в простых графах, таких как социальные сети, которые кодируют только отношения «дружбы», есть простые эвристики, основанные на том, сколько соседей имеют два узла, которые могут обеспечить высокую производительность

решения какой-либо задачи. С другой стороны, в более сложных наборах данных мультиреляционных графов, таких как графы биомедицинских знаний, которые кодируют сотни различных биологических взаимодействий, предсказание отношений может потребовать сложных стратегий рассуждений и выводов. Как и классификация узлов, предсказание отношений стирает границы традиционных категорий машинного обучения и требует индуктивных смещений (индуктивное смещение (также известен как предвзятость обучения) алгоритма обучения - это набор предположений, которые учащийся использует для прогнозирования выходных данных заданных входных данных, с которыми он не сталкивался) [3].

Кластеризация

И классификация узлов, и прогноз отношений требуют вывода недостающей информации, и во многих отношениях эти две задачи являются графовыми аналогами обучения с учителем. Обнаружение сообщества, с другой стороны, является графовым аналогом неконтролируемой кластеризации.

Предположим, у нас есть доступ ко всей информации о цитировании в Google академии, и мы создаем график сотрудничества, который связывает двух исследователей, если они совместно написали статью.

Если бы пришлось исследовать эту сеть, то нельзя утверждать, что каждый с равной вероятностью будет сотрудничать со всеми остальными. Более вероятно, что граф разделится на разные кластеры узлов, сгруппированных вместе по исследовательской области, по учреждениям или демографическим факторам. Другими словами, можно ожидать, что эта сеть, как и многие реальные сети, будет демонстрировать структуру сообщества, где узлы с гораздо большей вероятностью образуют ребра с узлами, принадлежащими к тому же сообществу.

Это общее представление, лежащее в основе задачи обнаружения сообщества. Задача обнаружения сообщества состоит в том, чтобы сделать вывод о скрытых структурах сообщества, учитывая только входной граф $G = (V; E)$. Многие реальные приложения обнаружения сообществ включают обнаружение функциональных модулей в сетях генетического взаимодействия или, например, обнаружение мошеннических групп пользователей в сетях финансовых транзакций .

Классификация графов, регрессия и кластеризация

Последний класс популярных приложений машинного обучения включает задачи классификации, регрессии или кластеризации по всем графам. Например, имея граф, представляющий структуру молекулы, можно захотеть построить регрессионную модель, которая могла бы предсказать токсичность или растворимость молекулы. Или можно захотеть построить классификационную модель, чтобы определить, является ли компьютерная программа вредоносной, путем ана-

лиза графического представления ее синтаксиса и потока данных.

В этих приложениях для классификации графов или регрессии стремяться учиться на данных графа, вместо того, чтобы делать прогнозы по отдельным компонентам одного графа (т. е. узлам или ребрам). Вместо этого нам дается набор данных из нескольких разных графов, и наша цель — сделать независимые прогнозы для каждого графа [4].

Из всех задач машинного обучения на графах регрессия графов и классификация являются, пожалуй, самыми прямыми аналогами стандартного обучения с учителем. Каждый граф является независимый и одинаково распределённый точкой данных, связанной с меткой, и цель состоит в том, чтобы использовать помеченный набор обучающих точек для изучения сопоставления графов с метками. Подобным образом кластеризация графа является прямым расширением неконтролируемой кластеризации для данных графа. Однако проблема в этих задачах на уровне графа состоит в том, как определить полезные функции, учитывающие реляционную структуру в каждой точке данных.