دانشگاه صنعتی سیرجان
گروه مهندسی کامپیوتر

پایان نامه کارشناسی
گرایش نرم افزار و هوش مصنوعی

دانشمند داده - معرفی

نگارش:
محمد گنجی نژاد

استاد راهنما:
امیر سالارپور

مرداد ۱۳۹۹


بسم الله الرحمن الرحيم

تقدیم به:

دوست خوبم امین شریفی

تشکر و قدردانی:

سپاس و تشکر فراوان از دکتر امیر سالار پور که بنده را در مسیری قرار دادند که مطـالبی ارزشـمند را بیاموزم و به درستی این همان چیزی است که شهید مطهری گفته اند :

" در زندگی چیزی برای ترسیدن وجود ندارد، بلکه باید درک شود، اگر نتوانید موضوعی را بـه سـادگی توضیح دهید، در حقیقت آن موضوع را درست نفهمیده اید. "

چکیده

علوم داده یک حوزه ی میان رشته ای که افراد با استفاده از روش ها، فرایندها، الگوریتم ها و سیستم ها از داده های آماده به نتایج و پیش بینی های جدید می رسند، در واقع این حوزه، دید افراد را نسب به اتفاقات آینده روشن تر می کند.

نمونه بارز این حوزه را می توان در پیش بینی های نیت سیلور (ملقب به پیامبر عصر جدید) بدانیم:

این شخص در انتخابات ریاست جمهوری آمریکا در سال ۲۰۰۸ نتیجه ی ۴۹ ایالت را از ۵۰ ایالت و در انتخابات سال ۲۰۱۲ کل ایالت ها را به درستی پیش بینی نمود.

همچنین در انتخابات کنگره ی آمریکا در سال ۲۰۰۸، تمام ۳۵ سناتور را به درستی پیش بینی نمود.

این شخص همچنین خالق دقیق ترین سامانه ی پیش بینی نتایج بازی بیسبال هست.

البته از این پیش بینی ها می توان به راحتی پول زیادی را بدست آورد، به همین دلیل این حوزه یکی از شاخص ترین حوزه ها در دنیای امروز می باشد.

در روند نگارش، ابتدا به بینش اولیه علوم داده و سپس درباره ی ابزارهای تحلیل می پردازیم.

واژه‌های کلیدی: علوم داده، تحلیل داده، پیش بینی، پردازش زبان طبیعی

فهرست مطالب

فهرست اشکال

فصل ۱: مقدمه

۱-۱- مقدمه

سالیان زیادی از اختراع کامپیوتر و پدیدن آمدن تکنولوژی های مرتبط با آن نگذشته است، اما در همین مدت انقلاب های زیادی رخ داده است.

هر نرم افزاری و وبسایتی که شما می بینید از مجموعه ای کد تشکیل شده است، که مثل آجر های و مصالحی هستند که یک خانه را تشکیل می دهد. داده ها در دنیای ما نقش بی نهایت مهمی دارند. از یک سیستم مدرسه شروع کنیم تا سازمان های بسیار بزرگ، که به نوعی همه و همه نیاز بسیار زیادی به داده ها دارند.

یک مدرسه را تصور کنید که لیست دانش آموزان را ندارد، یا لیست ساعات حضور معلم ها را آیا می شود بدون نگهداری داده های هر دانش آموز و تمام داده ها مورد نیاز دیگر یک مدرسه را اداره کرد؟

مسلما نه، ما در همه جا به مجموعه ای از داده ها و نگهداری آنها نیاز داریم. باید گاهی آنها را تحلیل کنیم گاهی آنها را ساده تر کنیم، گاهی آنها را حذف کنیم، گاهی آنها را بروز کنیم و غیره علم داده را می توان علم نوظهوری دانست که می تواند بسیاری از شرکت ها را نجات دهد و مشاغل زیادی را ایجاد کند و حتی دیگر علوم را بهتر کند.

در این رساله می خواهیم درباره ی متخصص داده۱ با پایتون بحث کنیم ابتدا لازم است که مواردی چون علم داده چیست؟ زبان های برنامه نویسی، ابزارها، کتابخانه هایی که هر متحصص داده باید تجربه ی کار با آنها را داشته باشد تا بتواند به اکتشاف داده ها بپردازد و به سوالاتی که درباره ی داده ها پرسیده می شود، به پاسخ درستی دست یافت.

۱-۲- علم داده چیست

اگر به ساده ترین شکل ممکن بگوییم که علم داده[۱] چیست؟ باید گفت که کسب آگاهی و دانش از مجموعه ای از داده ها یعنی شما با استفاده از داده هایی که دارید، چیزی را بفهمید یا چیزی را

[۱] Data science

کشف کنید. به زبان ساده تر، یعنی کسب اطلاعات به منظور استفاده ی از آن می باشد (این تعریف از علم داده تعریف بسیار ساده و قابل فهم برای هر کسی است، اما تعریف کاملی نیست) .

برای رسیدن به درک بهتر مثال ها و دیگر تعاریف از علم داده، بایستی با داده ها درگیر شوید و یک تجربه در این زمینه کسب نماید. در واقع علم داده دانشی میان رشته ای است. برای کسب دانش و آگاهی از داده ها و اطلاعات که از روش ها، الگوریتم ها سیستم های علمی و فرآیند ها برای کسب آگاهی و بینش از داده های ساختار یافته [۱]و ساختار نیافته[۲] استفاده می کند.

۱-۳- داده های ساختار یافته و بدون ساختار

در مورد داده های ساختار یافته و ساختار نیافته باید گفت که داده های ساختار یافته به داده هایی گفته می شود که، برای کامپیوتر قابل فهم هستند. یعنی کامپیوتر می تواند با این داده ها سریع کار پردازش را انجام دهد. مثل داده های موجود در پایگاه های داده یا مثلا داده های نرم افزار اکسل اما داده های ساختار نیافته مثل داده های هستند که در بانک های اطلاعاتی قرار ندارند.

یک مثال واضح تر داده های موجود در ویدیو ها، اخبار، آهنگ ها و غیره است. این داده ها برای اینکه بتوانیم آنها را در علم داده، داده کاوی و یادگیری ماشین استفاده کنیم، باید ساختار یافته شوند. به عبارتی دیگر هر نوع داده ای که کامپیوتر سریعا بتواند آن را با استفاده از الگوریتم های از پیش تعیین شده پردازش کند ساختار یافته است.

[۱] Structured data

[۲] Unstructured data

۱-۴- نیاز به علم داده

اگر هنوز هم متوجه ی مفهوم علم داده نشده اید، با یک مثال دیگر به درک آن نزدیک تر خواهیم شد. تصور کنید که علم آمار، ریاضی، برنامه نویسی و تجزیه و تحلیل داده را می دانید به این فکر کنید که این علوم تا چه اندازه می توانند کمک نمایند که با استفاده از داده هایی که به دست آورده اید(مثل اطلاعاتی که از شخصیت کاربران جمع آوری می شود) سبب موفقیت و رشد یک کسب و کار شوید.

به عنوان مثال شما می توانید بهترین محصولی که کاربران واقعا به آن نیاز دارند را عرضه کنید فقط با کمک علم داده.

۱-۵- هدف علم داده

شاید یکی از مهمترین مزایای علم داده برای شرکت ها را بتوان قدرت تصمیم گیری دانست. علم داده بی نهایت در تصمیم گیری های اصلی شرکت ها، مهم است. با وجود اطلاعات از داده های کاربران می توان بهترین تصمیم ها را گرفت و کسب و کار مورد نظر را تقویت کرد، تنها با پاسخ دادن به این پرسش ها:

اینکه کاربران چه می خواهند؟ در جستجوی چه هستند؟ چه را دوست دارند؟ به چه چیزهایی عادت دارند؟ میانگین سن اکثر کاربران یک محصول چند است؟ یا حتی مواردی مثل چه رنگ هایی در جذب کاربر بیشتر نقش دارند؟ نیز می تواند تصمیم گیری را برای شرکت ها ارائه دهنده خدمات بسیار آسان کند.

شرکت نتفلیکس[1] از داده هایی استفاده می کند تا سلیقه ی کاربران را در فیلم و سریال بداند و علاوه بر استفاده های متنوع از این اطلاعات، حتی این اطلاعات نقش مهمی در شروع پروژه ساخت یک فیلم جدید دارد. آنها همیشه دوست دارند بدانند کاربران بیشتر چه فیلمی را دوست دارند ساخته شود. علم داده را می توان مورد نیاز همه ی شرکت های بزرگ و کوچک برای ارتقا کسب و کار دانست. اما چون این کار در اکثر اوقات مخصوصا وقتی با داده های ساختار نیافته طرف هستیم، هزینه و زمان

[1] Netflix یک شرکت آمریکایی در سرویس تولید فیلم و سریال فعالیت می کند

زیادی می خواهد. فقط شرکت های کمی روی علم داده سرمایه گذاری می کنند. (این را هم در نظر بگیرید که استخدام متخصصین علم داده هزینه ی زیادی می خواهد و این افراد حقوق بسیار بالای را می خواهند چون برای متخصص شدن در علم داده باید چند علم را به خوبی یاد گرفت، سالهای زیادی را صرف یادگیری کرد و قطعا برای رسیدن به این شغل هوش بالایی نیاز است).

۱-۶- متخصص علم داده کیست و چه وظایفی دارد

به دلیل ماهیت این رشته تعریف واحدی برای متخصص علم داده یا همان دانشمند داده وجود ندارد. تعاریف و توصیفات، محدوده ی وسیعی را شامل می شود.

"متخصص علم داده یک فرد یا گروهی که با مهارت و ذهنیت علمی قادر به استفاده و ترکیب داده های قبلی و فعلی می باشد تا پرسش های درست را بپرسد (و در نهایت به آنها پاسخ دهد) تا اینکه بتواند آگاهانه ترین تصمیمات آتی را بگیرد." – دیدگاه نویسنده

نقل قول های مشهور زیادی از افراد ارشد این رشته برای تعریف متخصص داده وجود دارد. هرکدام از آنها یک جنبه را از یک منظر به خوبی توصیف می کند و همه ی آنها با هم می توانند توصیف خوبی از واقعیت مربوط به متخصص علم داده و دانشمند داده به ما بدهد. در ادامه چند نقل قول از متخصص های شایسته را که دارای اهمیت بالایی است را بیان می کنیم.

داده نفت جدید است؟ خیر. داده خاک جدید است. ـــدیوید مک کندلوس

متخصصین داده، درگیر جمع آوری دیتا و ماساژ دادن آن به یک فرم قابل انعطاف و قابل کنترل هستند، داده را وادار می کنند که داستانش را بگوید، و آن داستان را به دیگران ارائه می دهند. ـــ مایک لوکیدیس، نایب رئیس اورایلی[۱]

متخصصین داده، شخصی است که در آمار از هر آمارگری بهتر و از هر مهندس نرم افزاری و در مهندسی نرم افزار بهتر است. ـــ جاش ویلز

[۱] O'Reilly یک سرویس آموزش کسب و کار به روش انتشار کتاب

تا سال ۲۰۱۸ ایالات متحده آمریکا کمبود ۱۹۰ هزار متخصص علم داده با مهارت و همچنین یک و نیم میلیون مدیر و تحلیل گر با توانایی به دست آوردن بینش قابل اقدام از سیل کلان داده ها [۱]را تجربه خوهد کرد. — مک کینزی ریپورت

این رشته ی جدید مورد تقاضا (متخصص علم داده) وعده تغییرات اساسی و انقلابی در صنایع از تجارت تا دولت، و از نظام درمانی تا محیط دانشگاهی را می دهد. — نیویورک تایمز

متخصص علم داده باید در علوم کاربردی اطلاعات داشته باشد با تجربه ای گسترده در صنعت و آموزش. — خوآن اف سیا

۱-۷- داستان یک چالش واقعی

این داستان ترجمه ای از صحبت های دکتر مرتضی حیدر است که در مدرسه مدیریتی تدراجرز به عنوان دانشیار فعالیت می کند.

در شهر تورنتو[۲]، ترانزیت[۳] عمومی توسط کمیسیون ترانزیت تورنتو انجام می شود. ما آنها TTC [۴] می نامیم. این یکی از بزرگترین مقامات ترانزیت منطقه ، در آمریکای شمالی است. آن سازمان یک روز با من تماس گرفتند و گفتند: "ما مشکل داریم." و من گفتم ، "خوب ، مشکلی چیست؟" آنها گفتند: "خب ، ما شکایات زیادی در این منطقه داریم و می خواهیم آن را تجزیه و تحلیل کنیم ولی به کمک شما احتیاج داریم."

گفتم: "خب خوش حال می شوم به شما کمک کنم، چند شکایت دارید؟"

گفتند: "خیلی زیاد"

---

[۱] Big data

[۲] Toronto یکی از شهرهای کانادا

[۳] Transit بخشی از جابه جایی مسافر و کالا بین مبدا و مقصد که مستلزم عبور از کشور ثالث می باشد.

[۴] مخفف عبارت Toronto Transit Commission کمیسیون ترانزیت تورنتو

گفتم: "چندتا؟"

گفتند: "شاید نیم میلیون شکایت در همین یک سال !"

گفتم: "خب ، بیایید کار رو شروع کنیم"

بنابراین من داده ها را گرفتم و شروع به تجزیه و تحلیل کردم.

در ابتدا نگاه من به ساختار داده هایی بود که به عنوان شکایت ثبت شده بود و اساساً، آنها کار بزرگی را برای نگه داشتن برخی از داده ها در قالب جدولی که داده های بدون ساختار بودند انجام داده اند.
در پرانتز (داده های ساختار یافته به داده هایی گفته می شود که دارای فرمت خاصی هستند مثل تاریخ، ولی داده های ساختار نیافته از قالب فرمتی خاصی پیروی نمی کنند، همچون داده های متنی مثل نظرات کاربران).
در این حالت، جدول شامل داده های زمان شکایت، چه کسی آن را دریافت کرده ، نوع شکایت چیست، نتیجه ی شکایت، مقصر کیست؟ و در بخش بدون ساختار آن متن پاسخ ایمیل و پاسخ نمابر بود. بنابراین، تصور کنید که چطور نیم میلیون ایمیل رو بررسی نموده و پاسخی برای سوالاتی که در ذهن پدید می آید برسیم.

Id	زمان دریافت	شکایت کننده	نوع شکایت	برطرف شده	...	پاسخ ایمیل	پاسخ فکس
1	7:17 PM 2/24/2019	جان امیت	بسته بودن مسیر	خیر	...	متن گفتگوی ایمیل	متن گفتگوی فکس
2	6:15 PM 2/24/2019	کارل دین	تغییر مسیر	بله	...	متن گفتگوی ایمیل	متن گفتگوی فکس
...	...	...	...	...	...	...	...
58500	8:20 AM 12/25/2019	سارا مالیون	بسته بودن مسیر	خیر	...	متن گفتگوی ایمیل	متن گفتگوی فکس

شکل (۱-۱) دیدکلی از داده های شکایت

بنابراین من شروع به کار با آن کردم. اولین چیزی که می خواستم بدونم این بود که چرا مردم شکایت می کنند و آیا الگویی وجود داره یا اینکه آیا برخی از روزها شکایات بیشتری نسبت به سایرین وجود داره؟

من به داده ها نگاه کردم و آن را در همه قالب های مختلف مورد تجزیه و تحلیل قرار دادم.
ولی در نهایت متوجه انگیزه ای که تعداد شکایت در یک روز خاص رو بالاتر از بقیه ی روز ها یا حتی

ماه های دیگه میبره، نشدم!

پاسخ سوالاتی که از این داده ها باید استخراج میشن چی هستش، قبلش سوالات رو دوباره بررسی کنیم:

- چرا مردم شکایت می کنند؟
- آیا الگویی وجود دارد یا اینکه آیا برخی روزها شکایت بیشتری نسبت به سایرین وجود دارد؟

پس از آن ، یک روز در حال پیاده شدن از اتوبوس در تورنتو بودم و عمیقا به فکر حل آن مسئله بودم و بدون اینکه روی زمین را نگاه کنم ، بیرون رفتم و در یک گودال کوچک آب افتادم و قوزک پایم در آب فرو رفت ولی پای دیگرم کامل خشک بود.

از این اتفاق کاملا اذیت شدم بعد از آن در حال پیاده روی بودم که ایده ای به ذهنم خطور کرد؛

با خودم گفتم : "خوب ، یک ثانیه صبر کن. امروز باران غیر منتظره ای بارید و من برای آن آماده نبودم از این رو خیس شدم و اشتیاقی به این نداشتم، آیا رابطه ای بین آب و هوای شدید و نوع شکایاتی که TTC دریافت می کند، وجود دارد؟

بنابراین به وب سایت آب و هوای کانادا رفتم و داده هایی از باران و بارندگی، باد و نور دریافت کردم و در آنجا ، چیز جالبی پیدا کردم.

ده روز که بیشترین شکایت در آن ثبت شده بود با آب و هوای آن روز ها مقایسه شد و باران غیر منتظره، درجه حرارت شدید، برف خیلی زیاد و روزی که باد در آن شدید بود.

بنابراین برگشتم و به مدیران TTC گفتم: "من خبرهای خوب و بدی برای شما دارم"

خبر خوب این است که می دانم چرا مردم در روزهای معین شکایت بیش از حد می کنند و خبر بد این است که شما هیچ کاری در مورد آن نمی توانید انجام دهید.

۱-۸- تعریف واژه نامه ها

API: Application programming interface رابط برنامه نویسی کاربردی به صورت خلاصه رابط برنامه نویسی، مجموعه ای از روش های تعریف شده و شفاف به منظور ارتباط بین اجزا مختلف نرم افزار می باشد. یک API خوب با فراهم کردن سنگ بناهای لازم، توسعه یک نرم افزار کامپیوتری را آسان تر می کند. یک API می تواند برای یک سیستم تحت وب، سیستم عامل،

سیستم بانک اطلاعاتی، سخت افزار کامپیوتر و یا کتابخانه نرم افزار طراحی شده باشد. مشخصات API می تواند شکل های مختلفی داشته باشد اما این مشخصات اغلب شامل روال ها، ساختمان داده ها، دسته های اشیا، متغیر ها یا دستورات فراخوانی می باشد

Info graphic: اطلاعات تصویری یا اینفوگرافی، روشی برای بیان اطلاعات، داده‌ها یا مفاهیم پیچیده مربوط به یک دانش خاص است که به شیوه‌ی دیداری یا تصویری انجام می‌شود و از سرعت و وضوح بیشتری نسبت به سایر روش‌ها برخوردار است. استفاده از تصاویر در این روش سبب می‌شود که انسان به درک بهتری از الگوها و روندها دست یابد به بیان ساده تر یکی از روش‌های ارائه‌ی اطلاعات به‌صورت دیداری است. با این استدلال که ذهن انسان داده‌های تصویری را بهتر و سریع‌تر از داده‌های متنی درک و ذخیره می‌کند؛ امروزه اینفوگرافی کاربرد گسترده‌ای در بسیاری از زمینه‌ها از آموزش و فرهنگ‌سازی گرفته تا تبلیغات و بازاریابی داراست

متن باز: Open source به نرم افزارهایی که کدهای آنها قابل دسترسی بود و توسط یک برنامه نویس یا جامعه برنامه نویسی مورد بازبینی قرار می گیرند و هدف از متن باز کردن نرم افزارها، توسعه و بهبود یا شخصی سازی نرم افزار مورد نظر می باشد. بسیاری از کتابخانه ها و نرم افزارهای بزرگ بعد از یک بلوغ اولیه، متن باز می شوند.

BSD: مخفف عبارت Berkeley Software Distribution، خانواده ای از مجوزهای نرم افزاری است که محدودیت های کمتری را در استفاده و توزیع نرم افزارهای تحتت پوشش دارند.

طبقه بندی: Classification یک الگوریتم یادگیری ماشین در دسته ی نظارت شده که هدف آن پیش بینی دسته بندی اشیا با توجه به ویژگی هایی که برای آن مشخص شده است.

رگرسیون: Regression یک الگوریتم یادگیری ماشین در دسته ی نظارت شده که هدف آن پیش بینی و بیان یک متغیر نسبت به متغیر های دیگر.

خوشه بندی: Clustering یک تکنیک یادگیری ماشین در دسته ی بدون نظارت که هدف آن پیش بینی دسته بندی خودکار اشیا مشابه در مجموعه های اشیا.

کاهش ابعاد: Dimensional Reduction یک تکنیک یادگیری ماشین برای کاهش تعداد متغیر های تصادفی در مجوعه داده که هدف آن افزایش راندمان مدل ایجاد شده توسط ماشین.

انتخاب مدل: Model selection برای مقایسه، اعتبارسنجی و انتخاب پارامترها و مدل مناسب در مدل هایی که توسط ماشین ایجاد می شوند.

پیش پردازش: Preprocessing استخراج ویژگی و نرمال سازی داده ها، قبل از تولید مدل ماشینی.

۱-۹- خلاصه فصل‌ها

در بخش های بعد به تحصیلات یک متخصص داده و ابزارهای مورد نیاز و مهارت هایی که باید برای این مسیر بیاموزد، می پردازیم. همچنین در این مسیر با یک پروژه ی عملی با ابزارهای رایگان آشنا می شویم. با اتمام این مسیر درک و بینش بهتری از یک متخصص داده داشته خواهید داشت.

فصل ۲: قدم نخست برای یادگیری علم داده

۲-۱- تحصیلات

در ابتدا طبق آمار و مصاحبه هایی که از شاخص ترین افراد [1]در این زمینه فعالیت می کنند، ثبت شده؛ همگی از مسیر دانشگاهی و در مقطع دکتری وارد این حوزه شده و قابلیت برنامه نویسی و کار با ابزارهای علم داده را داشته اند، ازین رو می توان انتظار داشت که هر متخصصی که در یک زمینه علمی دارای تجربه کافی باشد به عنوان مثال، فیزیک‌دانان تجربی نیز باید تجهیزات را طراحی، داده‌ها را جمع‌آوری کنند و آزمایش‌های متعددی انجام دهند و نتایج خود را توضیح دهند. بنابراین، شرکت‌ها به دنبال استخدام افرادی هستند که می‌توانند با داده‌های پیچیده کار کنند، افرادی که سابقه ی تحصیلی و کاری در علوم فیزیکی یا علوم اجتماعی دارند. برخی از بهترین و قوی‌ترین متخصصین داده، دارای دکترا در زمینه‌های اکولوژی و بیولوژی هستند. جورج روملیوتیس[2]، که دارای مدرک دکترای اخترفیزیک است. خیلی تعجب نمی‌کنید، اگر بدانید که بسیاری از داده‌هایی که امروزه متخصصین داده در آن کار می‌کنند به طور رسمی در علوم رایانه، ریاضیات یا اقتصاد دیده می‌شود. آن‌ها می‌توانند در هر زمینه‌ای که دارای داده‌های قوی و تمرکز محاسباتی باشد، بیرون بیایند.

۲-۲- مهارت های مورد نیاز

این بخش به توضیح علومی می پردازد که یک متخصص علوم داده باید به طور پیش زمینه ای به آنها مسلط باشد.

۲-۲-۱- ریاضیات

برای درک درست از چگونگی وقایع آتی بایستی یک بینش ریاضی به مسائل اطراف داشته باشیم که در ادامه شاخه های مورد نظر بررسی می شوند.

---

[1] مصاحبه با ۲۵ نفر از با تجربه ترین و اولین افراد که در این زمینه فعالیت داشتند و نام آن در قسمت منابع ذکر شده.

[2] رئیس یک تیم علوم داده در شرکت اینتویت واقع در سیلیکون ولی است.

۲-۲-۱-۱- جبر خطی

جبر خطی شاخه ای از ریاضیات است که به فضاهای برداری می پردازد. پایه ای ترین جز محاسباتی آن، روی بردارها و ماتریس ها انجام می گیرد که این محاسبات شامل عملیات های پایه ای ریاضی تا محاسباتی همچون میانگین یا انحراف از معیار یا محاسباتی از این قبیل هستند و همچنین دارای خواصی همچون معکوس پذیری یا ترانهاده یا ترتیب پذیری می باشد بحث فعلی لزوماً یک موضوع عمیق نخواهد بود و لذا برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۱].

"آیا چیزی مفیدتر یا بی ارزش تر از جبر هست؟"-- بیلی کانولی

۲-۲-۱-۲- آمار

آمار به ریاضیات و تکنیک هایی که داده ها را درک می کنیم اشاره دارد. این یک زمینه غنی و عظیم می باشد، بحث فعلی لزوماً یک موضوع عمیق نخواهد بود و لذا برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۲].

"حقایق سرسخت است ، اما آمار پایدارتر است." -- مارک تواین

۲-۲-۱-۳- احتمالات

نظریه ی احتمالات به شاخه‌ای از ریاضیات گویند که با تحلیل وقایع تصادفی سروکار دارد.

احتمال معمولاً مورد استفاده برای توصیف نگرش ذهن نسبت به گزاره‌هایی است که ما از حقیقت آن‌ها مطمئن نیستیم. گزاره‌های مورد نظر معمولاً از فرم "آیا یک رویداد خاص رخ می‌دهد؟" و نگرش ذهن ما از فرم "چقدر اطمینان داریم که این رویداد رخ خواهد داد؟" است. میزان اطمینان ما، قابل توصیف به صورت عددی می‌باشد که این عدد مقداری بین ۰ و ۱ را گرفته و آن را احتمال می نامیم. هر چه احتمال یک رویداد بیشتر باشد، ما مطمئن تر خواهیم بود که آن رویداد رخ خواهد داد. در واقع میزان اطمینان ما از اینکه یک واقعه (تصادفی) اتفاق خواهد افتاد.

به طور کلی بررسی اطلاعات برای کشف دانش، بدون درک احتمال و ریاضیات آن کار سختی است و لذا برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۳].

"قوانین احتمال ، به طور کلی ، بسیار دقیق است ، ولی گاه با موارد خاص مغایر است."
-- ادوارد گیبون

۲-۲-۲- مهارت برنامه نویسی

برای پیاده سازی الگوریتم های ریاضی و کارهایی که می توانیم روی داده ها انجام دهیم نیاز به مهارت برنامه نویسی است، از آنجایی که زبان هایی زیادی برای این کار ایجاد شده اند ولی از مشهور ترین آنها می توان به پاینون و R نام برد، زبان پایتون که در همه ی زمینه های علمی و سخت افزاری قابل دسترسی و دارای ابزار های رایگان زیادی می باشد ولی زبان R تنها برای حوزه ی آماری ایجاد شده و هر دو گزینه های خوبی برای متخصصین داده می باشند و لذا برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۴].

۲-۲-۳- پردازش داده

برای اینکه یک متخصص داده شوید حتما به داده نیاز دارید. در واقع، به عنوان یک دانشمند داده ، بخش بزرگی از وقت خود را صرف جمع آوری، تمیز کردن و تبدیل داده ها می کنید.

انواع داده را به فرمت های متنی، صوتی، تصویری و چند رسانه ای در اختیار داریم که بخش اعظمی از فعالیت متخصصین داده در فایل های متنی صرف می شود که همین داده های متنی شامل دو دسته ی ساختار یافته و بدون ساختار تقسیم بندی می شوند که در مورد آنها مفصلا صحبت شد.

برای نگهداری داده ها، تکنولوژی هایی همچون SQL و NoSQL و Excel و غیره ایجاد شده که متخصص داده باید نحوه ی جستجوهای پیچیده را برای استخراج داده ها مسلط باشد و برای دسترسی یا جمع آوری آنها بایستی از API ها استفاده نمایند یا اگر داده هایی را که خودتان جمع آوری کرده اید و روی حافظه کامپیوتر نگه داری می کنید، می توان به طور مستقیم و از روی حافظه

کامپیوتر باگذاری نموده و مورد تحلیل قرار داد و لذا برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۵].

"برای نوشتن این کتاب، سه ماه طول کشید. برای تصور کردن آن ، سه دقیقه؛ برای جمع آوری داده های آن، تمام زندگی من زمان سپری شد." --اف. اسکات فیتزجرالد

۲-۲-۴- مصورسازی داده

نمایش تصویری اطلاعات را مصورسازی داده [۱]می‌نامند. در این شیوه، اطلاعات و داده‌ها، به واسطه تصاویر و شکل‌ها، قابل نمایش شده و بیننده قادر به درک سریع‌تر و بهتر اطلاعات نهفته در داده‌ها خواهد شد.

ارتباط تصویری، نگاشتی سیستمی، بین تصاویر و مقادیر متغیرها (کمی یا کیفی) است که در مصورسازی داده به بهترین نحو صورت می‌گیرد. ویژگی‌های متنوع در نمودارها[۲]، آن‌ها را به ابزاری مهم برای مصورسازی داده تبدیل کرده است. وجود رنگ، اندازه‌های مختلف برای هر دنباله از داده‌ها و همچنین نمایش روند تغییرات، از مواردی است که نمودارها را برای ارتباط تصویری بهتر، نسبت به جداول و گزارشات متنی، متمایز می‌کند.

نمودارهای آماری، اینفوگرافیک[۳]، شکل‌ها و نمادها[۴] ابزارهایی مهمی در مصورسازی داده محسوب می‌شوند. به این ترتیب، داده‌های عددی (کمی) یا اطلاعات کیفی، به صورت تصاویر، خطوط، میله یا قطاع‌هایی از دایره، تبدیل شده و اطلاعات مربوطه را منتقل می‌کنند.

هدف اصلی مصورسازی داده یا تجسم آن‌ها، برقراری ارتباط واضح و مؤثر از طریق ابزارهای گرافیکی است. این بدان معنا نیست که لزوما مصورسازی داده باعث ایجاد یک تصویر زیبا شود بلکه

---

[۱] Data visualization

[۲] Graphs

[۳] Info Graphics به تصویری که اطلاعات خاصی را در درون خود جای داده است.

[۴] Icons

درک اطلاعات به شیوه ساده و راحت منظور این روش توصیفی محسوب می‌شود. بـه طـور مـؤثر ، هـم فرم زیبا شناختی و هم عملکرد باید دست به دست هم دهند و با برقراری ارتباط بـا جنبـه‌های اصـلی آن به روشی بصری، اطلاعات نهفته در داده‌های نسبتاً پراکنـده و پیچیـده را ارائـه دهنـد. مصورسـازی داده که بدون هدف تولید شده و فقط جاذبه‌های بصری داشـته باشـند، منظـور نظـر مصورسـازی داده نخواهد بود و لذا برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهـای ایـن رسـاله مراجعـه نمایید [۶].

۲-۲-۵- یادگیری ماشین

"شاخه‌ای از علـم کـه بـه رایانـه‌ها توانـایی یـادگیری بـدون یـک برنامه‌نویسـی خـاص را مـی‌دهـد."
-- آرتور سموئل

تعریف کلی یادگیری ماشین[۱]، به مطالعـه‌ی علمـی الگوریتم‌هـا و مـدل‌های آمـاری مـورد اسـتفاده‌ی سیستم‌های کامپیوتری است که بجای استفاده از دستورالعمل‌های واضـح از الگوهـا و اسـتنباط بـرای انجام وظایف سود می‌برند. به بیـان دیگـر ایـن علـم بـه عنـوان زیـر مجموعـه‌ای از هـوش مصـنوعی، الگوریتم‌های ماشین را به یک مدل ریاضی بر اساس داده‌ها تبدیل می نماید.

شما احتمالاً چندین بار در روز از یادگیری ماشین استفاده می‌کنید، حتی بدون آنکه بدانید. هر بار کـه شما یک جستجوی اینترنتی در گوگل یا بینگ انجام می‌دهید، یادگیری ماشینی انجام می‌شود چراکـه نرم‌افزار یادگیری ماشینی آن‌ها چگونگی رتبه‌بندی برای یک صفحه وب را درک کرده‌است. هر بار کـه ایمیل خود را چک می‌کنید و فیلتر هرزنامه[۲] شما را از داشتن مجدد هزاران هرزنامه خلـاص مـی‌کنـد نیز به همین دلیل است که رایانه‌ی شما آموخته‌است که هرزنامه را از غیرهرزنامه تشخیص دهـد. ایـن همان یادگیری ماشین است. این علمی است که باعث می‌شود رایانه‌ها بدون نیاز به یک برنامـه صـریح در مورد یک موضوع خاص یاد بگیرند.

---

[۱] Machine learning

[۲] Spam مزاحم

یکی از تقسیم‌بندی‌های متداول در یادگیری ماشینی، تقسیم‌بندی بر اساس نـوع داده‌هـای در اختیـار کارگزار هوشمند قرار می دهند انجام می شود و به طور کلی داریم:

۱. یادگیری با نظارت[۱]: به رایانه آموزش می‌دهیم که چگونه کاری را خودش انجام دهد.

۲. یادگیری بی‌نظارت: به رایانه اجازه می‌دهیم که خودش یاد بگیرد.

۳. انواع دیگری همچون تقویتی [۲]و غیره.

برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۷].

۲-۲-۶- یادگیری عمیق[۳]

شاخه ای از بحث یادگیری ماشین و هوش مصنوعی و مجموعه ای از الگوریتم هایی اسـت کـه تلاش می کنند، مفاهیم انتزاعی سطح بالا را با استفاده از یادگیری در سطوح و لایه های مختلف مـدل کنند. مطالعات بالینی نشان می دهند، که ساختار مغز پستانداران از معماری شبکه های عصبی عمیـق بهره می برد که در آن، مفاهیم انتزاعی در لایه های مختلف، به ترتیب از مفاهیم و ویژگی هـای سـاده تا مفاهیم سطح بالا، در نواحی مختلف قشر مغز، پردازش می شوند. ایـده یـادگیری عمیـق بـا الهـام از ساختار طبیعی مغز انسان و به کمک امکانات و فـن آوری هـای جدیـد، توانسـته اسـت در بسـیاری از حوزه های مربوط به هوش مصنوعی و یادگیری ماشین، موفقیت های چشم گیری را کسب کند.
مهم ترین مزایای یادگیری عمیق عبارت اند از:

- یادگیری خودکار ویژگی ها
- یادگیری چند لایه ویژگی ها
- دقت بالا در نتایج
- قدرت تعمیم بالا و شناسایی داده های جدید

---

[۱] Supervisor learning

[۲] Reinforcement learning

[۳] Deep learning

- پشتیبانی گسترده سخت افزاری و نرم افزاری
- پتانسیل ایجاد قابلیت ها و کاربردهای بیشتر در آینده

در سال های اخیر، یادگیری عمیق، تحول بزرگی را در یادگیری ماشین و هوش مصنوعی ایجاد کرده است. از سال ۲۰۱۲ تا کنون، تمامی رتبه های برتر چالش شناسایی بصری[۱]، که به جام جهانی بینایی ماشین معروف است، از شبکه های عصبی عمیق استفاده کرده اند. همچنین، تمام روش های برتر در رقابت های دسته بندی تصاویر اعداد دست نویس [۲]نیز به مدل های شبکه عصبی عمیق تعلق دارد. از آن سال به بعد، شرکت های بزرگ نرم افزاری و سخت افزاری مانند Google, Microsoft, NVIDIA نیز بخش مهمی از فعالیت های پژوهشی و تجاری خود را به یادگیری عمیق اختصاص داده اند.

با این که یادگیری عمیق در سال های ابتدایی توسعه خود قرار دارد، اما روند تحقیقات، مقالات و سرمایه گذاری های شرکت های بزرگ در این حوزه، نشان دهنده گسترش روز افزون کاربردهای یادگیری عمیق است. یادگیری عمیق تا کنون در کاربردهای گوناگون داده کاوی، پردازش تصویر و صدا، رباتیک و پزشکی مورد استفاده قرار گرفته است. طبق پیش بینی های مراکز علمی، در سال های آینده، بسیاری از تحقیقات، کاربردها و مشاغل موفق، به طور مستقیم یا غیرمستقیم از یادگیری عمیق بهره خواهند برد، برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۸].

۲-۲-۷- پردازش زبان طبیعی

پردازش زبان‌های طبیعی یکی از زیرشاخه‌های بااهمیت در حوزه ی گسترده ی علوم رایانه، هوش مصنوعی، که به تعامل بین کامپیوتر و زبان‌های (طبیعی) انسانی می‌پردازد؛ بنابراین پردازش زبان‌های طبیعی بر ارتباط انسان و رایانه، متمرکز است. پس چالش اصلی و عمده در این زمینه درک زبان طبیعی و ماشینی کردن فرایند درک و برداشت مفاهیم بیان‌شده با یک زبان طبیعیِ انسانی است.

---

[۱] ImageNet

[۲] MINIST مخفف Modified National Institute of Standards and Technology database بازشناسی ارقام دست نویس

به تعریف دقیق‌تر، پردازش زبان‌های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و زبان نوشتاری. بدین معنی که رایانه‌ها را قادر سازیم که گفتار یا نوشتار تولید شده در قالب و ساختار یک زبان طبیعی را تحلیل و درک نموده یا آن را تولید نمایند. در این صورت، با استفاده از آن می‌توان به ترجمه ی زبان‌ها پرداخت، از صفحات وب و بانک‌های اطلاعاتیِ نوشتاری جهت پاسخ دادن به پرسش‌ها استفاده کرد، یا با دستگاه‌ها، مثلاً برای مشورت گرفتن به گفت‌وگو پرداخت. این‌ها تنها مثال‌هایی از کاربردهای متنوع پردازش زبان‌های طبیعی هستند.

هدف اصلی در پردازش زبان طبیعی، ایجاد تئوری‌هایی محاسباتی از زبان، با استفاده از الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم رایانه است. بدیهی است که در راستای تحقق این هدف، نیاز به دانشی وسیع از زبان است و علاوه بر محققان علوم رایانه، نیاز به دانش زبان شناسان نیز در این حوزه می‌باشد. با پردازش اطلاعات زبانی می‌توان آمار مورد نیاز برای کار با زبان طبیعی را استخراج کرد. کاربردهای پردازش زبان طبیعی به دو دسته کلی قابل تقسیم است:

۱. کاربردهای نوشتاری

۲. کاربردهای گفتاری.

از کاربردهای نوشتاری آن می‌توان به استخراج اطلاعاتی خاص از یک متن، ترجمه یک متن به زبانی دیگر یا یافتن مستنداتی خاص در یک پایگاه داده نوشتاری (مثلاً یافتن کتاب‌های مرتبط به هم در یک کتابخانه) اشاره کرد.

نمونه‌هایی از کاربردهای گفتاری پردازش زبان عبارتند از: سیستم‌های پرسش و پاسخ انسان با رایانه، سرویس‌های اتوماتیک ارتباط با مشتری از طریق تلفن، سیستم‌های آموزش به فراگیران یا سیستم‌های کنترلی توسط صدا. در سالهای اخیر این حوزه تحقیقاتی توجه دانشمندان را به خود جلب کرده‌است و تحقیقات قابل ملاحظه‌ای در این زمینه صورت گرفته است.

برای مطالعه دقیق تر می توانید به منابع معرفی شده در انتهای این رساله مراجعه نمایید [۹].

فصل ۳: استقرا و پیاده سازی

۳-۱- مقدمه

برای جمع آوری و نگهداری داده ها روی سیستم و همچنین پردازش آنها و رسیدن بـه یـک تحلیـل نیاز به پیاده سازی و پیش نیاز هایی می باشد که در ادامه آنها را بررسی می نماییم.

۳-۲- سیستم عامل

سیستم عامل، نرم‌افزار سیستمی ای است که مدیریت منابع رایانه را به عهده گرفته و بسـتری را فراهم می‌سازد که نرم‌افزار کاربردی اجرا شده و از خدمات آن استفاده کنند.

برای استفاده از نرم افزارها و ابزارهای دانشمند داده می توان از هر سیستم عاملی استفاده کرد ولی بـه طور خاص می توان از سیستم عامل لینکوس[۱] استفاده نمـود کـه پیشـنهاد خـوبی دانسـت، هـر چنـد نسبت به رایگان بودن آن در کنار سیستم عامل هایی که توسط شرکت مایکروسافت یـا اپـل طراحـی کرده دارای ابزارهای زیادی می باشد و به همان مقدار دارای هسته ی[۲] قدرتمند می باشد و بـه تـازگی شرکت مایکروسافت برای نسخه ی ویندوز[۳] خود یک هسته ی لینوکس هم در کنار هسته ی سیسـتم عامل خود در اختیار کاربران قرار داده و آنها مـی تواننـد بـا داشـتن ایـن سیسـتم عامـل، همزمـان از هسته ی لینوکسی هم استفاده کنند.

[۱] Linux operating system

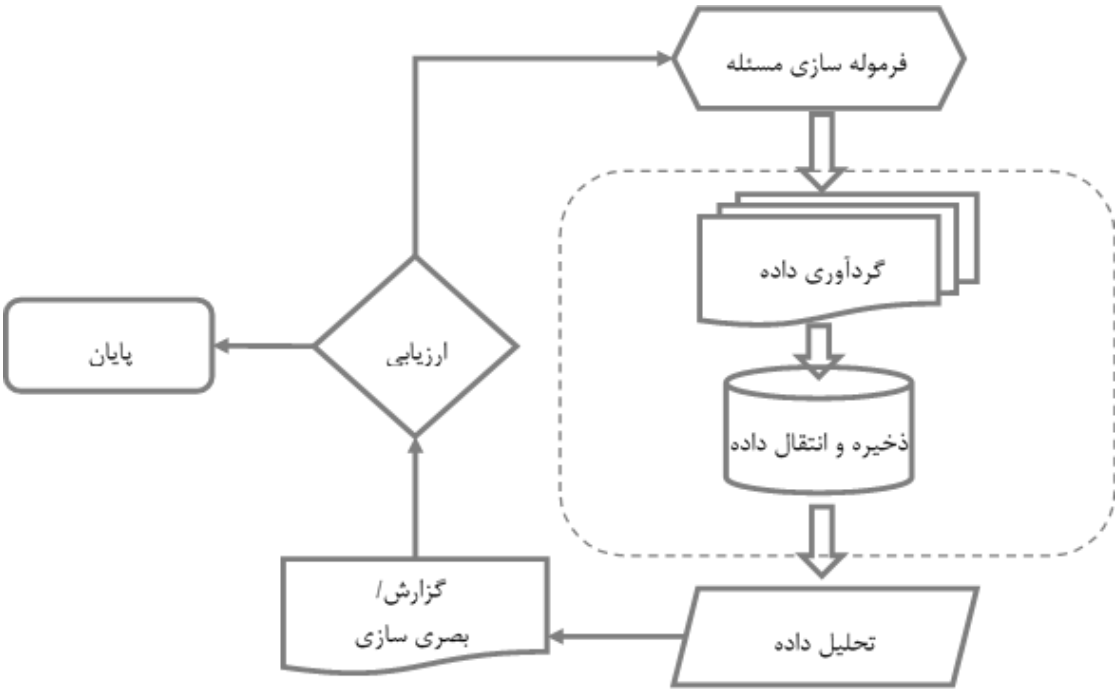
[۲] Kernel

[۳] Windows

۳-۳- زبان برنامه نویسی

زبــان پــایتون [۱] و زبــان R[۲] بــرای قابلیــت تحلیــل داده و فرمــول ســازی مســئله و گــزارش یــا بصری سازی داده ها بکار گرفته می شود و به کمک کتابخانه هــای رایگــان ایــن دو زبــان، مــی تــوان مدل ها و توابع آماری را بکار گرفته و در نهایت از داده ها به عنوان راهی برای کشف دانش یا پاســخی به پرسش های ما استفاده نماید.

تشریح جربان داده در پروژه هایی که با داده های کلان در ارتباط هست در شکل زیر آمده است.



شکل (۳-۱) روند جریان داده های کلان در استقرار پروژه

[۱] Python

[۲] زبان R یک زبان آماری می باشد که قابلیت هایی همچنون پردازش داده و مصور سازی داده ها رو انجام می دهد.

۳-۴- مدیریت داده ها

پس از انتخاب یک پروژه عملی می توان راجب فرمت داده هـا و نحـوه ی بارگـذاری آنهـا در برنامـه و همچنین پیش پردازش آنها و انتخاب ویژگی هایی که می تواند در رسیدن به فرضـیه ی مـورد انتظـار راهنمای ما باشد، صحبت نمود.

۳-۴-۱- آپاچی هدوپ

آپاچی [۱۰] یک نرم افزار سرور وب می باشد که منبع باز [۱] و آزاد تحت گـواهی آپـاچی [۲] منتشـر مـی شود و توسط یک جامعه ی متن باز، توسعه و نگه داری می شود؛ این درحالی است که آپـاچی هـدوپ [۱۱] به عنوان یکی از پروژه های این جامعه متن باز بوده و نقش آن، یـک چـارچوب مـدیریت داده های توزیع شده است که توسط یک مدل سـاده برنامـه نویسـی پیـاده سـازی شـده و داده هـا را روی مجموعه های کوچک تری به با خوشه تقسیم بندی می نماید و به گونه ای طراحی شده که خرابی هر بخش به صورت خودکار توسط چارچوب کنترل می شود و بدون متوقف شدن و با داشتن چنـد کپـی از داده هایی که خراب شده اند آنها را دوباره به سیستم بازگردانی می نماید و در ادامه اجزای تشـکیل دهنده [۱۲] این چارچوب داریم:

HDFS: سیستم فایل توزیع شده هدوپ

YARN: یک میانجی منبع باز

MapReduce: پردازش داده های مبتنی بر برنامه نویسی

Spark: پردازش داده های درون حافظه

PIG, HIVE: پردازش داده مبتنی سرویس پرس و جو

---

[۱] Open source متن باز

[۲] Apache License 2.0

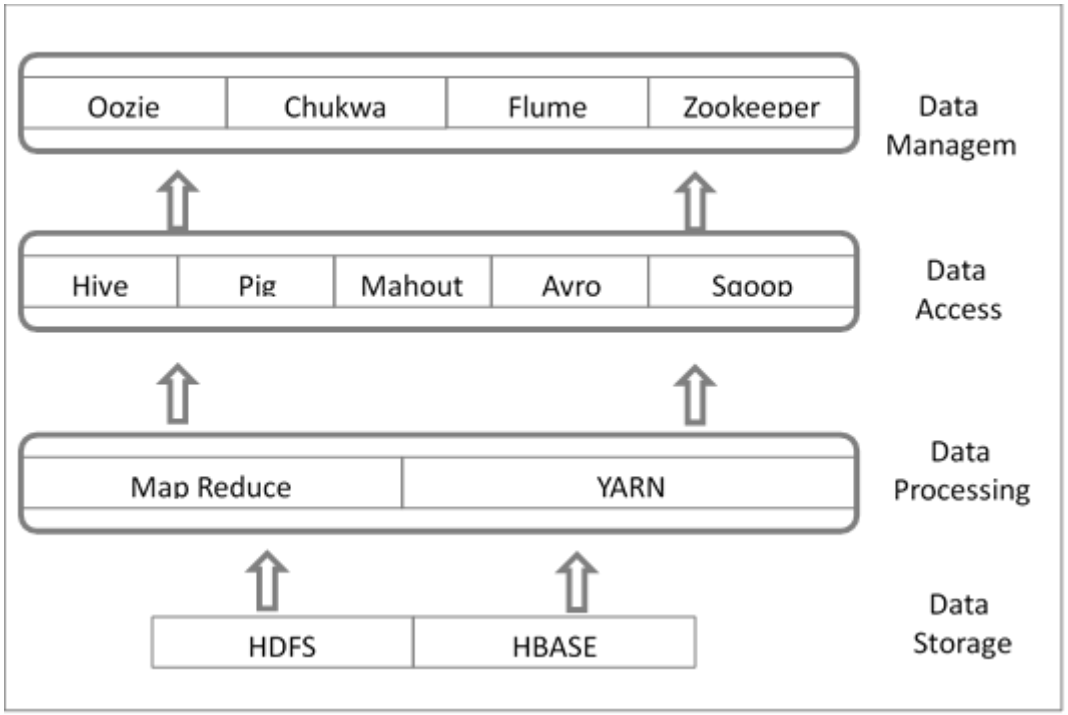
HBase: بانک اطلاعاتی NoSql

Mahout, Spark MLLib: کتابخانه های الگوریتم یادگیری ماشین

Solar, Lucene: جستجو و نمایه سازی

Zookeeper: مدیریت خوشه بندی[۱]

Oozie[۸]: سیستم زمانبندی گردش کار برای مدیریت مشاغل



شکل (۳-۲) اجزای چارچوب هدوپ

۳-۴-۲- ژوپیتر نوت بوک[۲]

منظور از این کلمه بزرگترین سیاره منظومه شمسی نیست و بلکه وب نرم افـزار مـتن بـاز کـه قابلیـت نوشتن کدهای پایتون را به صورت پیش فـرض پشـتیبانی مـی کنـد، البتـه زبـان هـای دیگـری نیـز

[۱] Cluster management

[۲] Jupyter Notebook

پشتیبانی می کند و هدف کلی آن تعامل با کاربر برای یک فایل، کد، عکس و چیزهایی از این قبیل را در مرورگر ویرایش و اجرا کنیم.

ژوپیتر نوت بوک ها یکی از ابزارهای اصلی تقریبا همه ی متخصصین علـم داده هسـتند و از دو جهـت تاثیر گسترده ای بر حوزه علم داده ها داشته اند: اول از همه ژوپیتر نوت بوک ها فضایی بـرای تکـرار و آزمایش فعالیت های مختلف علـم داده هـا را در اختیـار مـا قـرار مـی دهنـد و همـین پیـاده سـازی فرآیندهای تکرارشونده علم داده ها را برای ما راحت تر می کند. ثانیا دلیل دیگـر تاثیرگـذاری ژوپیتـر نوت بوک این است که آن ها از زبان نشانه گذاری مارکدان [۱]پشتیبانی می کنند؛ یعنی به عبارتی دیگر ما هم می توانیم در آن ها هم کد بنویسیم و هم می توانیم برای کدهایمان (فراتر از کامنت هـایی کـه قبلا یا الان استفاده می کنیم) توضیحاتی مثل متن و شکل قرار دهیم. همین مساله باعث شده کـه بـه طور خاصی ژوپیتر نوت بوک ها در آموزش دادن مفاهیم علم داده ها، برنامه نویسی، پیاده سازی مقاله ها و غیره نقشی انقلابی داشته باشند.

ژوپیتر نوت بوک ها از دو مولفه اصلی تشکیل می شوند: کرنل[۲] و داشبورد[۳].

کرنل وظیفه این را دارد که کدی که ما نوشته ایم را اجرا کند.

داشبورد هم به ما این امکان را می دهد که نوت بوک ها را ببینیم، ویرایش کنیم و حتی کرنـل مـورد استفاده را تغییر دهیم یا ببندیم [۱۳].

۳-۴-۳- کتابخانه ها

Pandas

یک کتابخانه متن باز با گواهی BSD[۱] برای دستکاری و تجزیه و تحلیل داده ها در زبان پـایتون ایجـاد شده و به طور خاص دادها را به صورت جداول اعداد و سری های زمانی ارائه می دهند [۱۴].

[۱] Markdown زبان نگارشی که هدف آن نوشتن توضیحات دقیق تر با قابلیت استفاده تصاویر و متن و کد

[۲] Kernel هسته ی پردازشی

[۳] Dashboard میزکار

یک کتابخانه متن باز با گواهی BSD برای عملیات ریاضی سطح بالا در زبان پایتون ایجـاد شـده و بـه طور خاص با آرایه ها و ماتریس ها تعامل دارد [۱۵]. ۱۱

Scikit-learn

یک کتابخانه متن باز با گواهی BSD برای یادگیری ماشین در زبان پایتون ایجاد شده و به طـور کلـی قابلیت هایی برای طبقه بندی، رگرسیون، خوشه بندی، کاهش ابعاد، انتخاب مدل، پیش پـردازش داده های ارائه می دهند [۱۶]. ۱۲

Keras

یک کتابخانه متن باز با گواهی MIT برای یادگیری عمیق در زبان پایتون ایجاد شده و بـه طـور کلـی قابلیت هایی برای طبقه بندی، رگرسیون، خوشه بندی، کاهش ابعاد، انتخاب مدل، پیش پـردازش داده های ارائه می دهند [۱۷]. ۱۳

Mathplotlib

یک کتابخانه متن باز و جامع برای ایجاد تصاویر استاتیک، متحرک و تعاملی در پایتون است [۱۸]. ۱۴

Spacy

یک کتابخانه متن باز با گواهی MIT برای پردازش پیشرفته زبان طبیعی به زبان پایتون ایجـاد شـده و به طور کلی قابلیت هایی برای طبقه بندی، رگرسیون، خوشه بندی، کاهش ابعاد، انتخاب مـدل، پـیش پردازش داده های ارائه می دهند [۱۹]. ۱۵

Lightgbm

یک کتابخانه متن باز که مبتنی بر الگـوریتم درخـت تصـمیم گیـری، طبقـه بنـدی و رتبـه بنـدی در یادگیری ماشین با زبان های پایتون و C++ و R ایجاد شده است [۲۰]. ۱۶

[۱] Berkeley Software Distribution گواهی انتشار نرم افزار بِرکِلی

فصل ۴: نتیجه گیری

۴-۱- اصول داستان گویی

با تکمیل شدن تحلیل و رسیدن به پاسخِ پرسش ها، بایستی نتایج را به گونه ای بیـان کـرد کـه به سادگی برای همگان قابل فهم باشد و بدین منظور می توان از اصول زیر پیروی کرد.

۴-۱-۱- مصورسازی نتایج

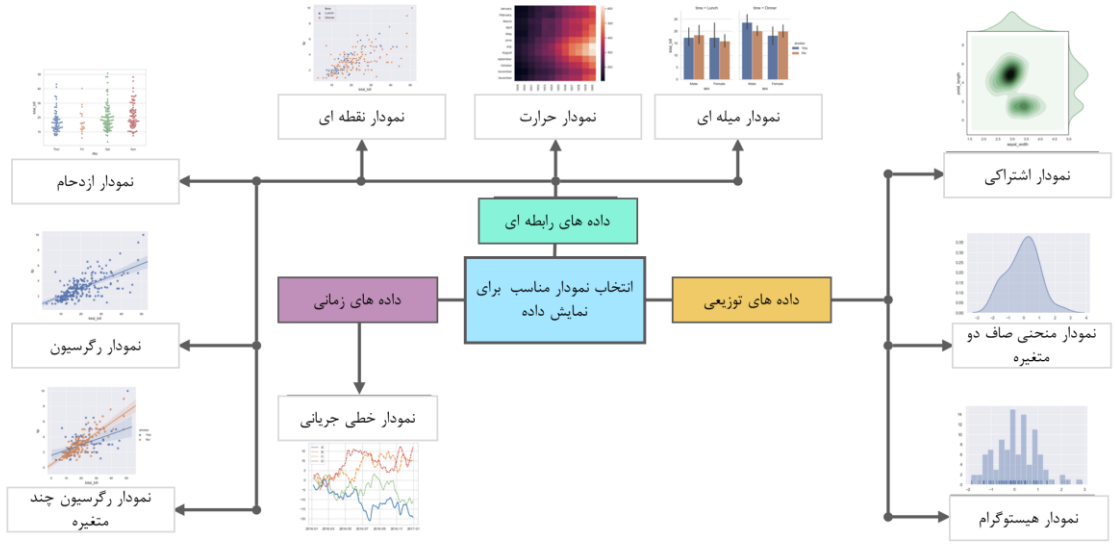
در انتخاب نوع نمودار یا طراحی اینفوگرافیک بایستی به نکات زیر دقت نمایید:

- برای پریسش های هدفمند فیلدهای که از داده انتخاب می شود را به درستی بیابیم، بـه بیـان دیگر موجودیت های فعلی می توانند به تنهای پاسخ ما را بدهند و یا نیاز است آنهـا را تغییـر دهیم یا شاید نیاز باشد اطلاعات بیشتری را جمع آوری نموده و با آنها ترکیـب نمـاییم، بـرای این کار نیاز است داستان ها و پرسش هایی که توسط دیگر متخصصین علوم داده انجام شـده را مطالعه نماییم و به درک و بینش عمیق تری نسبت به داده ها برسیم.
- انتخاب الگوریتم مناسب برای رسیدن به نتایج دقیق تر می توانـد بـه پاسـخی دقیـق تـر و بـا ارزش تر کمک نماید و همین نتیجـه بـا مطالعـه تکنولـوژی هـای جدیـد تـر حاصـل شـود و می تواند به رسم نمودار های دقیق تر کمک نمود.
- برای تصمیم گیری انتخاب بهترین نمودار برای داده ها می توان سـه دسـته بنـدی زیـر را در نظر گرفت.

۱. روند جریانی داده : یک روند به عنوان الگویی از تغییر تعریـف مـی شـود. نمودارهـای خط برای بهتر نشان دادن روندها در یک دوره زمانی انتخاب مناسبی است و چنـدین خط می توانند برای نمایش روندها در بیش از یک گروه استفاده شوند.

۲. ویژگی های مرتبط به هـم (رابطـه ای) : انـواع نمودارهـای مختلفـی وجـود دارد کـه می توانید از آنها برای درک روابط بین متغیرها در داده های خود استفاده کنید.

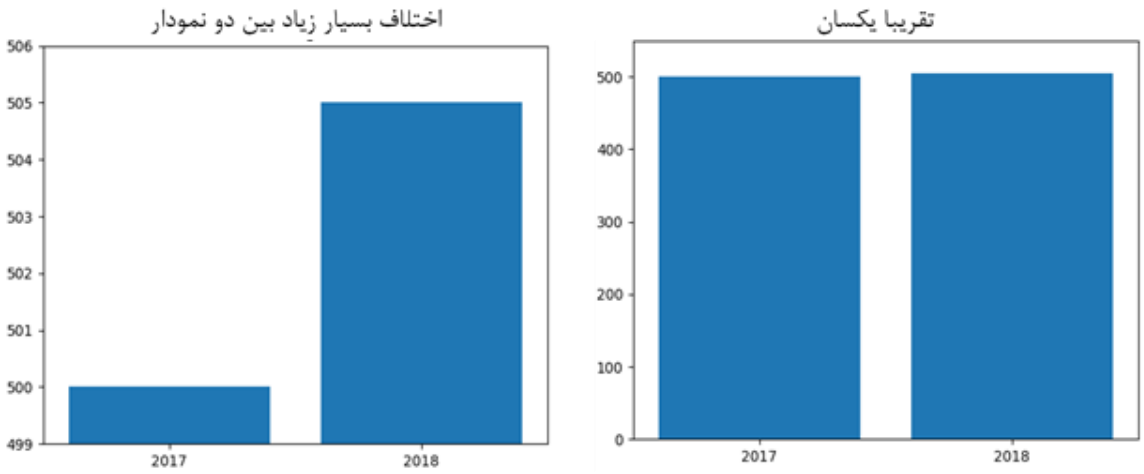
۳. داده های توزیعی : با مصورسازی داده ها، می توان مقادیر احتمالی مورد نظر را بـرای یک متغییر انتظار داریم مشاهده کنیم، به همراه اینکـه چقـدر احتمـال دارد درسـت باشد.



شکل (۴-۱) راهنمای انتخاب نمودار مناسب برای نمایش داده ها

۴-۱-۲- اشتباهات رایج

گاه در مصور سازی داده ها و انتخاب یک نمودار مناسب، ممکن است انتخـاب درسـتی داشـته باشـید ولی در نحوه ی معیار بندی آن دچار خطا شوید، منظور آن در زیر مشخص است:



شکل (۴-۲) مصورسازی یکسان با مقیاس بندی متفاوت

با اینکه هر دو نمودار با یک داده ی یکسان مصور سازی شده اند ولی در شکل سمت راست هر دو نمودار میله ای بسیار به هم نزدیک هستند و در شکل سمت چپ این حس به بیننده تلقین می شود که دو نمودار میله ای دارای اختلاف زیادی هستند، پس بایستی در انتخاب معیار معقولانه عمل نماییم و شروع مقیاس بندی را از صفر شروع نموده تا بیننده دچار خطا نشود.

.

مراجع

مراجع

[۱]منابع برای آموختن جبر خطی

1. Linear Algebra, by Jim Hefferon (Saint Michael's College) as joshua.smcvt.edu/linearalgebra.
2. Linear Algebra, by David Cherney, Tom Denton, Rohit Thomas, and Andrew Waldron (UC Davis) as math.ucdavis.edu/~linear.
3. Linear Algebra Done Wrong, by Sergei Treil (Brown University), is a more advanced introduction. as math.brown.edu/~treil/papers/LADW/LADW_2017-09-04.pdf

[۲]کتاب و کتابخانه های برنامه نویسی به زبان پایتون برای آمار

1. SciPy, pandas, and StatsModels all come with a wide variety of statistical functions.
2. Introductory Statistics, by Douglas Shafer and Zhiyi Zhang (Saylor Foundation) as open.umn.edu/opentextbooks/textbooks/introductory-statistics
3. OnlineStatBook, by David Lane (Rice University) as onlinestatbook.com
4. Introductory Statistics, by OpenStax (OpenStax College) as openstax.org/details/introductory-statistics

[۳]کتاب و کتابخانه هایی که به زبان پایتون نوشته شده برای درک احتمالات

1. scipy.stats contains PDF and CDF functions for most of the popular probability distributions as docs.scipy.org/doc/scipy/reference/stats.html

2. Introduction to Probability, by Charles M. Grinstead and J. Laurie Snell (American Mathematical Society) as dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html

[۴]منبع آموزش زبان پایتون و آر

1. main reference for Python as docs.python.org
2. main reference for R as r-project.org/other-docs.html
3. Kaggle course online python as kaggle.com/learn/python

[۵]منبع آموزشی برای پیش پردازش داده

1. Preprocessing course online as kaggle.com/learn/intro-to-machine-learning
2. Preprocessing course online as kaggle.com/learn/intermediate-machine-learning
3. Data cleaning course online as kaggle.com/learn/data-cleaning

[۶]منبع آموزشی و کتابخانه برای مصورسازی داده

1. Data Science from Scratch First principle by Joel Grus, chapter 3. Visualizing Data.
2. Data Visualization course online as kaggle.com/learn/data-visualization.
3. The matplotlib Gallery will give you a good idea of the sorts of things you can do with matplotlib (and how to do them) as matplotlib.org/gallery.htm.
4. Seaborn is built on top of matplotlib and allows you to easily produce prettier (and more complex) visualizations as seaborn.pydata.org.
5. Altair is a newer Python library for creating declarative visualizations as altair-viz.github.io
6. D3.js is a JavaScript library for producing sophisticated interactive visualizations for the web. Although it is not in Python, it is widely used, and it is well worth your while to be familiar with it as d3js.org.
7. Bokeh is a library that brings D3-style visualizations into Python as docs.bokeh.org/en/latest.

[۷]منبع آموزشی برای یادگیری ماشین

1. Kaggle course online intro to machine learning as kaggle.com/learn/intro-to-machine-learning

2. Kaggle course online intermediate machine learning as kaggle.com/learn/intermediate-machine-learning
3. Kaggle course online pandas as kaggle.com/learn/pandas
4. Kaggle course online data visualization as kaggle.com/learn/data-visualization
5. Kaggle course online feature engineering as kaggle.com/learn/feature-engineering
6. Kaggle course online advanced sql as kaggle.com/learn/advanced-sql
7. Data Science from Scratch First principle by Joel Grus, chapter 3. Visualizing Data.

[۸]منبع آموزشی برای یادگیری عمیق

1. Kaggle course online deep learning as kaggle.com/learn/deep-learning
2. Data Science from Scratch First principle by Joel Grus, chapter 19. Deep learning.
3. The canonical textbook Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (MIT Press), is freely available online. It is very good, but it involves quite a bit of mathematics as deeplearningbook.org.
4. Francois Chollet's Deep Learning with Python (Manning) is a great introduction to the Keras library, after which our deep learning library is sort of patterned as manning.com/books/deep-learning-with-python.
5. PyTorch for deep learning. Its website has lots of documentation and tutorials as pytorch.org.

[۹]منبع آموزشی برای پردازش زبان طبیعی

1. Kaggle course online natural language processing as kaggle.com/learn/natural-language-processing.
2. Data Science from Scratch First Principle by Joel Grus, chapter 21 Natural Language Processing
3. NLTK is a popular library of NLP tools for Python. It has its own entire book, which is available to read online, as nltk.org and nltk.org/book.
4. Gensim is a Python library for topic modeling, which is a better bet than our from-scratch model as radimrehurek.com/gensim.
5. SpaCy is a library for "Industrial Strength Natural Language Processing in Python" and is also quite popular as spacy.io.
6. Andrej Karpathy has a famous blog post, "The Unreasonable Effectiveness of Recurrent Neural Networks", that's very much worth reading as karpathy.github.io/2015/05/21/rnn-effectiveness
7. My day job involves building AllenNLP, a Python library for doing NLP research. (At least, as of the time this book went to press, it did.) The library is quite beyond the scope of this book, but you might still find it interesting, and it has a cool interactive demo of many state-of-the-art NLP models.
8. Repository at NLP course and data set as github.com/lpln25/NLP-course.

Apache.org [۹]

Hadoop.apache.org [۱۰]

Geeksforgeeks.org/hadoop-ecosystem [۱۱]

Oozie.apache.org [۱۲]

Dataio.ir [۱۳]

Pandas.pydata.org [۱۴]

Numpy.org [۱۵]

Scikit-learn.org [۱۶]

Keras.io [۱۷]

Matplotlib.org [۱۸]

Spacy.io [۱۹]

Lightgbm.readthedocs.io [۲۰]

[۲۱] کتاب مصاحبه با ۲۵ متخصص علم داده

The Data Science Handbook, Advice and Insights from 25 Amazing Data Scientists, By Carl Shan, Henry Wang, William Chen and Max Song.

[۲۲]مقاله علوم داده به عنوان جذاب ترین شغل قرن ۲۱

Data Scientist: The Sexiest Job of the 21st Century, by Thomas H. Davenport and D.J. Patil

Blog.faradars.org [۲۳]

Fa.wikipedia.org [۲۴]

Apache.org [۲۵]