بسمه تعالی

مقدمه ای بر یادگیری ماشین

نیمسال اول ۹۸-۹۹

تمرین عملی سری سوم

مدرس: صابر صالح

● مهلت تحویل تمرین ها: ۱۳۹۸/۹/۱۲ ●

تمرین‌های تحلیل داده

○ مقدمه

در دنیای امروز علوم دادگان (Data Science) به زمینه ی مهمی در تکنولوژی تبدیل شده و بسیاری از فن آوری های جدید بر این پایه شکل گرفته اند. در علوم داده داشتن دیتای مفید و استفاده ی درست از آن اهمیت ویژه ای دارد به طوری که بسیاری از غول های دنیای تکنولوژی بر پایه ی استخراج و بهره برداری از دادگان بنا شده اند. در این میان، شناخت درست داده و استفاده ی صحیح و مطلوب از آن، امری بسیار حیاتی است. اگر شناخت کافی از یک داده نداشته باشیم و به مفاهیم پایه ای در مورد آن مسلط نباشیم، هیچ گاه نخواهیم توانست همه ی آن ارزشی که درون داده وجود دارد را بهره برداری کنیم. آنالیز درست داده موجود به اندازه ی ساختن مدل های یادگیری ماشین مهم است. متخصصین علوم داده پیش از شروع به استفاده از الگوریتم های یادگیری ماشین به تحلیل اکتشافی داده (Exploratory Data Analysis) می پردازند. تحلیل اکتشافی داده، یکی از ابزارهای مهم برای شناخت داده است که هر چند در نگاه اول شاید روش مهم یا پیچیده ای به نظر نیاید ولی بخش مهمی از هر فرآیند یادگیری یا تحلیل داده را تشکیل می دهد. در این تمرین این مهارت مهم را تجربه خواهیم کرد. این تمرین نیمه‌ی اول از یک تمرین دو قسمتی می‌باشد. در این قسمت روی دیتاست موجود کار کرده و تحلیل اکتشافی داده را تمرین می کنید. در تمرین بعد با استفاده از دستاوردهای این تمرین روی ساخت مدل مناسب یادگیری ماشین کار خواهید کرد.

○ آشنایی با مسئله

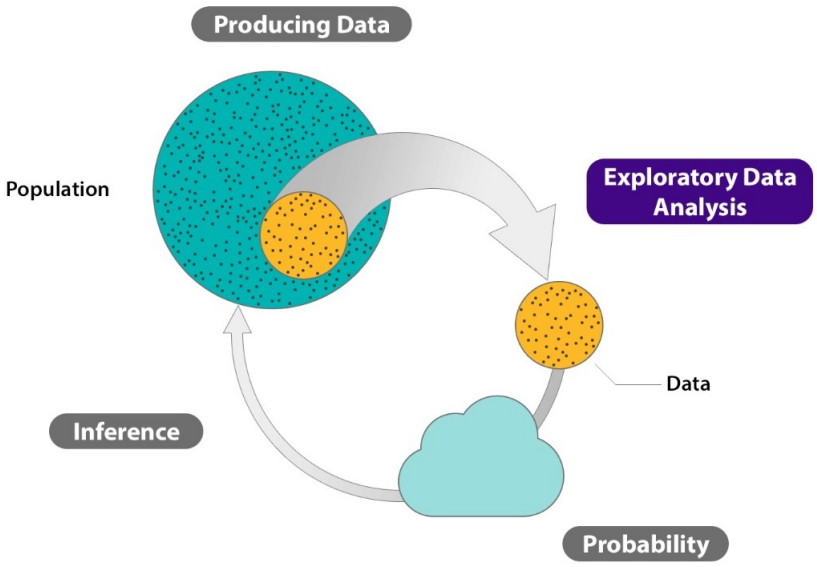
در این تمرین قصد داریم آماده سازی و تحلیل داده را با استفاده از مجموعه دادگان مسافران کشتی تایتانیک تمرین کنیم. تایتانیک کشتی بزرگی بود که در سال ۱۹۱۲ پس از برخورد با یک کوه یخی غرق شد. به دلیل نبود قایق نجات به تعداد کافی، ۱۵۰۲ نفر از ۲۲۲۴ نفر افراد سوار کشتی غرق شدند. در این مجموعه دادگان مشخصات اشخاص حاضر در کشتی (مانند سن، جنسیت،...) ثبت شده است. هدف در این تمرین این است که به مدلی برسیم تا بتوانیم شانس زنده ماندن افراد را بر اساس خصوصیات آن‌ها حدس بزنیم.
در انتهای این تمرین باید گزارش کار کاملی از تحلیل های انجام شده و پاسخ به سوالات و نمودارهای رسم شده را در قالب یک فایل PDF تحویل دهید. همچنین نوتبوک شامل تمام کدهای زده شده و فعالیت‌های خود را آپلود کنید.

○ پیش از شروع : آشنایی با داده ها

داده ها را می توان به روش های متنوعی ذخیره کرد و انتقال داد. در تمرین های قبل استفاده از دادگان جدا شده با کاما (Comma Separated Values) یا CSV را دیدید. فایل های CSV از رایج ترین و ساده ترین شیوه های ذخیره دادگان در دنیای یادگیری ماشین می باشند. در این تمرین نیز از همین فرمت استفاده خواهیم کرد.

داده ها به فرمت بردارهای به صورت خطوط جدا شده اند که نماینده ی سطرهای داده هستند. ستون های داده هم با کاما از یک دیگر جدا شده اند. برای کار با این داده ها، کتابخانه Pandas در پایتون پیشنهاد می شود که در تمرین های قبل با نحوه کار با آن آشنا شدید. برای مطالعه بیشتر درباره توابع این کتابخانه، به مستندات کتابخانه Pandas مراجعه کنید.
مجموعه دادگان در دو فایل train.csv و test.csv در اختیار شما قرار گرفته است. یک رویکرد رایج در مسایل یادگیری ماشین تقسیم دیتا به دوسته train و test است تا مدل روی داده های train آموزش داده شده و عملکرد آن روی داده test امتحان شود. همچنین بسیاری اوقات از دسته سومی از داده ها تحت عنوان Cross Validation set استفاده میشود. در مورد فلسفه و روابط ریاضی این سه دسته در ساختن یک مدل یادگیری ماشین در درس پیشتر آشنا شدید.
دادگان موجود دارای ۱۲ ستون میباشد که ۱۲ ویژگی هر یک از مسافرین است:

- Passenger Id شماره مسافر - این ستون فقط Index هر سطر از دیتا میباشد.
- Survived - زنده ماندن پس از حادثه : عدد یک به معنای زنده ماندن می باشد. این ستون در حقیقت جوابی است که مدل ما قرار است بتواند آن را از روی بقیه ی ستون ها پیش بینی کند. توجه کنید که این ستون از دیتای تست حذف شده است.
- Pclass: کلاس بلیط



- Name: نام کامل مسافر (در یک رشته)
- Sex: جنسیت
- Age: سن
- SibSp: جمع تعداد همسر و برادر و خواهر مسافر حاضر در کشتی
- Parch: جمع تعداد والدین و فرزندان مسافر حاضر در کشتی
- Ticket: شماره بلیط
- Fare: قیمت بلیط
- Cabin: نام کابین
- Embarked: نام بندر محل سوار شدن

- مرحله اول : پیش پردازش داده‌ها

آشنایی با انواع داده ها

داده ها به صورت متنوعی گزارش می شوند. قبل از کار با یک داده باید نوع آن را فهمید و روش درست را به کار گرفت. در این جا فهرست مختصری از این داده ها آمده است:

- داده های عددی یا **Numerical**

این داده ها در واقعیت عدد هستند و مقادیر احتمالا پیوسته ای را نشان می دهند. چیزی که این نوع داده را مشخص می کند امکان عملیات های ساده ی ریاضی روی آن است به نحوی که هم چنان معنا داشته باشند. از این عملیات ها می توان به جمع و ضرب اسکالر اشاره کرد. برای مثال شدت رنگ یک کمیت عددی است چرا که می توان دو شدت رنگ را با یک دیگر جمع کرد و باز شدت رنگ به دست آورد و یا ده برابر شدت بیش تر یک رنگ معنی دارد. اما خود رنگ کمیت عددی نیست و مثلا ده زرد معنی رنگ را با خود حمل نمی کند.کمیت های عددی دیتاست موجود را برای خودتان جدا کنید.

- داده های دسته ای یا **Categorical**

به داده هایی گفته می شود که برچسب آن ها به جای یک عدد، یک طبقه یا دسته است مانند جنسیت یا حالت مو. بر خلاف قد که با یک عدد بیان می شود، حالت مو با دسته هایی مانند فر، صاف و غیره مشخص می شود. این داده ها به صورت مجموعه ای متناهی از حالت ها ذخیره شده اند. برخی از الگوریتم های یادگیری ماشین مانند درخت تصمیم گیری (Decision Tree) می توانند مستقیم با داده های دسته ای کار کنند ولی بیشتر الگوریتم ها فقط با اعداد کار می کنند. بنابراین برای استفاده درست از محتوای این دسته ها باید آن ها را به نحوی به اعداد تبدیل کرد.
البته بسیاری مواقع این حالت ها با اعداد نماینده آن ها نشان داده میشوند. اگر اعداد نسبت داده شده معنای ترتیبی خود را حفظ کرده باشند، این اعداد ممکن است صورت مناسب نمایان گر داده باشند. مثلا اگر مدرک تحصیلی افراد را به ترتیب از دبستان با صفر شماره دهی کنیم، شماره های این داده ی فهرستی، معنی ترتیبی خود را حفظ کرده اند. یعنی وقتی ۳ بزرگ تر از یک است یعنی کسی که مدرک کاردانی دارد از کسی که مدرک سیکل دارد دانش بیش تری دارد که خب این صحیح است. ولی در مورد مثال مو این گونه نیست. شما نمی توانید ترتیبی ذاتی بین حالت های فر، صاف و غیره پیدا کنید. در این موارد باید احتیاط بیش تری در تبدیل داده به خرج داد که یک روش ساده One-hot-encoding است. در این روش هر حالت به صورت یک ستون جدا معرفی می شود و مقادیر صفر و یکی داده می شود. این کار دو مزیت دارد. اول این که فرض ترتیبی که در حالت اعداد صحیح وجود دارد را از بین می برد و دوم داده شکل شبه پراکنده پیدا می کند که ذخیره و پردازش آن به مراتب ساده تر است. اما توجه کنید که این روش در برخی حالات ممکن است تعداد ستون ها را خیلی زیاد کند. چرا که برخی دسته های کمیاب (rare categories) تبدیل به ستون های جدیدی میشوند که ممکن است مفید نباشند.

- داده های تنک یا **Sparse**

این نوع داده به صورت ذاتی به طور متفاوتی ذخیره شده است. در واقع مجموعه ای از مشخصه های وابسته به یک دیگر، به صورت یک ماتریس نمایش داده شده اند که تعداد زیادی از خانه های این ماتریس مقدار صفر دارد. برخی از این خانه ها فقط مقدار ناصفر دارند که تعداد آن ها ناچیز است. در این تمرین داده ی تنک نداریم.

شروع کار با دادگان

۱. اگر وضعیت داده های هر ستون را بررسی کنید، می بینید که برخی از داده ها ثبت نشده یا گم شده هستند. این به معنی است که این بخش از داده ها جمع آوری نشده، نامعتبر بوده یا از دست رفته است. در هر صورت شما هیچ اطلاعی از مقدار واقعی آن ندارید و بنابراین ارزش اطلاعاتی خود را از دست داده است. سطرها یا ستون هایی که دارای تعداد زیادی از N/A هستند، فرآیندهای یادگیری را دچار اختلال می کنند و باید برای آن ها چاره ای اندیشید .
در ابتدا می خواهیم مشخص کنیم که کدام ویژگی ها دارای خانه خالی هستند (NaN). این ویژگی ها را مشخص کرده و تعداد خانه‌های خالی هریک را بدست آورید. کدام ویژگی دارای درصد زیادی خانه‌ی خالی می‌باشد؟

۲. ویژگی name را در نظر بگیرید. در این ستون اسم افراد با title های مختلفی آغاز شده است (مانند Mr، Ms و ...). در این بخش هدف جدا کردن title اسم افراد است. به این منظور ویژگی title را ایجاد کنید و title اسم هر فرد را در این ستون بنویسید. ویژگی name را در نظر بگیرید. در این ستون اسم افراد با title های مختلفی آغاز شده است (مانند Mr، Ms و ...). در این بخش هدف جدا کردن title اسم افراد است. به این منظور ویژگی title را ایجاد کنید و title اسم هر فرد را در این ستون بنویسید.

۳. تعداد تکرار title های مختلف را نمایش دهید (برای مثال چند بار Mr تکرار شده است). با توجه به تنوع تکرار title ها، می‌خواهیم آن ها را به ۳ دسته‌ی کلی تقسیم نماییم. چه پیشنهادی برای این ۳ دسته دارید؟ (راهنمایی: می‌توانید دسته‌ای به نام others ایجاد کنید و title های کم تکرار را در این دسته قرار دهید). این دسته بندی را به ستون title اعمال نمایید.

۴. برای این که بتوان از یک ویژگی جهت پیشبینی استفاده کرد، نیاز است تا تدبیری برای خانه‌های خالی اندیشیده شود. توجه کنید که کیفیت یادگیری شما می تواند به این پارامتر هم وابسته باشد. به این منظور راه‌های مختلفی از جمله:
   - حذف ویژگی‌های دارای خانه‌ی نامشخص
   - حذف سطرهای دارای خانه‌ی نامشخص
   - استفاده از سطر قبلی یا بعدی برای تعیین خانه‌های نامشخص
   - پر کردن خانه‌های نامشخص یک ویژگی با مقدار ثابت (میانگین، میانه و ...)

استفاده کرد. فقط حتما توجه کنید که برای این کار باید توجیه داشته باشید. انجام هر عملیات بدون توجیه بر روی داده می تواند اثرات مخربی روی دقت یادگیری داشته باشد . می‌خواهیم ستون age را که دارای خانه‌های خالی است با استفاده از ستون title پر کنیم. به این منظور، میانگین سن افراد دارای title یکسان را محاسبه نموده و در پر کردن خانه‌های نامشخص ستون age استفاده نمایید.

۵. برای پر کردن خانه‌های نامشخص ستون fare، می‌توان از ویژگی pclass استفاده کرد، زیرا این متغییر تا حدودی موقعیت اقتصادی فرد را مشخص می‌نماید. برای این کار boxplot مربوط به fare برای هر pclass را رسم کنید. از مقدار میانگین fare های هر pclass برای پر کردن خانه‌های خالی استفاده نمایید.

۶. برای استفاده از ویژگی هایی categorical مانند embarked، نیاز است تا آن‌ها را کمی کنیم. هریک از این ویژگی‌ها چند category مختلف دارند؟ برای مثال در کمی کردن ویژگی embarked که ۴ مقدار به خود میگیرد، از دو ستون ویژگی جدید که هرکدام دارای عناصر ۰ و ۱ هستند استفاده کنید. اگر برای کمی کردن این ویژگی از اعداد ۰،۱،۲،۳ استفاده می‌کردید چه مشکلی پیش می‌آمد؟ ویژگی های categorical را با اضافه کردن ستون های مناسب کمی نمایید.

۷. ویژگی‌هایی را که کمی کرده‌اید و دیگر به آنها نیازی ندارید را از جدول حذف نمایید.

- مرحله دوم : تحلیل آماری داده‌ها

آشنایی با تحلیل آماری

در این بخش قرار است ابتدا با برخی آماره ها آشنا شویم و با رسم تعدادی نمودار با ماهیت داده بیش تر آشنا شویم. همان طور که در بخش قبل نیز دیدید یک راه خوب برای پیدا کردن درک درست از ویژگی ها و روابط بین آن ها استفاده از نمودارها است. عملیاتی که در این بخش یاد می گیرید در مورد هر داده ای می تواند مفید باشد.

آماره های مهم:

آماره ها در واقع توابعی هستند از فضای آماری داده ها به اعداد حقیقی که پارامتری از داده را محاسبه می کنند. در اولین برخورد با یک داده باید آماره های متنوع ولی ساده ای را در مورد آن به دست آوریم و تحلیل کنیم. مهم ترین آماره ها میانگین، واریانس، مد و میانه هستند. به عنوان تمرین، آماره ها فوق را برای ویژگی های Age,Sex(F or M),SibSp,Parch یک بار به طور کلی و یک بار برای دو دسته ی Survived 0,1 محاسبه کرده و نتایج حاصله را تحلیل کنید. منظور از تحلیل این است که برای مثال بیان کنید میانگین ویژگی افراد زنده مانده چه تفاوتی با افراد غرق شده دارد یا چه گزاره هایی در مورد شانس زنده ماندن درست است؟

نمودارهای مهم:

با وجود این که آماره ها، اعداد بسیار مهمی در کار با داده هستند اما آن ها روح زنده ی نمودارها را با خود ندارند. در این بخش قرار است با رسم نمودارهای مختلف داده را ارزیابی کنیم. نمودارهای نام برده در هر بند را برای داده ی خواسته شده رسم کنید و نتایج آن را تحلیل کنید. در تمرین های قبل کار با کتابخانه ی matplotlib برای رسم نمودار را دیدید. کتابخانه ی seaborn نیز می‌تواند مفید باشد. برای آشنایی با نحوه کار آن به مستندات این کتابخانه مراجعه کنید. البته تا زمانی که نمودارهای رسم شده صحیح باشد شما مقید به استفاده از کتاب‌خانه ی خاصی نیستید.

۱. با استفاده از ویژگی title که در بخش‌های قبل بدست آوردید نمودار میله‌ای افراد زنده مانده و نمانده با هر یک از عنوان ها(title) را رسم کنید.

۲. نمودار جعبه‌ای (Boxplot) :این نمودار بر حسب مفاهیم میانه و چارک کشیده می‌شود. در نمودار جعبه‌ای چارک‌های اول و سوم و میانه به همراه داده‌ی کمینه و بیشینه رسم می شود. این نمودار شهود بسیار خوبی در مورد پراکندگی مقادیر داده می دهد. نمودار جعبه ای مقدار Fare را برای ۳ گروه Pclass و برای دو حالت Survived 0,1 (یعنی در مجموع ۶ نمودار) رسم کنید. چه نتیجه ای میتوان گرفت؟

۳. نمودار هیستوگرام (Histogram): با این نمودار میتوان توزیع داده های عددی را در تعداد دسته بندی دلخواه مشاهده کرد. هیستوگرام age را رسم کنید و توزیع جمعیت مسافران را مشاهده نمایید (برای سن از بازه های مناسب استفاده نمایید). یک بار نیز این کار را برای دو دسته ی زنده ماده و نمانده انجام ندهید. حال این نمودارها را با جداسازی جنسیت بکشید تا چهار نمودار داشته باشید. مشاهدات خود را ثبت کنید.

۴. نمودار Heatmap یک نمایش گرافیکی دو بعدی از دادگان است که ارتباط ویژگی های مختلف را نشان می دهد. علاوه بر ارتباط بین هر ویژگی با زنده ماندن، بین خود ویژگی ها نیز ممکن است ارتباط معناداری باشد. برای دریافتن این ارتباطات میتوان همبستگی یا Correlation بین ستون ها را مقایسه کرد. میتوان برای مقایسه جامع از Heatmap استفاده کرد و همبستگی بین ستون ها را نشان داد. Correlation Heat map ویژگی‌ها و همچنین survival را رسم نمایید. کدام ویژگی‌ها دارای همبستگی زیادی هستند؟ آیا می‌توان یکی از این ویژگی‌ها را حذف کرد یا خیر؟

- مرحله سوم : استخراج ویژگی‌های جدید

۱. - می‌خواهیم با ترکیب دو ویژگی sex و age، ویژگی جدیدی بسازیم که دارای ۳ دسته‌ی man، woman و child باشد. این ستون جدید ویژگی را ایجاد نمایید.

۲. با ترکیب دو متغییر SibSp و Parch، متغییر جدیدی ایجاد نمایید. این متغییر می‌خواهد تاثیر تعداد همراهان را در نجات یافتن و یا نیافتن فرد نشان دهد. همچنین می‌توانید متغیرهای دیگری مانند مادر بودن، همراه داشتن یا نداشتن و ... را ایجاد نمایید.

۳. در ستون ویژگی cabin، تعداد زیادی خانه‌ی خالی وجود دارد. یک راه پیشنهادی برای این ستون، حذف کامل آن است. اما باید توجه داشت که موقعیت cabin نقش مهمی در نجات یافتن و یا نیافتن افراد داشته است. با توجه به اینکه کشتی از نقطه‌ای خاص شروع به غرق شدن کرده است، افراد آن نقطه احتمالا شانس کمتری برای نجات داشته اند. پس در این بخش هدف پیشبینی کابین افراد است. در این بخش می‌توانید با کمک ویژگی pclass و دیگر ویژگی‌ها، کابین افراد را پیش بینی نمایید.

.

- مرحله چهارم : ادامه‌ی راه

با توجه به مشاهداتی که در تحلیل های آماری داشتید و شهودی که نسبت به دادگان پیدا کرده اید دو ویژگی جدید پیشنهاد کرده و بسازید. با رسم نمودارهای مناسب و بیان آماره ها توضیح دهید که چرا این ویژگی ها باید موثر باشند. تکنیک های بخش های قبل را در این جا به سلیقه‌ی خودتان استفاده کنید. در تمرین بعدی که ساخت مدل و گرفتن جواب است از این ویژگی ها استفاده خواهید کرد.