دانشگاه تهران
دانشکده‌ی مهندسی برق و کامپیوتر

آشنایی با هوش مصنوعی

پروژه صفر

پیش‌بینی قیمت خانه

مهلت تحویل: سه‌شنبه شب، ۹ مهر

طراحان:

احمد پوری‌حسینی

محمدرضا طیرانیان

۱. مقدمه

هدف اصلی این پروژه آشنایی شما با برخی کتابخانه‌ها و ابزارهایی است که در مسیر یادگیری مفاهیم هوش مصنوعی و مخصوصا یادگیری ماشین کمک زیادی به شما خواهند کرد. به عنوان یک تمرین عملی برای استفاده از این کتابخانه‌ها، یک پیاده‌سازی نسبتا ساده نیز از یک الگوریتم هوش مصنوعی برای شما در نظر گرفته شده، که در آن شما باید مدلی را طراحی کنید که با ورودی گرفتن چند مشخصه درباره‌ی یک خانه، قیمت آن را پیش‌بینی بکند.

۲. کتابخانه‌ها و ابزارها

کتابخانه‌هایی که در این پروژه قصد معرفی آن‌ها را داریم، کتابخانه‌های numpy ، pandas و matplotlib به همراه ابزار jupyter notebook است. Pandas یک کتابخانه‌ی متن باز و پرسرعت است که data structure ها و ابزارهای تحلیل داده‌ی پرقدرتی را ارائه می‌کند که استفاده از آن‌ها بسیار راحت است. ما در این پروژه بیش‌تر از آن برای کار کردن با داده‌های csv استفاده خواهیم کرد. Matplotlib نیز کتابخانه‌ی مشهور python به منظور کشیدن انواع نمودارها است. در مورد pandas و jupyter notebook نیز در ادامه توضیحات مفصل تر داده خواهد شد.

توصیه می‌شود برای نصب این کتابخانه‌ها از ابزار conda استفاده کنید. این ابزار، نصب بسیاری از کتابخانه های مورد نیاز را برای شما راحت‌تر خواهد کرد و همچنین قابلیت نصب کتابخانه‌ها را به صورت یک virtual environment به شما خواهد داد. در هنگام

ساخت virtual environment توصیه می‌شود از پایتون ۳ برای آن استفاده کنید. برای نصب conda از این لینک کمک بگیرید. هم‌چنین برای نصب کتابخانه‌های مختلف می‌توانید از دستور conda install یا از رابط کاربری گرافیکی conda استفاده کنید. جزئیات بیشتر نصب توسط conda برای هر کتابخانه در سایت آن وجود دارد.

۳. صورت مسئله

a. آشنایی اولیه با داده‌ها:

همانطور که در مقدمه نیز ذکر شد، در این پروژه شما باید یک مدل برای پیش‌بینی قیمت خانه طراحی بکنید. این مدل به عنوان ورودی تعدادی از مشخصه‌های مربوط به یک خانه را دریافت کرده، و به عنوان خروجی قیمت تخمین زده شده برای خانه را برمی‌گرداند.

داده‌های مربوط به این تمرین، زیرمجموعه‌ای از Ames Housing Dataset است که به منظور استفاده‌ی راحت‌تر شما کمی خلاصه شده است و البته نسخه‌ی کامل آن به همراه توضیحاتی کلی راجع به این مجموعه را می‌توانید در این لینک مشاهده بکنید. مجموعه‌ای که شما باید با آن کار بکنید در فایل houses.csv قرار گرفته است که به همراه صورت پروژه به شما داده شده است. محتوای این فایل را با استفاده از کتابخانه‌ی pandas بخوانید. کتابخانه‌ی pandas به شما data frame ای را خروجی خواهد داد که می‌تواند با استفاده از توابع head چند سطر اول آن را ببینید یا با تابع describe اطلاعاتی راجع به توزیع داده‌های هر ستون به دست بیاورید.

همان‌طور که می‌بینید این مجموعه داده به ازای هر خانه به جز خصیصه‌ی هدف، یعنی قیمت خانه، ۱۱ ویژگی دیگر نیز دارد. برخی از این ویژگی‌ها عددی هستند،و برخی دیگر طبقه‌ای[1] . استفاده از داده‌های عددی در طراحی مدل‌های یادگیری ماشین، نسبتا ساده‌تر است. ولی استفاده از داده‌های طبقه‌ای نیاز به تکنیک‌های خاصی دارد. با جلو رفتن در صورت پروژه و دیدن نحوه‌ی استفاده از داده‌های عددی، سعی کنید حدس بزنید که به چه روش‌هایی می‌توان داده‌های طبقه‌ای را برای استفاده در یک مدل یادگیری ماشین مناسب کرد. (نیاز نیست در این مورد چیزی در گزارش کار ذکر کنید.) در این پروژه قرار نیست درگیر کار با این داده‌ها بشوید، در نتیجه کدی بنویسید که این داده‌ها را از dataframe شما حذف کند.

یک نکته‌ی دیگر که معمولا در مجموعه‌داده‌های واقعی با آن روبه‌رو می‌شوید، مسئله‌ی داده‌های گمشده[2] است، که حتی در دقیق‌ترین مجموعه‌داده‌ها نیز خود را نشان می‌دهد. اگر با دقت به محتوای dataframe خود نگاه کنید، خواهید دید که pandas این مقادیر را با NaN نشان می‌دهد. روش‌های مختلفی برای هندل کردن این مسئله وجود دارد، ولی شما باید در این پروژه همه‌ی این مقادیر را با میانگین ستون مربوطه جایگزین کنید. برای این کار از توابع مخصوص pandas که برای همین منظور طراحی شده اند استفاده کنید.

از آنجایی که هدف ما پیش‌بینی قیمت خانه بر حسب ویژگی‌های آن است، خوب است که رابطه‌ی بین هر کدام از ویژگی های عددی را با قیمت خانه را با استفاده از رسم نمودار توسط matplotlib مشاهده بکنید. قطعه کدی بنویسید که همین کار را انجام دهد، یعنی ۹ نودار تولید کند که هر کدام رابطه‌ی یکی از ۹ خصیصه را با قیمت خانه به صورت یک scatterplot نشان بدهد. این ۹ نمودار را در گزارش کار خود بیاورید.

---

[1] categorical

[2] missing data

b. تخمین‌گر خطی

شما در این مرحله باید به منظور تخمین قیمت خانه‌ها یک تخمینگر خطی طراحی کنید. فرمول توصیف کننده‌ی ساختار یک تخمینگر خطی در حالت کلی به شکل زیر است:

$$\hat{y} = wx + b$$

که در آن $\hat{y}$ خروجی مدل، w وزن ورودی (شیب خط)، و b مقدار bias (عرض از مبدأ) است. در حالت کلی ورودی مدل می‌تواند بیش‌از یک عدد باشد و در واقع یک بردار باشد، که در این حالت مقادیر w و b نیز برداری خواهند بود، اما در این پروژه به منظور سادگی فرض میکنیم که ورودی مدل صرفا یک عدد باشد. چیزی که دقت یک تخمینگر خطی را تعیین می‌کند، پارامترهای آن یعنی مقادیر w و b هستند. شما در این بخش باید ابتدا با نگاه کردن به نمودارهای تولید شده در بخش قبل تشخیص بدهید که رابطه‌ی کدامیک از خصیصه ها با قیمت خانه شباهت نسبتا خوبی را به یک رابطه‌ی خطی دارد. سپس سعی کنید مقادیر w و b را طوری حدس بزنید که خط رسم شده توسط آن‌ها به داده‌های واقعی نزدیک باشد. سپس به ازای همه‌ی داده‌های ورودی، نمودار قیمت واقعی در کنار قیمت تخمین زده شده را بر حسب مقدار خصیصه‌ی انتخاب شده رسم کنید.

برای اینکه عملکرد مدل خود را اندازه گیری بکنید، از معیار RMSE[3] استفاده بکنید. تعریف این معیار به شکل زیر است:

$$L_{RMSE}(\hat{y}, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

که در آن N تعداد داده‌ها، $\hat{y}_i$ پیش‌بینی مدل برای داده‌ی i ام، و y_i مقدار واقعی قیمت برای داده‌ی i ام است. شما باید مقادیر پارامترها را طوری تنظیم کنید که مقدار این معیار کمتر از ۹۰ باشد. در نهایت مقدار پارامتر ها، نمودار حاصل و مقدار معیار RMSE را در گزارش کار خود بیاورید.

توجه‌کنید که در این بخش هیچ الزامی به استفاده از هیچ کتابخانه‌ای نیست و صرف اینکه کدتان کار مورد نظر را انجام بدهد کافیست.

c. معرفی numpy و برداری‌سازی[4]

numpy اصلی‌ترین و مهم‌ترین کتابخانه‌ی محاسبات عددی در پایتون است. مهمترین ویژگی‌های این کتابخانه یک کلاس بسیار قدرتمند نمایانگر آرایه‌های N-بعدی، و مجموعه‌ی بسیار بزرگی از انواع توابع پیچیده، از جمله توابع broadcasting است.تقریبا می‌توان گفت آشنایی با این کتابخانه برای برنامه نویسی در زمینه‌ی هوش مصنوعی الزامی است. از آنجایی که آموزش‌های بسیار زیادی در مورد این کتابخانه موجود هستند، در این‌جا به دادن یک لینک از چنین آموزش هایی بسنده می‌کنیم. برای مقدمه‌ای بر numpy می‌توانید به این لینک و برای توضیحات دقیق‌تر در مورد broadcasting به این لینک مراجعه کنید.

[3] root mean square error
[4] vectorization

کتابخانه‌ی numpy علاوه بر ویژگی‌هایی که تا کنون ذکر شد، به شما قابلیت انجام عمل برداری‌سازی را نیز می‌دهد. معرفی این مفهوم به منظور آشنایی شما با برنامه‌ی jupyter notebook تحت عنوان یک notebook با اسم vectorization.ipynb آماده شده است. برای کارکردن راحت تر با این بخش، پیشنهاد می‌شود که ابتدا این لینک که معرفی کوتاهی بر jupyter notebook است را مطالعه کنید. (مطالعه‌ی بخش‌های "Adding Rich Content" و به بعد، اختیاری است.) فایلnotebook این بخش در کنار صورت پروژه به شما داده شده است. برای باز کردن آن کافیست پس از نصب برنامه، به پوشه‌ی حاوی فایل رفته، دستور" . jupyter notebook" را اجرا کرده، و از پنجره‌ی مرورگر باز شده، فایل vectorization.ipynb را انتخاب کنید.

d. بازگشت به تخمین‌گر خطی

حال شما باید با توجه به چیزهایی که در بخش قبل یادگرفته‌اید، و به کمک کتابخانه‌ی numpy کد بخش b را طوری تغییر دهید که فاقد هرگونه ساختار حلقه (for یا while) باشد. استفاده از این ساختار ها در این بخش موجب کسر بخش قابل توجهی از نمره‌ی این بخش خواهد شد. نیازی نیست مجددا چیزی را گزارش کنید، صرف اصلاح کردن کد بخش b کافیست.

e. K-nearest-neighbors

در بخش a با تخمینگرهای خطی آشنا شدید. این تخمینگرها یکی از انواع مدل‌های پارامتری هستند، ولی گاهی اوقات استفاده از مدل‌های غیر پارامتری برای ما سودمند تر است. یکی از این مدل ها، مدل k-nearest-neighbors است. در این مدل، پارامتری وجود ندارد که مقدار بهینه‌ی آن را یادگرفته و برای خانه‌های جدید با استفاده از آن‌ها تخمین را محاسبه کنیم، بلکه تخمین قیمت هر خانه‌ی جدید به صورت مستقیم از روی داده‌هایی داریم محاسبه می‌شود. به این صورت که ابتدا نزدیک‌ترین k نقطه (خانه) را به خانه‌ی جدید پیدا کرده، و میانگین قیمت این خانه‌ها را به عنوان قیمت خانه‌ی جدید گزارش می‌کنیم.

حال شما باید تابعی بنویسید که با گرفتن اطلاعات مربوط به یک خانه‌ی جدید در قالب یک dataframe تک سطری (که فاقد ستون ID و SalePrice است)، مقدار تخمینی برای قیمت آن خانه را گزارش بدهد. مقدار k را در این تابع برابر ۱۰ فرض کنید.

به منظور قابل قضاوت بودن پیاده‌سازی ها، شما باید از فاصله‌ی اقلیدسی به عنوان معیار فاصله استفاده کرده، و مطابق توضیحات پایانی این لینک مقدار همه‌ی خصیصه‌ها را استانداردسازی کنید که خصیصه‌هایی که مقادیر بزرگ‌تری دارند اثر سایر خصیصه‌ها را از بین نبرند.

دقت کنید که در این بخش نیز تا جای ممکن از ساختار حلقه به منظور انجام محاسبات عددی دوری کنید. هر ساختار حلقه‌ای که جایگزین معقولی در کتابخانه‌ی numpy یا pandas داشته باشد باعث کسر نمره از شما خواهد شد.

به فرمت ورودی تابع خود دقت کنید، چرا که این تابع به صورت اتوماتیک تست خواهد شد.

همچنین برای ارزیابی تابع خود می‌توانید نمودار قیمت تخمین زده شده را بر حسب یکی از ویژگی‌های دلخواه رسم کرده، و با نمودار قیمت واقعی بر حسب همان ویژگی مقایسه کنید.

۴. گزارش کار

گزارش کار در همه پروژه ها باید کامل باشد و تصحیح از روی آن انجام می شود. نمودارها و تحلیل‌هایی که در هر مرحله به دست می‌آید، در آن ضمیمه شده باشد. فرمت نهایی گزارش کار باید pdf یا HTML باشد.

۵. نکات پایانی

- تاخیر به ازای روز اول و دوم هر کدام ۱۰ درصد و روز سوم به بعد هر روز ۱۵درصد خواهد بود. برای مثال سه روز تاخیر ۳۵درصد از نمره دریافتی شما را کم میکند.
- برای ما مهم است که حاصل کار خودتان را به ما تحویل دهید. در صورت تقلب برای بار اول به هر دو طرف نمرهی ۱۰۰- تعلق میگیرد و بار دوم معرفی به دانشگاه و ثبت نمره ۰.۲۵ به عنوان تقلب انجام می شود.
- در صورتی که سوالی داشتید در فروم درس مطرح کنید که دیگران هم از جواب آن استفاده کنند.

موفق باشید!