به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

پروژه پایانی

خرداد ماه ۱۴۰۲

*فهرست

مقدمه

در صنعت بانکداری روزانه داده‌های زیادی از طریق تراکنش‌ها و فرآیندهای بانکی تولید می‌شود و همین موضوع، آن را به یک کاندید مناسب برای تحلیل‌های داده‌کاوی تبدیل کرده است. در این پروژه، روی یک مجموعه داده که به صورت عمومی در اینترنت وجود دارد، کار می‌کنیم که شامل اطلاعاتی درباره کمپین‌های بازاریابی یک مؤسسه بانکداری پرتغالی است. این مجموعه داده شامل ستون‌هایی مثل سن، شغل، وضعیت تأهل و ... است و ستون هدف، نشان می‌دهد که آیا مشتری در سپرده مدت‌دار اشتراک گرفته است یا خیر. در این پروژه شما با مشکلاتی که ممکن است در طی فرآیند طبقه‌بندی ایجاد شود(مانند زیاد بودن تعداد ویژگی ها) مواجه شده و برای حل آن‌ها اقدام می‌کنید.

شما در این پروژه، ابتدا یک روش مناسب برای آماده‌سازی ویژگی‌ها انتخاب می‌کنید، سپس یک مدل مناسب برای این مسأله ارائه می‌دهید و نتایج بدست آمده را بررسی می‌کنید. آنالیزهای این‌چنینی می‌توانند به بانک‌ها در بهبود کمپین‌های بازاریابی کمک شایانی کنند.

شرح پروژه

در صنعت بانکداری، یکی از روش‌های جذب مشتری، استفاده از کمپین‌های بازاریابی است. میزان کارایی این کمپین‌ها، در صورتی که بدون آنالیز جامعه هدف انجام شود، می‌تواند بسیار پایین باشد. باید در نظر داشت که هزینه‌ی این کمپین‌ها نیز مقادیر ناچیزی نیستند. در این پروژه می‌خواهیم با تحلیل مجموعه داده یاد شده و استفاده از ویژگی‌های آماده شده، پیش‌بینی کنیم که یک مشتری در سپرده مدت‌دار اشتراک می‌گیرد یا خیر. این مسأله را می‌توان را به صورت یک مسأله‌ی طبقه‌بندی دو کلاسه[1] مدل کرد. برای بررسی درست بودن پاسخ نیز از ستون هدف استفاده می‌کنیم: در صورتی که مقدار این ستون صفر باشد، به این معنی است که اشتراک گرفته نشده و ستون یک نیز به این معنی است که اشتراک گرفته شده است.

مجموعه داده شامل اطلاعات مشتریان است. هر ردیف از آن، نشان‌دهنده‌ی یک مشتری خاص در کمپین است.

۲۰ ویژگی[2] در این مجموعه وجود دارد که در ادامه به توضیح آن‌ها می‌پردازیم:

۱- age : سن مشتری

۲- job : نوع شغل مشتری

۳- marital : وضعیت تأهل مشتری

۴- education : بالاترین سطح تحصیل مشتری

۵- default : اینکه مشتری به صورت پیش‌فرض اعتبار دارد

۶- housing : اینکه مشتری وام مسکن دارد

[1] Binary Classification
[2] Feature

۷- loan : اینکه مشتری وام شخصی دارد

۸- contact : روش تماس با مشتری

۹- month : ماه تماس با مشتری

۱۰- day_of_week : روز در هفته تماس با مشتری

۱۱- duration : مدت زمان تماس با مشتری، به ثانیه

۱۲- campaign : تعداد تماس‌های برقرار شده در این کمپین با مشتری

۱۳- pdays : تعداد روزهای سپری شده از آخرین تماس برقرار شده در کمپین قبلی با مشتری

۱۴- previous : تعداد تماس‌های برقرار شده در کمپین قبلی با مشتری

۱۵- poutcome : نتیجه کمپین قبلی

۱۶- emp_var_rate : نرخ تنوع اشتغال - محاسبه شده در هر سه ماه

۱۷- cons_price_idx : شاخص قیمت مشتری - محاسبه شده در هر ماه

۱۸- cons_conf_idx : شاخص اطمینان مشتری - محاسبه شده در هر ماه

۱۹- euribor3m : نرخ Euribor 3 - محاسبه شده به صورت روزانه. این نرخ به صورت کلی به نرخ بهره‌ای که بانک‌ها برای هم به مدت سه ماه در نظر می‌گیرند، اشاره می‌کند.

۲۰- nr_employed : تعداد کارمندان بانک در آخرین تماس با مشتری

این ویژگی‌ها بازه‌ی وسیعی از انواع متغیرها را شامل می‌شوند و یکی از چالش‌های این پروژه، نحوه برخورد با آن‌هاست.

با استفاده از این متغیرها می‌خواهیم مسأله حدس زدن اشتراک یک مشتری در سپرده مدت‌دار را مدل کنیم. توجه داشته باشید که قسمت اعظم نمره این پروژه، تحلیل و توضیح شما درباره هر قدم از حل مسأله است و تمامی نتایج باید تحلیل شوند.

مرحله اول: آماده‌سازی ویژگی‌ها

در ابتدا باید داده‌ها را برای استفاده در یک مدل یادگیری ماشین آماده کنیم. با توجه به اینکه ویژگی‌های داده شده شامل انواع مختلفی از متغیرها هستند، نیاز است برای تبدیل آنها به متغیرهای عددی تمهیداتی اندیشیده شود، زیرا اکثریت مدل‌های مبتنی بر یادگیری ماشین تنها بر روی داده‌های عددی کار می‌کنند. در این مرحله شما وظیفه دارید ویژگی‌های داده شده را برای استفاده در مدل‌های یادگیری ماشین آماده کنید. به علاوه، با توجه به تعداد زیاد ویژگی‌ها، ممکن است استفاده از تمامی آن‌ها در یک مدل یادگیری ماشین به بیش‌برازش[3] منجر بشود. بنابراین، یکی دیگر از وظایف شما در این بخش کاهش ابعاد ویژگی‌ها می‌باشد. پس از بررسی کافی، در این مورد به سوالات زیر پاسخ دهید:

- برای هرکدام از ویژگی‌های لیست شده در بالا شیوه‌ی مناسب تبدیل آن به ویژگی عددی (در صورت نیاز) را شرح دهید.

- با توجه به اینکه ویژگی‌های داده شده در بازه‌های گوناگون قرار دارند، در صورت استفاده از این ویژگی‌ها به صورت خام ممکن است به خاطر تفاوت بزرگی، یکی از این ویژگی‌ها بر سایر ویژگی‌ها غالب شود و تاثیر بیشتری بر خروجی داشته باشد. برای حل این مشکل چه پیشنهادی دارید؟[4]

- مساله‌ی دیگر تعداد زیاد ویژگی‌های پیش‌رو است. برای این مشکل راه‌حل‌های گوناگونی وجود دارد که بهتر است در مورد آن کمی جستجو کنید. به عنوان مثال، می‌توانید تعدادی از ویژگی‌ها را حذف کنید، یا اینکه با استفاده از روش‌های کاهش بعد از ابعاد ویژگی‌ها بکاهید. به علاوه، میتوانید با دسته‌بندی ویژگی‌ها، برای هر دسته یک روش کاهش بعد جداگانه ارائه دهید. با توجه به نتایج تحقیقات خود و شناختی که از داده‌ها دارید کدام روش را پیشنهاد می‌دهید؟

- چالش نهایی، برخورد با داده‌های پرت در مجموعه‌ی داده است. برای این مشکل نیز راه‌حلی ارائه دهید.

دقت کنید پاسخ شما به هر کدام از سوالات بالا باید همراه با توضیحات باشد. روش‌های استفاده شده را مختصرا توضیح دهید و همچنین برای تصمیم‌گیری‌های خود دلایل کافی و منطقی ارائه کنید. همچنین، توجه کنید که الزاما یک پاسخ صحیح برای سوالات بالا وجود ندارد. بنابراین، با جستجو و مطالعه‌ی کافی به سوالات بالا پاسخ دهید، و تمامی دلایلی که به ذهنتان می‌رسد را مطرح کنید.

[3] Overfitting
[4] به مبحث Feature Scaling رجوع کنید

مرحله دوم: یادگیری و انتخاب مدل مناسب

در این مرحله قصد داریم یک مدل مناسب برای یادگیری بیابیم. مدل‌های مدنظر ما در این بخش شامل مدل های زیر است:

- مدل Support Vector Machine
- مدل Multi-layer Perceptron با یک لایه‌ی پنهان[5] و تابع فعال‌سازی Tanh[6]
- مدل Classification Based on Assocation

برای پیاده‌سازی این مدل‌ها می‌توانید از هر کتابخانه‌ای که می‌خواهید، استفاده کنید. پیشنهاد ما، کتابخانه‌ی sklearn و pyarc است. هرکدام از این مدل‌ها تعدادی فراپارامتر[7] دارند که نیاز به یافتن مقدار مناسب برای آن‌ها داریم. در مدل اول باید کرنل مناسب و پارامترهای مناسب آن را بیابید. در مدل دوم باید اندازه‌ی لایه‌ی پنهان را پیدا کنید. در مدل سوم نیز باید مقدار support و confidence مناسب را بیابید. ابتدا در مورد هرکدام از این مدل‌ها و فراپارامترهای ذکر شده تحقیق کرده و نتیجه را گزارش دهید. دقت کنید که با توجه به واریانس موجود در عملکرد مدل دوم به خاطر مقداردهی اولیه و غیر محدب بودن مدل، باید میانگین و انحراف از معیار عملکرد آنها برای چندین مقداردهی اولیه مختلف ذکر شود. در هنگام انتخاب مدل مناسب این تفاوت عملکرد و بزرگی واریانس را در نظر بگیرید.

ابتدا داده‌هایی که در اختیار شما قرار داده شده‌اند را به سه بخش یادگیری[8]، ارزیابی[9] و آزمون[10] طبقه‌بندی کنید. این طبقه‌بندی باید به نسبت ۱:۱:۳ باشد. به علاوه دقت کنید که طبقه‌بندی به صورت Stratified باشد. توجه داشته باشید که داده‌ی ارزیابی، باید برای تعیین فراپارامترهای مطلوب استفاده شود. با توجه به انتخاب‌هایی که در قسمت قبل برای ویژگی‌ها داشتید، برای هرکدام از مدل‌های ذکر شده و برای هر انتخاب فراپارامتر یک مدل یادگیری کرده و عملکرد مدل را روی داده‌های ارزیابی گزارش کنید. معیارهای مورد استفاده برای عملکرد مدل‌ها معیارهای Accuracy و Micro-F1 و Macro-F1 باشد. برای مدل SVM فقط کرنل‌های rbf، Linear، Polynomial و Gaussian را بررسی کنید. برای مدل MLP اندازه‌ی لایه‌ی پنهان ۸، ۱۶، ۳۲ و ۶۴ را بررسی کنید. برای مدل CBA نیز تنها الگوریتم m1 را برای پارامتر algorithm این مدل درنظر بگیرید و برای هر کدام از support و confidence سه مقدار دلخواه را بررسی کنید؛ طوری که نتیجه بهتری حاصل شود. نتیجه‌ی عملکرد مدل‌ها را توجیه کنید. دقت کنید که بسته به انتخاب‌های شما در قسمت قبل ممکن است عملکرد مدل‌ها متفاوت باشد. به همین خاطر تحلیل‌های خود را با توجه به انتخاب‌های خود در قسمت قبل و مطالعاتی که درباره‌ی هرکدام از مدل‌ها در این قسمت انجام دادید بیان کنید. همچنین، عملکرد مدل‌ها طبق معیارهای مختلف میتواند متفاوت باشد، که این مساله نیز در تحلیل‌های شما باید مورد بررسی قرار بگیرد.

---

[5] Hidden Layer
[6] Activation Function
[7] Hyperparameter
[8] Train
[9] Validation
[10] Test

مرحله سوم: یک مدل تحلیل‌پذیر (امتیازی)

علت عملکرد خوب روش‌های مبتنی بر یادگیری عمیق ایجاد ویژگی‌های پیچیده در لایه‌های پنهان شبکه است. یکی از مشکلات روش‌های مبتنی‌بر شبکه‌های عصبی عدم تحلیل‌پذیری عملکرد آن‌ها خصوصا در لایه‌ی پنهان می‌باشد. این به این معنی است که پیاده‌سازی آن‌ها در محیط‌های مخاطره آمیز مانند مساله‌ی پیش روی ما ممکن نیست، زیرا به خاطر این عدم تحلیل‌پذیری امکان بررسی عملکرد مدل در تمام حالات وجود ندارد. به علاوه، این عدم تحلیل پذیری امکان بررسی دقیق اهمیت و تاثیر ویژگی‌ها در حل مساله را نیز از ما خواهد گرفت .

یکی از جایگزین‌های روش‌های مبتنی بر شبکه‌های عصبی استفاده از روش‌های مهندسی ویژگی است. مهندسی ویژگی شامل طراحی ویژگی‌های مراتب بالاتر[11] و غیر خطی از ویژگی‌های خام و سپس انتخاب آن‌ها بر اساس عملکرد می‌باشد. به عنوان مثال، یکی از روش‌های معمول مهندسی ویژگی استفاده از توابع اسکالر با فرم بسته (مانند توابع سینوسی، توابع نمایی، و توابع چندجلمه‌ای) برای انتقال ویژگی‌ها و بررسی عملکرد آن‌ها روی داده‌های ارزیابی است.

در این بخش، شما باید با استفاده از ابزار AutoFeat[12] ویژگی‌های مناسب را برای مساله‌ی پیش رو بیابید. این روش به طور اتوماتیک یک مجموعه ویژگی مناسب مهندسی کرده و به شما می‌دهد. مسائل از این دست که در آن‌ها هدف طراحی اتوماتیک بخشی از پایپ لاین یادگیری ماشین است را AutoML نامیده‌اند.

ابتدا در مورد جزییات عملکرد این روش توضیح بدهید. سپس توضیح دهید چرا استفاده از ابزاری مانند PCA یا Linear Discriminant Analysis در بعضی مسائل نمی‌توانند جایگزین مناسبی برای روش‌های مهندسی ویژگی مبتنی‌بر توابع غیرخطی باشد.

سپس، ابزار AutoFeat را یک مرحله روی ویژگی‌های خام اجرا کرده و ویژگی‌های انتخاب شده را تحلیل کنید. همچنین، با استفاده از یک مدل Logistic Regression و ویژگی‌های مهندسی شده یک مدل جدید یادگیری کرده و عملکرد آن را با مدل بخش قبل مقایسه و تحلیل کنید.

دقت کنید که بخش اصلی نمره‌ی شما در این قسمت به تحلیل‌های شما از مشاهدات تعلق می‌گیرد.

[11] Higher Order

[12] https://arxiv.org/pdf/1901.07329.pdf - https://github.com/cod3licious/autofeat

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_PRJ_StudentID.zip تحویل داده شود.

- این فایل فشرده، باید حاوی یک فایل با فرمت PDF (گزارش تایپ شده) و یک پوشه به نام Codes باشد که کدهای نوشته شده را شامل می‌شود. در صورتی که از Jupyter Notebook استفاده می‌کنید نیازی به ارسال جداگانه کدها و گزارش بخش عملی نیست و هر دو را می‌توانید در یک فایل Notebook قرار دهید. حتما خروجی html فایل Notebook خود را نیز همراه فایل Notebook ارسال کنید.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- گزارش به صورت تایپ شده در قالب PDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد. دقت داشته باشید که در تمامی تمرین‌ها، نمره‌ی اصلی به تفسیر و تحلیل شما تعلق می‌گیرد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. مجموعا ۱۴ روز برای تمامی تمرین‌ها و پروژه‌ی درس به عنوان Grace day در نظر گرفته می‌شود و پس از پایان مجموعا ۱۴ روز، برای هر تمرینی که پس از زمان اختصاص یافته ارسال شود روزی ۱۵ درصد از نمره آن تمرین کسر خواهد شد.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:mohammad.saadati80@gmail.com

mailto:taha.fakharian@gmail.com

مهلت تحویل بدون جریمه: ۲۵ / ۰۳ / ۱۴۰۲