
بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده علوم ریاضی

عنوان:

پروژه درس یادگیری ماشینی

استاد درس:

دکتر رضا رضازادگان

بهمن ۱۴۰۱

لطفا پیش از شروع پروژه، من را بخوانید!!

سلام! ما اینجا می‌خوایم پروژه انتهایی درس که شامل سه مرحله هست رو با هم بررسی کنیم. فصل اول که پرکارترین فصل هست؛ مربوط به مفاهیم کلیدی مطالبی هست که در طول ترم یاد گرفتید. اکثر نمره هم به این فصل اختصاص داره که احتمالا حدود ۷۵ درصد نمره رو خواهد بود. فصل دوم، مشابه حالت با فصل اول هست؛ اما این بار دیگه مسئله شما باید بدون label در نظر گرفته بشه و در انتها هم، باید خودتون یک سوال خوب (تاکید می‌کنم خوب!) طرح کنید و بهش پاسخ بدید. می‌تونید برای طرح سوال، یک نگاهی به سوالات فصل سوم هم بندازید تا بهتون ایده بده. این فصل، احتمالا ۱۵ درصد از نمرتون رو خواهد داشت. خب تا اینجای کار احتمالا یکم دست به سرچتون قوی شده! اما چون یکی از مهارت‌های اصلی یک مهندس یادگیری ماشین! این هست که بتونه به خوبی سرچ کنه و چیزایی که مد نظر داره رو پیدا کنه؛ فصل سوم رو هم به پروژه اضافه کردیم. این فصل، خیلی ساده و در عین حال جذاب هست و بیشتر مرتبط با مفاهیم آماری یادگیری ماشین هست. در توضیحات خود فصل، دقیقا می‌گیم که باید چی کار کنید اما اگر بخوام به عنوان یک توضیح کلی بگم؛ هدف ما یادگیری یک سری مفاهیم آماری هست که اون هم به سادگی با سرچ کردن به دست میاد و خبر خوب اینکه توی این فصل، فقط نیاز هست که کدها، نتایج و یک سری توضیحات مختصر رو بهمون بگید و چون درس ما آمار نیست؛ قاعدتا ازتون انتظار نداریم که قسمت نظری کدها رو هم یاد بگیرید و به یک سری توضیحات و نتایج کلی بسنده خواهیم کرد. این فصل هم، ۱۰ درصد نمره رو به خودش اختصاص میده. پس اگر بخوایم یک جمع بندی کنیم؛ شما با داشتن اینترنت! و سرچ کردن، به راحتی می‌تونید ۲۵ درصد نمره رو به خودتون اختصاص بدید و تمامی چالش کار، روی فصل اول، یعنی فصلی که مطالبش رو یاد گرفتید؛ خواهد بود. پیشنهاد خود من، این هست که سعی کنید در اسرع وقت، فصل دوم و سوم رو انجام بدید و تمرکز خودتون رو بذارید روی فصل اول! البته این به این معنی نیست که اون فصل‌ها رو سرسری بگیرید؛ بلکه اشاره به راحتی اون‌ها داشتم و قطعا باید چیزی که مد نظر ما هست رو بنویسید تا نمره کاملش، بهتون تعلق بگیره. در مورد مشورت کردن هم خوبه که یک توضیحات کوتاهی بدیم. قطعا اینکه شما از سایت‌ها و کمک دوستانتون استفاده کنید؛ هیچ اشکالی نداره؛ اما باید بگم که اگر متوجه بشیم که کد سایت‌ها رو بدون فهمیدن و عوض کردن، کپی کردید؛ متاسفانه نمره‌ای رو بهتون نمیدیم:) اگر هم دیدیم که کدهاتون رو در اختیار همدیگه قرار دادید و یک نفر بدون فهمیدن از دوستش کپی کرده هم مجبوریم که از هر دو نفر نمره کسر کنیم:(این اتفاق زیاد دیده میشه که بچه‌ها به خاطر دوستی خارج از حیطه درسیشون به هم کمک می‌کنن. اگر شما جزو این دسته از افراد هستید؛ مطمئن بشید

که دوستتون کد رو فهمیده و خدای نکرده قرار نیست کد شما رو کپی کنه چون برای ما فرقی نداره که کی کد رو داده و کی کد رو گرفته و مجبوریم از هر دو نفر نمره کم کنیم. مورد دیگه اینکه من توقعاتم از هر بخش رو تا حدی عنوان می‌کنم تا بدونید که دقیقا باید چه کاری رو انجام بدید. اونجا احتمالا یک سری راهنمایی‌های ریز هم در مورد پروژه خواهم کرد. که انتهای هر متن، باز موارد رو به صورت تیتر وار مینویسم. پس خوندنشون خالی از لطف نیست؛ اما اگر فکر می‌کنید که به اندازه کافی مسلط هستید؛ می‌تونید از خوندن اون‌ها صرف نظر کنید. اگر موفق به نصب ژوپیتر نشدید؛ در گروه یا با آیدی @nsrmelikaaaa در تلگرام یا از طریق ایمیل mnasirian77@gmail.com با من در ارتباط باشید و یا اگر ابهامی توی سوال‌ها بود؛ با من در گروه در میون بگذارید. در ادامه هم یک سری از مواردی رو میگم که قبلا هم توی گروه گفتیم و برای تاکید بیشتر و محکم کاری! دوباره می‌گیم. زبان برنامه نویسی شما حتما باید به زبان پایتون باشه؛ قرار نیست پروژه‌ها به راحتی کوییزها تصحیح بشن و نسبت بهشون، سخت‌گیر خواهیم بود؛ در انتهای زمان تحویل پروژه‌ها، طی دو روز، یک سری زمان مشخص می‌کنیم و شما باید کارهایی که کردید رو برای ما توضیح بدید. در حین این توضیحات، ما یک سری سوال که لزوما از کدی که زدید نیست رو از شما می‌پرسیم تا اطمینان حاصل کنیم که کدها رو خودتون زدید و خب اگر قانع نشدیم؛ متاسفانه نمره‌ای بهتون تعلق نخواهد گرفت؛ پروژه تحویلی باید به صورت یک jupyter notebook باشه و از سلول‌های markdown و کامنت‌ها، جهت توضیح کدها استفاده کنید.

خب دیگه فکر می‌کنم وقت اون رسیده که بریم سراغ سوالا!

فصل ۱

بررسی مدل‌ها در حالت Supervised

برای شروع کار، ازتون می‌خوام که یک سری توضیحات در مورد ستون‌های داده‌هاتون بنویسید. یعنی هدف این هست که اصلا بدونید دیتا در مورد چیه! اینکه اسم ستون‌هارو بنویسید و یک توضیح یکی دو خطی هم بنویسید کافیه. همونطور که می‌دونید برای انجام این قسمت، باید از حالت markdown استفاده کنید.

* نوشتن توضیحات مختصر در مورد ستون‌ها

۱-۱ تمیز کردن داده‌ها

خب قاعدتا طبق مطالبی که سر کلاس‌های حل تمرین، با هم بررسی کردیم؛ توقع ما از شما این هست که بتونید دیتایی که بهتون می‌دیم رو تمیز کنید. یعنی چی؟ یعنی اینکه اگر داده‌ای دارید که غیرعددی هست باید به عدد تبدیل بشه. اگر سلولی توی دیتاتون خالی هست! باید یک فکری براش بکنید! این موضوع که چطوری بررسی کنیم یا چی کار کنیم رو توی کلاس حل تمرین با هم بررسی کردیم. یک مورد دیگه هم حذف ستون‌های بی ربط هست که نباید به اشتباه بعضی از ستون ها رو حذف کنید و باید تشخیصتون درست باشه و از این قبیل کارهای مشابه!! من سعی کردم به یک تعدادیشون اشاره کنم که یادتون بیاد منظورمون از تمیز کردن داده چی هست.

* تبدیل داده غیرعددی به عددی

* پر کردن جای خالی داده‌های گم‌شده با عدد مناسب!

*حذف ستون‌های بی‌ربط و ...

۱-۲ بررسی کوریلیشن بین ستون‌ها

این دستور رو هم که به همراه رسم شکلش توی کلاس بررسی کردیم و خب عینا می‌خوایم که شما هم این کار رو انجام بدید:)) (سوال نمره بیار!)

* به دست آوردن کوریلیشن و رسم نمودار آن

۱-۳ بصری‌سازی

می‌دونید که با پایتون میشه نمودارهای خیلی جالبی رو کشید. خیلی روی این بخش سخت‌گیری نداریم؛ اما اگر دو سه تا نمودار بکشید که ارتباط ستون‌هاتون رو به خوبی نشون بده؛ نظر ما رو هم به توانایی‌هاتون جلب خواهید کرد:)

*رسم یک یا دو نمودار جالب

۱-۴ برازش مدل‌ها و بررسی دقت آن‌ها

خب توی این درس، مدل‌های خیلی زیادی رو با هم بررسی کردیم. قبل از برازش مدل، باید بتونید تشخیص بدید که مسئله شما regression هست یا classification ! اگر در این دو مورد اشتباه کنید؛ متاسفانه کل پروژه شما غلط خواهد بود! هم مقیاس کردن داده‌ها هم فراموش نشه (طبیعتا اگر لازمه و خب اگر استفاده می‌کنید یا نمی‌کنید باید دلیل بیارید که باز این رو هم قبلا بررسی کردیم:)) ازتون میخوایم که به دلخواه، سه مدل رو با توجه به مسئله خودتون و با استفاده از کتابخونه‌ای که خودتون میدونید چیه:) (scikit learn) برازش کنید. فقط علاوه بر این‌ها، چند تا چیز دیگه هم می‌خوام از شما. توی این سه مدل، حتما یکی از مدل‌ها باید ویژگی که در ادامه می‌گم رو داشته باشه.باید توی برازش یکی از مدل‌ها، نیاز به هایپر پارامتر باشه. که در واقع اون عدد رو نباید به صورت تصادفی انتخاب کنید

و باید هایپر پارامتر بهینه رو پیدا کنید که کار راحتیه! به عنوان یک راهنمایی و شاید هم جواب عینی سوال :))، به جملات بعدی من دقت کنید. مثلا توی مدل KNN، شما نیاز دارید که تعداد همسایگی‌ها رو مشخص کنید. یک روش این هست که با استفاده از حلقه‌ها، یک بازه برای این تعداد مشخص کنید و تعداد همسایگی بهینه رو پیدا کنید. تعداد همسایگی بهینه، قاعدتا با مقایسه معیار مورد بررسی شما به دست میاد. معیار شما میتونه بیشترین دقت یا کمترین MSE یا معیارهای دیگه باشه که توی جلسه کارگاهی که توی Naive Bayes داشتیم؛ حوالی توضیحات confusion matrix می‌تونید اون‌ها رو پیدا کنید. البته که برای هر مسئله‌ای یک معیار، بهینه هست اما ما از این موضوع صرف نظر می‌کنیم و شما هر معیاری رو بررسی کنید؛ قبول خواهیم کرد! یا مثلا اگر از مدل SVM استفاده می‌کنید؛ نیاز دارید که یک کرنل مشخص کنید که آوردن یک کرنل از هوا:))) باعث کسر نمره میشه. چندتا کرنل بیشتر نیست؛ شما می‌تونید با هر کرنل، معیارتون رو بسنجید و در نهایت با مقایسه اون، کرنل بهینه رو به ما معرفی کنید. دقت کنید که بیست درصد داده‌ها رو به عنوان داده تست انتخاب کنید و دانه رو هم برابر با ۱۲۳۴ قرار بدید. چون این بخش رو خیلی راهنمایی کردم؛ قطعا اگر مطابق با اون عمل نشه؛ نمره زیادی کسر خواهد شد:) فقط چون انتخاب مدل‌ها به عهده خودتون هست؛ ما یک نمره امتیازی برای کسایی در نظر خواهیم گرفت که به سراغ مدل‌های پیچیده‌تر خواهند رفت.

* انتخاب ۲۰ درصد از داده‌ها به عنوان داده‌های تست

*انتخاب مقدار ۱۲۳۴ برای دانه

* انتخاب سه مدل دلخواه(که یکی از آن‌ها باید دارای هایپر پارامتر باشه)، برازش آن‌ها با استفاده از کتابخانه‌ها و به دست آوردن مقدار عددی معیار مورد بررسی

*به دست آوردن مقدار بهینه هایپرپارامتر

*امتیازی بودن انتخاب مدل‌های سخت‌تر

۱-۵ پیاده‌سازی انیمیشن و الگوریتم

در مورد این بخش نمی‌تونم راهنمایی زیادی کنم. اما از بین اون مدل‌هایی که در بخش قبلی برازش کردید و مثلا دقت اون‌ها رو بررسی کردید؛ الگوریتم اونی که نسبت به بقیه بهینه هست رو بزنید (این یعنی شما این بار، مجاز به استفاده از کتابخونه نیستید). در مورد انیمیشن هم به لینک استاد در گروه

مراجعه کنید و سوالی در موردش پاسخ داده نمیشه:)). اجرای انیمیشن هم چون پیچیدگی‌های مختص به خودش رو داره؛ امتیازی در نظر گرفته شده اما چون سطح اون در مقایسه با سوال‌های امتیازی دیگه، خیلی بالاتر هست؛ نمره امتیازی اون هم خیلی خیلی بیشتر خواهد بود:).

* پیاده‌سازی الگوریتم مدل بهینه از میان مدل‌های انتخابی شما در بخش ۱-۴

* امتیازی بودن اجرای الگوریتم: در صورت نیاز برای ساختن انیمیشن در matplotlib می‌تونید مثال موجود توی لینک زیر رو ببینید.

https://www.geeksforgeeks.org/using-matplotlib-for-animations/amp/

۱-۶ روش‌های کاهش بعد و بررسی مدل‌ها

توی این بخش ازتون می‌خوایم که از feature selection با استفاده از روش‌های backward و forward استفاده کنید و نتایج اون‌ها رو با هم مقایسه کنید. به عنوان یک بخش امتیازی و اضافی هم خوبه که روشی که به صورت ترکیبی هست رو تست کنید. (طبیعتا توی این بخش، استفاده از کتابخونه‌ها آزاده!) یعنی شما باید از بین ستون‌هاتون با استفاده از این سه روش، یک تعدادی ستون رو مشخص کنید و بعد از اون، یکی از مدل‌هایی که توی بخش ۱-۴ برازش کردید رو مجددا با متغیرهای پیشگوی جدید برازش کنید و با حالت قبلی مقایسه کنید.

* انتخاب متغیرها با استفاده از روش‌های backward و forward

*برازش یکی از مدل‌های انتخابی در بخش ۱-۴ و مقایسه معیار مورد بررسی با حالت ابتدایی

۱-۷ بررسی بهینگی پارامتر k در روش k-fold crossvalidation

می‌دونیم که این روش برای جداسازی داده های تست و آموزشی هست ولی اینکه ما چه نسبتی از داده‌ها رو به تست و یا آموزش نسبت بدیم؛ یک چالش هست که معمولا میگن که اگر $k = ۱۰$ باشه؛ حالت بهینه هست. خوبه که با استفاده از حلقه‌ها و مقایسه معیار مورد بررسیتون، مقدار بهینه k رو به دست بیارید. برای بررسی این بخش هم استفاده از هر مدلی که دوست دارید؛ مجاز هست.

* پیدا کردن مقدار بهینه برای پارامتر k در روش k-fold crossvalidation

۱-۸ استفاده از bootstrap

توی این بخش ازتون می‌خوام که فقط ۲۰ درصد از داده هاتون رو نگه دارید و مابقی رو حذف کنید. انتخاب این بیست درصد رو به صورت تصادفی انجام بدید. بعد از اون با استفاده از روش bootstrap، داده تولید کنید و باز هم معیارتون رو با یکی از مدل‌های دلخواهی که قبلا با استفاده از داده‌های کامل، تست کردید؛ مقایسه کنید و ببینید که چه تغییری توی معیار مورد بررسیتون رخ داده.

* حذف ۸۰ درصد از داده‌ها و تولید داده با استفاده از bootstrap

*برازش یک مدل دلخواه و بررسی معیار مورد بررسی با حالت ابتدایی

۱-۹ استفاده از روش‌های انقباضی

این بخش، فقط مربوط به کسایی میشه که مسئله اون ها regression هست. پس اگر مسئله شما classification هست؛ این قسمت رو رد بشید و در نظر نگیرید. ازتون می‌خوایم که هر دو روش ridge regression و lasso regression رو روی یک مدل پیاده‌سازی کنید و بگید که کدوم یکی بهتره (با دلیل!).

*مقایسه روش‌های ridge regression و lasso بر یک مدل دلخواه

۱-۱۰ بررسی bias و variance

این بخش، امتیازی هست و احتمالا برای کسایی که تاحالا دست به کد نشدن؛ یکم زمان بره. برای همین، پیشنهاد من به شما اینه که اگر زیاد مسلط نیستید؛ این بخش رو بذارید کنار و اگر وقت اضافه آوردید؛ انجام بدید. از لحاظ محتوایی اصلا پیچیده نیست اما ممکنه توی پیاده‌سازی، یکم چالش داشته باشید. اما اگر یکم تلاش کنید؛ احتمالا بتونید ۲ یا نهایتا ۳ ساعته، به هدف مد نظر ما برسید. هدف این بخش، اینه که با استفاده از یکی از مدل‌های Ensemble Learning، $bias(\hat{y})^2$ و $var(y, \hat{y})$ رو نشون بدید. قاعدتا منظور از y، مقادیر ستون label ها خواهد بود.

* بررسی $bias(\hat{y})^2$ و $var(y, \hat{y})$ با استفاده از شبیه‌سازی یکی از مدل‌های Ensemble Learning

فصل ۲

بررسی مدل‌ها در حالت UnSupervised

۲-۱ روش PCA و رسم آن

خب همونطور که می‌دونید؛ یکی از روش‌های کاهش بعد، PCA هست. ازتون می‌خوایم که این روش رو روی یکی از مدل‌هایی که توی بخش ۱-۴ برازش کردید؛ در دو بعد و سه بعد پیاده‌سازی کنید و بعد از اون، شکلش رو برای ما رسم کنید:).

* پیاده‌سازی روش PCA و رسم شکل آن در دو یا سه بعد

۲-۲ clustering

این بخش، فقط مربوط به کسایی میشه که مسئله اون‌ها classification هست. پس اگر مسئله شما regression هست؛ این قسمت رو رد بشید و در نظر نگیرید. خب همونطور که می‌دونید یکی از ستون‌های شما، اسمش label هست و شما دسته‌بندیتون رو طبق اون انجام دادید. ازتون می‌خوام که اون ستون رو کلا حذف کنید و فرض کنید که دیگه label ندارید. حالا از یکی از مدل‌های clustering استفاده کنید و در نهایت مشخص کنید که نسبت به یکی از مدل‌ها توی حالتی که مسئله شما، label داره؛ کدوم یکی بهتر هستند و چرا؟ (البته همونطور که می‌دونید؛ این کار، بهینه نیست؛ چون دارید مدل‌های انتخابیتون رو دلخواه در نظر می‌گیرید و حالت‌های بهینه هر دو روش رو با هم مقایسه نمی‌کنید. اما از

این موضوع صرف نظر می‌کنیم و همین که شما یک مدل از clustering رو برای ما برازش و مقایسه کنید؛ کافی خواهد بود.)

*حذف ستون labelها

* بررسی یکی از مدل‌های clustering و مقایسه اون با یک مدل از classification

۲-۳ طرح یک سوال و پاسخ به آن

خب این بخشم واضحه دیگه! یک سوال خوب طرح کنید و بهش پاسخ بدید. اگر سوال رو نپسندیدیم؛ نمره داده نمیشه:)

.

فصل ۳

مفاهیم آماری

خب رسیدیم به آسون‌ترین و نمره آورترین فصل پروژه! احتمالا اگر از قبل، پیشینه آماری نداشته باشید؛ مواری که ازتون خواسته شده رو برای بار اوله که می‌بینید و شما توی این قسمت، باید قدرت سرچ خودتون رو به ما نشون بدید!

۳-۱ بررسی نرمال بودن داده‌های یک ستون با استفاده از تست‌های آماری و نمودار QQ-plot

توی این بخش، یکی از ستون‌های دادتون رو به دلخواه انتخاب می‌کنید؛ بعد از اون میاید و با استفاده از یکی از تست‌های shapiro-wilk و d-agostino k-squared و anderson-darlinf و با استفاده از نمودار QQ-plot نشون می‌دید که داده‌های اون ستون از توزیع نرمال میان یا نه. هر کدوم از موارد بالا با سرچ در گوگل، به راحتی قابل دستیابی هستن و هدف ما از آوردن این بخش‌ها، یادگیری شما بوده. هر قسمت چندین خط کد بیشتر نداره و آوردن همون‌ها کافیه. فقط ازتون می‌خوام که در مورد هر کدوم از تست‌های انتخابی و نمودار مذکور، دو جمله‌ای توضیح بنویسید و اون چند خط کد رو هم فهمیده باشید!

* رسم qq-plot
* بررسی یکی از تست های آماری جهت تعیین نرمال بودن داده‌ها

سه بخش بعدی امتیازی هستند!

۳-۲ بررسی استقلال داده‌های دو ستون

برای این کار، دو تا ستون رو در نظر بگیرید و از تست CHI-SQUARE استفاده کنید. توی این بخش هم، آوردن چند خط کد و چند جمله‌ای توضیح در مورد این تست، کافیه!

۳-۳ بررسی توزیع داده‌های یک ستون

این بخش رو هم مشابه با بخش‌های قبلی سرچ کنید و از نتیجه لذت ببرید!:)))) ستون انتخابی هم، دست خودتون هست و هر ستونی که می‌خواید رو می‌تونید بررسی کنید.

۳-۴ بررسی رابطه خطی بین دو ستون با استفاده از پارامترهای آماری

این بخش امتیازی هست ولی قبل از اینکه ازش رد بشید! به شدت پیشنهاد می‌کنم که برید دنبالش چون باز هم شما با سرچ عباراتی که میگم می‌تونید به راحتی به پاسخ این سوال، دسترسی پیدا کنید. چندین خط کد هست که منجر به کشیدن یک جدولی میشه که یک سری مفاهیم آماری مثل T-statistics و F-statistics و p-value توی اون نشون داده شده. یک سری موارد دیگه هم هست که کاری به اون‌ها نخواهیم داشت. این ها یک سری مفاهیم آماری هستند که شما می‌تونید با بررسی اینکه اون‌ها از چه میزانی کمتر یا بیشتر هستند؛ نشون بدید که رابطه خطی بین دو تا از ستون‌های شما برقرار هست یا نه! دقت کنید که این دو ستون لزوما نباید از ستون پیشگو و ستون هدف شما انتخاب بشه و دوتاش می‌تونه از ستون‌های پیشگوی شما باشه. مورد جالبیه و نمره بیار!

* بررسی مقادیر T-statistics و F-statistics و p-value و بررسی خطی بودن یا نبودن رابطه بین دو ستون

امیدوارم که موفق باشید:))