University of Tehran

School of Electrical and Computer Engineering

Data Mining

Final Project

مقدمه

در این پروژه قصد داریم به تبدیل مساله‌ی تخمین خسارت وارد شده به ساختمان‌ها در اثر زلزله براساس ویژگی‌های آن‌ها در قالب یک مساله‌ی طبقه‌بندی بپردازیم. از جمله مواردی که می‌توانند بر روی میزان خسارتی که به یک ساختمان در زلزله وارد می‌شود تاثیر بگذارند می‌توان به منطقه‌ی جغرافیایی که ساختمان در آن قرار دارد، تعداد طبقات ساختمان، مصالح مورد استفاده در ساختمان، هدف ساخت ساختمان، تعدا افراد و خانواده‌هایی که در ساختمان زندگی می‌کنند و وضعیت مالکیت قانونی زمین اشاره کرد. در این پروژه شما با مشکلاتی که ممکن است در طی فرآیند طبقه‌بندی ایجاد شود (مانند زیاد بودن تعداد ویژگی‌ها، مشکل Label Imbalance، و غیره) مواجه شده و برای حل آن‌ها اقدام می‌کنید.

مجموعه داده‌ی این پروژه اطلاعات ساختمان‌های آسیب دیده در زلزله‌ی Gorkha جمع‌آوری شده ست. این پروژه شامل سه مرحله است که مرحله‌ی پایانی امتیازی است. در مرحله‌ی اول، شما باید یک روش مناسب برای آماده‌سازی ویژگی‌ها ارائه دهید، در مرحله‌ی دوم یک مدل مناسب برای انجام هر کدام از مسائل بیابید. نهایتا در مرحله‌ی سوم شما با استفاده از مهندسی ویژگی‌ها، یک مدل تحلیل پذیر با عملکرد بالا بدست خواهید آورد.

شرح پروژه

یکی از مشکلاتی که در مواجهه با بلایای طبیعی مانند زلزله با آن مواجه هستیم تخمین میزان خسارتی است که این زلزله به بار می‌آورد. با داشتن این اطلاعات می‌توان از خسارت‌های وحشتناک بعدی با افزایش مقاومت ساختمان‌ها یا پیش‌بینی‌های درست جلوگیری کرد. در این پروژه مجموعه اطلاعات مرتبط با ساختمان‌ها و میزان خسارت آن‌ها در زلزله‌ی Gorkha در اختیار شما قرار گرفته است. میزان خسارت وارد شده به ساختمان‌ها به صورت یک عدد بین یک تا سه داده شده است که عدد ۱ بیانگر خسارت کم، عدد ۲ بیانگر خسارت متوسط و عدد ۳ بیانگر تخریب کامل ساختمان است. در این پروژه ما به پیش‌بینی میزان خسارت وارد شده به ساختمان‌ها در زمان زلزله می‌پردازیم که می‌توان مساله‌ی تخمین آن را به صورت یک مساله‌ی طبقه‌بندی چندکلاسه[1] مدل کرد.

[1] MulticlassClassification

مجموعه داده شامل اطلاعات مربوط به ساختار ساختمان‌ها و مالکیت قانونی آن‌هاست. هر ردیف در مجموعه داده نشان دهنده‌ی یک ساختمان خاص در منطقه است که در زلزله‌ی Gorkha آسیب دیده است.

۳۹ ویژگی[2] در این مجموعه وجود دارد که ستون Bulding_id یک شناسه‌ی یکتا و تصادفی برای هر ساختمان است. ۳۸ ویژگی باقی‌مانده را در لیست زیر مشاهده می‌کنید:

۱. geo_level_1_id, geo_level_2_id, geo_level_3_id: منطقه جغرافیای که ساختمان در آن قرار دارد و شامل سه سطح می‌باشد.
۲. count_floors_pre_eq: تعداد طبقات ساختمان قبل از زلزله
۳. age: سن ساختمان
۴. area_percentage: مساحت نرمال‌شده‌ی ساختمان
۵. height_percentage: ارتفاع نرمال‌شده‌ی ساختمان
۶. land_surface_condition: وضعیت سطح زمینی که ساختمان در آن ساخته شده است. (n, o, t)
۷. foundation_type: نوع پی استفاده شده در ساخت ساختمان. (h, r, i, u, w)
۸. roof_type: نوع سقف مورد استفاده در ساخت و ساز (n, q, x)
۹. ground_floor_type: نوع طبقه همکف (f, m, v, x, z)
۱۰. other_floor_type: نوع سازه‌های مورد استفاده در طبقات بالاتر از همکف به جز سقف اصلی (j, q, s, x)
۱۱. position: موقعیت ساختمان (j, o, s, t)
۱۲. plan_configuration: نوع پیکربندی ساختمان (a, c, d, f, m, n, o, q, s, u)
۱۳. has_superstructure_adobe_mud: نشان می‌دهد که آیا روبنا از گل و خشت ساخته شده است یا خیر.
۱۴. has_superstructure_mud_mortar_stone: نشان می‌دهد که آیا روبنا از ملات گل و سنگ ساخته شده است یا خیر.
۱۵. has_superstructure_stone_flag: شان می‌دهد که آیا روبنا از سنگ ساخته شده است یا خیر.
۱۶. has_superstructure_cement_mortar_stone: نشان می‌دهد که آیا روبنا از ملات سیمان و سنگ ساخته شده است یا خیر.
۱۷. has_superstructure_mud_mortar_brick: نشان می‌دهد که آیا روبنا از ملات گل و آجر ساخته شده است یا خیر.
۱۸. has_superstructure_cement_mortar_brick: نشان می‌دهد که آیا روبنا از ملات سیمان و آجر ساخته شده است یا خیر.
۱۹. has_superstructure_timber: نشان می‌دهد که آیا روبنا از چوب ساخته شده است یا خیر.
۲۰. has_superstructure_bamboo: نشان می‌دهد که آیا روبنا از بامبو ساخته شده است یا خیر.

[2] Feature

۲۱. has_superstructure_rc_non_engineered: نشان می‌دهد که آیا روبنا از بتن مسلح غیر مهندسی ساخته شده است یا خیر.
۲۲. has_superstructure_rc_engineered: نشان می‌دهد که آیا روبنا از بتن مسلح مهندسی شده ساخته شده است یا خیر.
۲۳. has_superstructure_other: شان می‌دهد که آیا روبنا از مواد دیگری ساخته شده است یا خیر.
۲۴. legal_ownership_status: وضعیت مالکیت قانونی زمینی که ساختمان در آن قرار دارد (a, r, v, w)
۲۵. count_families: تعداد خانواده‌هایی که در ساختمان زندگی می‌کنند
۲۶. has_secondary_use: این ویژگی نشان می‌دهد که آیا ساختمان برای اهداف دیگری ساخته شده است یا خیر.
۲۷. has_secondary_use_agriculture: آیا ساختمان برای اهداف کشاورزی ساخته شده است یا خیر.
۲۸. has_secondary_use_hotel: آیا ساختمان به عنوان هتل استفاده می‌شود یا خیر.
۲۹. has_secondary_use_rental: آیا ساختمان برای مقاصد اجاره‌ای استفاده می‌شود یا خیر.
۳۰. has_secondary_use_institution: آیا ساختمان به عنوان محلی برای تاسیس یک موسسه استفاده می‌شود یا خیر.
۳۱. has_secondary_use_school: آیا ساختمان برای مدرسه استفاده می‌شود یا خیر.
۳۲. has_secondary_use_industry: آیا ساختمان برای مقاصد صنعتی استفاده می‌شود یا خیر
۳۳. has_secondary_use_health_post: آیا ساختمان برای مقاصد بهداشتی و سلامت استفاده می‌شود یا خیر
۳۴. has_secondary_use_gov_office: آیا ساختمان برای یک اداره‌ی دولتی استفاده می‌شود یا خیر
۳۵. has_secondary_use_use_police: آیا ساختمان به عنوان ایستگاه پلیس استفاده می‌شود یا خیر
۳۶. has_secondary_use_other: آیا ساختمان برای اهداف دیگری استفاده شده است یا خیر.

دقت کنید که ویژگی‌های بالا بازه‌ی وسیعی از انواع متغیرها را شامل می‌شوند، و بخشی از مساله چگونگی برخورد با این متغیرها می‌باشد.
با استفاده از این متغیرها قصد داریم مساله‌ی حدس زدن میزان خسارتی که به یک ساختمان در طول این زلزله وارد شده است را پیش بینی کنیم.
لازم به ذکر است در این پروژه بخش زیادی از نمره‌ی شما را کیفیت تحلیل شما از نتایج مشخص خواهد کرد. بنابراین، باید تمامی نتایج بدست آمده تحلیل شوند.

مرحله‌ی 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 مبتنی‌بر 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 بدهید:

- 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؟

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] می‌توانید با جستجوی کلیدواژه‌ی Feature-scaling در این زمینه اطلاعات بیشتری بدست بیاورید.

مرحله‌ی دوم: یادگیری و انتخاب مدل مناسب

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

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

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. 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 ۳:۱:۱ باشد. به علاوه دقت کنید که طبقه‌بندی به صورت Stratified باشد. در صورتی که تعداد داده‌ها به نسبت منابع در دسترس شما زیاد است می‌توانید از یک نمونه‌ی به اندازه‌ی کافی بزرگ از داده‌ها (Stratified بر اساس کلاس‌ها) استفاده کنید.

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، MacroF، و MicroF۱ باشد. برای مدل SVM فقط کرنل‌های Linear، Polynomial ، و Gaussian را بررسی کنید. برای مدل MLP اندازه‌ی لایه‌ی پنهان ۸، ۱۶، ۳۲، و ۶۴ را بررسی کنید. 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

مرحله‌ی سوم: یک مدل تحلیل‌پذیر (امتیازی)

علت عملکرد خوب ورش‌های مبتنی‌بر 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 مبتنی‌بر 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 غیر خطی 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

- گزارش خود را به صورت یک فایل پی دی اف همراه با یک فایل جوپیتر حاوی کدها در سایت Elearn بارگزاری کنید. فایل نهایی باید در قالب یک فایل زیپ و به نام studentID.zip باشد که در آن studentID شماره‌ی دانشجویی شماست.
- مهلت تحویل تمرین تا پایان روز جمعه ۲۰ خرداد ماه می‌باشد. آخرین مهلت تحویل تمرین تا پایان روز جمعه ۲۷ خرداد ماه با جریمه‌ی ۳۰ درصدی می‌باشد. (دقت کنید که این جریمه به صورت روزانه محاسبه نمی‌شود و حتی یک روز تاخیر نیز مشمول جریمه‌ی ۳۰ درصدی می‌شود.)
- برای هر بخش از کدهای خود توضیحات مختصری ارائه دهید. این توضیحات باید عملکرد قطعه کد مورد نظر را به طور مختصر و مفید توضیح بدهد.
- در صورت بروز هر مشکل، با یکی از این دو ایمیل در تماس باشید:
    - s.movahedi94@gmail.com
    - m.biarinezhad@gmail.com