تکلیف نهایی شبکه‌های پیچیده

پارسا اسدنژاد

فهرست

[فهرست اشکال 2](#_Toc203468052)

[شبکه‌های BA 3](#_Toc203468053)

[شبکه‌های WS 3](#_Toc203468054)

[شبکه‌های ER 4](#_Toc203468055)

[استخراج ویژگی 4](#_Toc203468056)

[مقایسه مدل‌ها 5](#_Toc203468057)

[ویژگی‌های بیشتر 5](#_Toc203468058)

[مدل SIR 8](#_Toc203468059)

[پارامترها 8](#_Toc203468060)

[تحلیل نمودار 9](#_Toc203468061)

[ایجاد مجموعه داده 9](#_Toc203468062)

[پیاده‌سازی مدل یادگیری ماشین 10](#_Toc203468063)

[پیش پردازش داده‌ها 10](#_Toc203468064)

[مدل Linear Regression 11](#_Toc203468065)

[مدل Random Forest Classifier 11](#_Toc203468066)

[مدل XGBoost 11](#_Toc203468067)

# فهرست اشکال

[شکل 1 مدل SIR در شبکه‌های متفاوت 8](#_Toc203468121)

[شکل 2 اطلاعات دیتاست 10](#_Toc203468122)

[شکل 3 خروجی مدل یادگیری ماشین Linear Regression 11](#_Toc203468123)

[شکل 4 خروجی مدل یادگیری ماشین Random Forest 11](#_Toc203468124)

[شکل 5 خروجی مدل یادگیری ماشین XGBoost 12](#_Toc203468125)

شبکه‌ها Erdos-Renyi, Barabasi-Albert و Watts-Strogatz سه نوع مدل شبکه هستند که برای تولید شبکه‌های پیچیده با ویژگی‌های خاص استفاده می‌شوند.

# شبکه‌های BA

این مدل بر اساس دو اصل رشد (Growth) و اتصال ترجیحی (Preferential Attachment) عمل می‌کند. شبکه با تعداد کمی گره شروع می‌شود و گره‌های جدید به تدریج اضافه می‌شوند.. هر گره جدید به گره‌های موجود با احتمال متناسب با درجه (تعداد اتصالات) آن‌ها متصل می‌شود. این مدل ویژگی‌های زیر را داراست:

* توزیع درجه گره‌ها از قانون توان (Power-Law) پیروی می‌کند، یعنی تعداد کمی گره با درجه بالا (هاب‌ها) و تعداد زیادی گره با درجه پایین وجود دارد.
* جهان کوچک یا (Small World) است یعنی فاصله متوسط بین گره‌ها نسبتا کم است.
* در مدل سازی شبکه‌های واقعی مانند اینترنت، شبکه‌های اجتماعی و شبکه‌های بیولوژیکی که دارای هاب‌های بزرگ هستند کاربرد دارد.

این شبکه دارای پارامترهای زیر است:

* تعداد گره‌ها
* تعداد یال‌هایی که هر گره جدید به گره‌های موجود متصل می‌کند.

# شبکه‌های WS

این مدل برای تولید شبکه‌های جهان کوچک طراحی شده است. ابتده یک شبکه منظم مانند حلقه با n گره ایجاد می‌شود که هر گره به k همسایه نزدیک خود متصل است. سپس، برخی یال‌ها با احتمال p به صورت تصادفی بازآرایی(Rewiring) می‌شوند. این شبکه دارای ویژگی‌های زیر است.

* از لحاظ توزیع درجه یکنواخت است، اما با افزایش p، تنوع در درجه گره‌ها افزایش می‌یابد.
* فاصله متوسط بین گره‌ها کم است و ضریب خوشه‌بندی (Clustering Coefficient) بالایی دارد.
* در مدلسازی شبکه‌هایی که ترکیبی از نظم و تصادف هستند، مانند شبکه‌های اجتماعی یا عصبی کاربرد دارند.

که پارامترهای زیر را دارد:

* تعداد گره‌ها
* تعداد همسایگان اولیه هر گره k
* احتمال بازآرایی یال‌ها p

# شبکه‌های ER

این مدل یک شبکه تصادفی است که در آن n گره وجود دارد و هر جفت گره با احتمال pبه هم متصل می‌شوند (یا به‌صورت معادل، m یال به‌صورت تصادفی انتخاب می‌شود.( این شبکه دارای ویژگی‌های زیر است.

* از لحاظ توزیع درجه از توزیع پواسون (Poisson) پیروی می‌کنند.
* فاصله متوسط بین گره‌ها کم است اما ضریب خوشه‌بندی (Clustering Coefficient)معمولا کم است.
* در مدلسازی شبکه‌های کاملا تصادفی یا به عنوان مبنای مقایسه‌ای با دیگر مدل‌ها استفاده می‌شود.

که دارای پارامترهای زیر است.

* تعداد گره‌ها
* احتمال اتصال بین هر جفت گره p یا m (تعداد یال‌ها)

# استخراج ویژگی

برای استخراج ویژگی‌های این شبکه، معیارهای زیر معمولا بررسی می‌شود.

|  |  |  |
| --- | --- | --- |
| نام | معادل فارسی | توضیح |
| Degree Distribution | توزیع درجه | نشان‌دهنده توزیع اتصالات گره‌ها |
| Clustering Coefficient | ضریب خوشه بندی | میزان تشکیل خوشه‌های محلی(مثلث‌ها) |
| Average Path Length | متوسط طول مسیر | میانگین کوتاه‌ترین مسیر بین تمام جفت گره‌ها |
| Diameter | قطر شبکه | حداکثر فاصله بین هر دو گره |
| Density | چگالی | نسبت یال‌های موجود به یال‌های ممکن |

# مقایسه مدل‌ها

در جدول زیر به مقایسه مدل‌های ایجاد شبکه‌های پیچیده می‌پردازیم.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| مدل | توزیع درجه | ضریب خوشه‌بندی | طول مسیر متوسط | کاربرد اصلی |
| Barabási-Albert | قانون توان | پایین | کم | شبکه‌های بدون مقیاس (Scale-Free) |
| Watts-Strogatz | تقریباً یکنواخت | بالا | کم | شبکه‌های جهان کوچک |
| Erdős-Rényi | پواسون | پایین | کم | شبکه‌های تصادفی |

# ویژگی‌های بیشتر

علاوه بر معیارهای ذکر شده (توزیع درجه، ضریب خوشه بندی، طول مسیر متوسط و تعداد یال‌ها) استفاده کردن از ویژگی‌های زیر نیز می‌تواند مفید باشد.

1. مرکزیت (Centrality Measures)

* Degree Centrality
* Betweenness Centrality
* Closeness Centrality
* Eigenvector Centrality

1. اجزای همبند(Connected Components)

* تعداد و اندازه اجزای همبند در شبکه، به ویژه در شبکه ER که ممکن است به دلیل تصادفی بودن، چندین جزء مجزا داشته باشد.

1. ماژولاریتی(Modularity)

* معیاری برای ارزیابی کیفیت تقسیم بندی شبکه به جوامع (Communities). این معیار نشان می‌دهد که آیا گره‌ها در شبکه به صورت خوشه‌های متراکم گروه بندی شده‌اند یا خیر.

1. استحکام شبکه (Robustness)

* بررسی مقاومت شبکه در برابر حذف گره‌ها یا یال‌ها. این ویژگی به ویژه در شبکه‌های BA به دلیل وجود هاب حساس است.

1. ضریب هم‌گروهی (Assortativity)

* معیاری که نشان می‌دهد گره‌ها با درجه مشابه نمایل دارند به هم متصل شوند یا به گره‌های با درجه متفاوت.

1. توزیع طول مسیر (Path Length Distribution)

* توزیع فواصل کوتاه‌ترین مسیرها بین تمام جفت گره‌ها در شبکه.

1. شعای شبکه (Radius)

* کمترین فاصله‌ای که یک گره می‌تواند به تمامی گره‌های دیگر داشته باشد.

|  |  |
| --- | --- |
| مدل | ویژگی برجسته |
| Erdős-Rényi | گسترش سریع اطلاعات، فشردگی بالا، خوشه‌بندی کم |
| Watts-Strogatz | خوشه‌بندی بالا، مسیرهای بلند، ساختار شبیه به شبکه‌های اجتماعی واقعی |
| Barabási-Albert | وجود نقاط مرکزی (hubs)، توزیع قدرت‌قانونی، ماژولاریتی بالا |

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Erdős-Rényi | Watts-Strogatz | Barabási-Albert |
| Number of Nodes | 100 | 100 | 100 |
| Number of Edges | 493 | 200 | 196 |
| Density | 0.0996 | 0.0404 | 0.0396 |
| Average Clustering | 0.0941 | 0.4167 | 0.1578 |
| Average Path Length | 2.2432 | 5.5251 | 3.1905 |
| Network Diameter | 5 | 12 | 6 |
| Network Radius | 3 | 9 | 4 |
| Degree Centrality | 0.1818 | 0.0606 | 0.2222 |
| Betweenness Centrality | 0.0381 | 0.2165 | 0.2790 |
| Closeness Centrality | 0.5103 | 0.2409 | 0.4783 |
| Eigenvector Centrality | 0.1741 | 0.2134 | 0.4562 |
| Number of Connected Parts | 1 | 1 | 1 |
| Length of Biggest Part | 100 | 100 | 100 |
| Assortativity | -0.0887 | -0.0067 | -0.1116 |
| Modularity | 0.2591 | 0.7283 | 0.4782 |

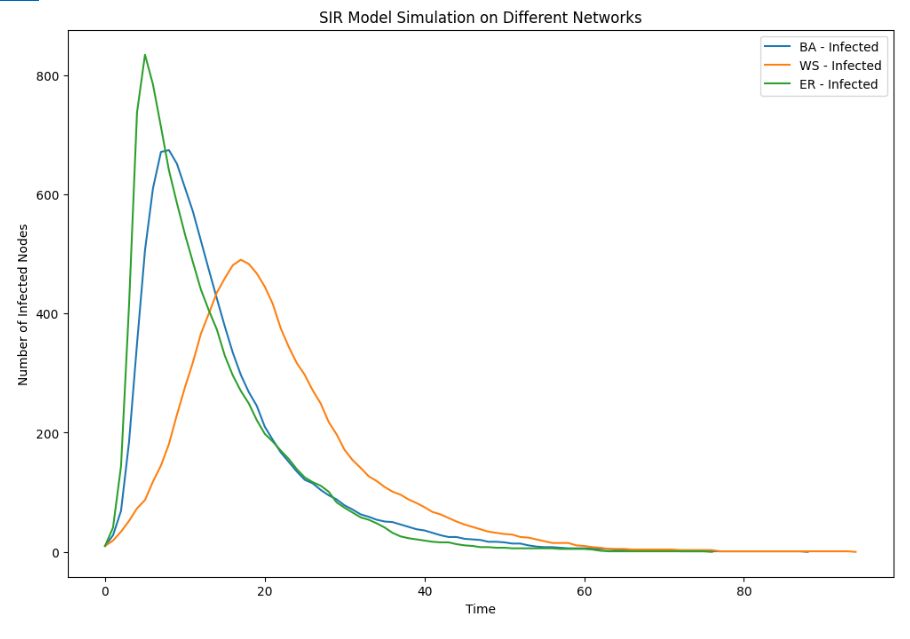
# مدل SIR

یکی از ساده‌ترین و پرکاربردترین مدل‌های اپیدمیولوژی ریاضی است که برای شبیه‌سازی گسترش بیماری‌های عفونی در میان جمعیت استفاده می‌شود. نام SIR از سه وضعیت یا کلاس مختلف که افراد جامعه می‌توانند در آن قرار گیرند، گرفته شده است.

* حساس یا مستعد (Susceptible)
* آلوده یا بیمار (Infected)
* بهبود یافته (Recovered)

## پارامترها

* بتا (Beta) که نرخ انتقال بیماری ( احتمال آلوده شدن در تماس با یک فرد آلوده) را نشان می‌دهد.
* گاما (Gamma) که نرخ بهبودی (احتمال اینکه یک فرد آلوده در یک واحد زمان بهبود یابد.) را نشان می‌دهد.
* نسبت Beta بر Gamma که اگر بزرگ‌تر از 1 باشد، بیماری گسترش می‌یابد.



شکل 1 مدل SIR در شبکه‌های متفاوت

## تحلیل نمودار

به طور کلی هر سه شبکه با تعداد کمی گره مبتلا شروع می‌کنند و سپس تعداد گره‌های مبتلا افزایش می‌یابد و به اوج می‌رسد.

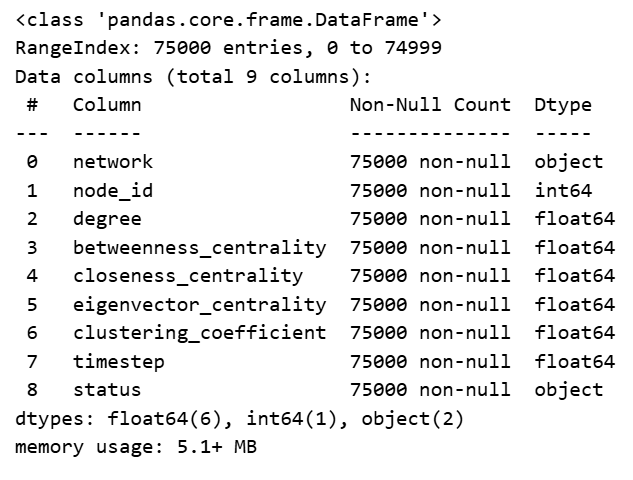
همانطور که در شکل دیده می‌شود نمودار مرتبط با مدل ER نسبت به سایر مدل‌ها با شدت بالاتری رشد می‌کند که این عمل می‌توانند به دلیل داشتن تعداد یال بیشتر و در نتیجه چگال تر بودن نسبت به سایر مدل‌ها باشد. علاوه بر این، همانطور که در جدول ویژگی‌ها می‌توان مشاهده نمود، طول فاصله بین نودها در این مدل از بقیه کمتر است، لذا نموارد مطابق منطق پیش رفته است.

با بررسی نمودار مرتبط با مدل BA می‌توان مشاهده نمود که این مدل هم نسبتا رشد بالایی داشته است و در مقایسه با مدل WS شدت بسیار بالاتر دارد که این رخداد می‌تواند به دلیل پیروی از قانون Power-Low باشد چرا که در این مدل به علت وجود چند Hub نرخ انتقال می‌تواند بالا برود.

در انتها نیز مدل WS به دلیل ساختار منظم، فاصله طولانی تر بین نودها و چگال نبودن نسبت به سایر مدل‌ها رشد کمتر را تجربه کرده است که با توجه به اطلاعات جدول منطقی به نظر می‌رسد.

# ایجاد مجموعه داده

برای تولید یک مجموعه داده در ابتدا بایستی تعدادی ویژگی یا feature مناسب جمع آوری کنیم. به طور کلی این ویژگی‌ها می‌توانند از دو نوع متفاوت باشند. یک دسته از ویژگی‌ها شامل ویژگی‌های کلی گره است، همانند: degree, betweenness centrality, closeness centrality, eigenvector centrality و clustering coefficient. دسته دیگر از ویژگی‌ها به کمک مدل SIR به دست می‌آید و شامل وضعیت هر گره در یک واحد زمان است. این وضعیت یا status می‌تواند یک از حالت‌های S, I و R که به ترتیب نشان دهنده حالت بیمار، مستعد بیمار شدن و بهبود یافته می‌شند. علاوه بر موارد ذکر شده تعدادی ستون برای نشان دادن id گره و اینکه گره به چه مدلی (ER, WS, BA) متعلق است اختصاص پیدا کرده است.



شکل 2 اطلاعات دیتاست

# پیاده‌سازی مدل یادگیری ماشین

من از سه مدل یادگیری ماشین اعم از Linear Regression, Random Forrest و XGBoost استفاده کردم، که در ادامه به بررسی نتایج بدست آمده از هر کدام می‌پردازم.

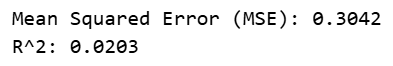
## پیش پردازش داده‌ها

همانطور که در شکل 2 مشخص است، دو نوع داده عددی و غیر عددی داریم، در ابتدا بایستی داده‌های غیر عددی خود را به داده‌های عددی تبدیل کنیم. لذا برای این منظور من از کتابخانه scikit-learn و کلاس Label Encoder و تابع fit transform استفاده کردم. در این مجموعه داده‌ها تنها دو ستون غیر عددی داریم: status و network، که پس از تبدیل به اعداد 0 تا 2 مقدارشان تغییر خواهد نمود.

داده‌های عددی نیز به کمک کتابخانه scikit-learn، کلاس Standard Scaler و تابع fit transform به مقادیر مناسب‌تری تغییر نمودند.

## مدل Linear Regression

مدل **رگرسیون خطی** یکی از ساده‌ترین و پرکاربردترین الگوریتم‌های یادگیری ماشین و آمار است که برای پیش‌بینی مقدار یک متغیر (متغیر هدف) بر اساس یک یا چند ویژگی ورودی استفاده می‌شود. این مدل تلاش می‌کند خطی را بیابد که به بهترین شکل داده‌ها را مدل‌سازی کند، به‌طوری‌که مجموع مربع خطاها (فاصله بین داده واقعی و پیش‌بینی‌شده) کمینه باشد. خروجی مدل معمولاً ترکیبی خطی از ویژگی‌هاست که هر کدام دارای ضریب خاصی هستند که نشان‌دهنده اهمیت آن ویژگی در پیش‌بینی است. با وجود سادگی، رگرسیون خطی پایه‌ بسیاری از مدل‌های پیچیده‌تر محسوب می‌شود.



شکل 3 خروجی مدل یادگیری ماشین Linear Regression

## مدل Random Forest Classifier

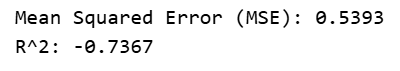
مدل Random Forest یکی از الگوریتم‌های قدرتمند یادگیری ماشین بر پایه درخت تصمیم است که برای مسائل طبقه‌بندی و رگرسیون به‌کار می‌رود. این مدل مجموعه‌ای از درخت‌های تصمیم مستقل می‌سازد (اصطلاحاً یک "جنگل") و خروجی نهایی را با میانگین گرفتن (در رگرسیون) یا رأی‌گیری اکثریت (در طبقه‌بندی) تعیین می‌کند. هر درخت از بخشی تصادفی از داده‌ها و ویژگی‌ها استفاده می‌کند که باعث کاهش بیش‌برازش (overfitting) و افزایش دقت مدل می‌شود. قدرت Random Forest در توانایی آن برای تحلیل روابط پیچیده و غیرخطی بین ویژگی‌هاست، بدون نیاز به تنظیمات دقیق اولیه.



شکل 4 خروجی مدل یادگیری ماشین Random Forest

## مدل XGBoost

مدل XGBoost مخفف (eXtreme Gradient Boosting) یکی از قدرتمندترین الگوریتم‌های یادگیری ماشین در حوزه یادگیری تقویتی (Boosting) است که برای مسائل طبقه‌بندی، رگرسیون، رتبه‌بندی و حتی بقا (survival analysis) استفاده می‌شود. این مدل مجموعه‌ای از درخت‌های تصمیم را به‌صورت ترتیبی آموزش می‌دهد، به‌طوری‌که هر درخت جدید تلاش می‌کند خطاهای درخت‌های قبلی را اصلاح کند. XGBoost با استفاده از تکنیک‌های پیشرفته مثل regularization، pruning، و parallel processing، دقت بالا و سرعت آموزش فوق‌العاده‌ای دارد. همچنین به‌خاطر توانایی‌اش در مقابله با داده‌های ناقص و بزرگ، در بسیاری از رقابت‌های داده‌کاوی و پروژه‌های صنعتی به عنوان الگوریتم برتر شناخته شده است.



شکل 5 خروجی مدل یادگیری ماشین XGBoost