به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

هوش مصنوعی، ترم بهار ۹۷-۹۸

پروژه آشنایی با پایتون، مهلت: از سه شبه ۱۶ بهمن تا شب دوشنبه ۲۲ بهمن

پیش بینی قیمت تاکسی تلفنی

مقدمه

در این پروژه شـــما مقداری با پایتون و کتاب خانه های مفید آن آشـــنا شـــده و تا حدی هم با مفاهیم و کاربردهای هوش مصنوعی روبرو می شوید. مسئله این پروژه تولید یک سیستم پیش بینی کننده قیمت تاکسی تلفنی هست. در واقع باید تابعی ساده طراحی کنید که با گرفتن زمان یا مسافت احتمالی یک سفر قیمت حدودی آن را به عنوان خروجی بازگرداند. این خروجی باید بر اساس داده سایر سفرهایی که انجام شده اند، تخمین زده می شود.

توضیحات دقیق تر مسئله

۱. فایل ride.csv در کنار پروژه قرار گرفته است که حاوی حدود ۵۰۰ مسافرت تاکسی تلفنی هست. در هر سطر این فایل یک رکورد از یک سفر آمده که در آن قیمت، مسافت و زمان آن ثبت شده است.

```
price,duration,distance
25500,2635,22392
18500,1665,21117
```

۲. دو ورودی اصـلی تابع پیش بینی قیمت، زمان و مسـافت هسـتند و خروجی آن قیمت پیش بینی شـده اسـت. برای راحتی کار در نظر بگیرید، جواب نهایی شـــما فقط از یکی از این عوامل باید اســـتفاده کند که انتخاب آن بر عهده خودتان هست.

۳. برای راحتی بیشتر هم فرض کنید تابعی که قیمت سفر را تخمین می زند، خطی است. یعنی با افزایش پارامتری که در بخش قبل انتخاب کرده اید – مثلا زمان – قیمت هم به صـــورت خطی افزایش پیدا می کند. در نتیجه تابع سیستم شما به شکل فرمول زیر خواهد بود که فقط باید دو پارامتر عرض از مبدا (intercept) و شیب(slope) آن را به دست آورید.

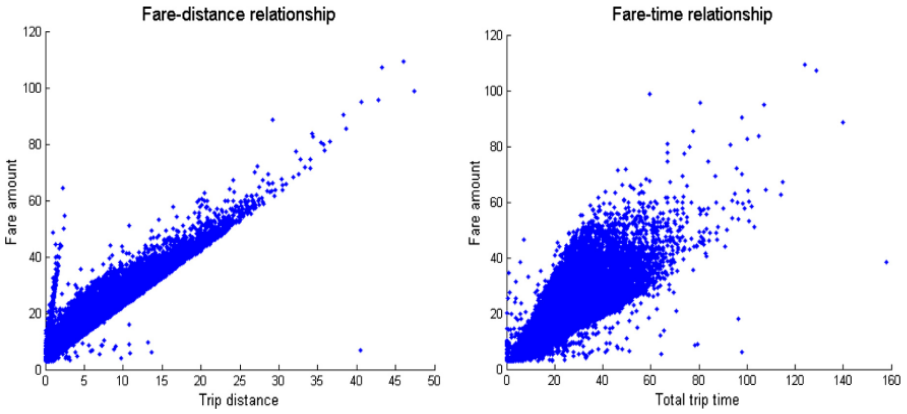
$$fare = time * slope + intercept$$

۴. معیار ارزیابی تابع، خطای آن (اختلافش با واقعیت) است. برای آن که تابع مناسب باشد، باید در صورت اعمال بر روی داده های ride.csv به کمترین خطا برسد. این خطا به صورت فرمول زیر تعریف می شود که real قیمت واقعی یک سفر و estimated تخمین شما از قیمت آن – درواقع خروجی تابع بالا- است.

$$error = \sum(real - estimated)^2$$

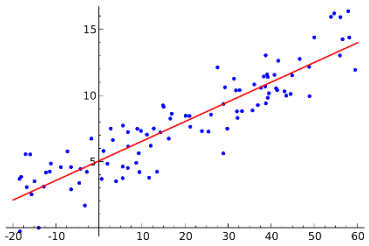
روش حل مسئله

ابتدا فرض کنید کل داده ها را در نمودار قیمت بر حسب زمان و قیمت بر حسب مسافت رسم کرده اید. (تصویر شماره ۱) در این نمودار همان طور که مشاهده می کنید، کلیت داده ها حالت خطی دارند. حالا طبق توضیحات قسمت قبلی تابع هزینه به صورت خطی است که از میان این داده ها می گذرد به طوری که خطای آن کمینه باشد.



تصویر شماره ۱

حالت کلی این مسئله Linear Regression نام دارد که در آن برنامه هوشمند باید خطی را پیدا کرده که از میان نقاط گذشته و با آن ها کمترین فاصله را داشته باشد (تصویر شماره ۲) اما در این پروژه هدف پیاده سازی آن نیست.



تصویر شماره ۲

بلکه به جای آن که برنامه - تو سط الگوریتم یادگیری ما شین - ضرایب را پیدا کند کافی است خودتان به صورت دستی ضرایب شیب و عرض از مبدا فرمول اول را تغییر داده تا خطای محاسبه شده در فرمول دوم کمینه شود. (Handy Linear Regression)

**نکته مهم: تصویر شماره یک مربوط به داده های تاکسی رانی نیویورک و تصویر شماره دو یک رگرشن کلی و نامرتبط است و اطلاعات آن برای استفاده در این پروژه مناسب نیست. شما باید کار مشابه را بر روی داده هایی که از تهران جمع آوری شده و داخل فایل* ride.csv *است، انجام دهید.*

خروجی نهایی

۱۰ عدد در بازه مناسب تولید کرده (به صورت رندوم یا منظم) و خروجی تابع خطی را به ازای آن ها ذخیره کنید. جدولی از فایل ذخیره شده رو در گزارش کار خود قراردهید.

هیستوگرام (امتیازی)

برای تحلیل بهتر داده های – برای مثال توزیع زمان تاکسی ها – کشیدن هیستوگرام آن ها مناسب است. از نامپای و متپلات برای آن استفاده کنید.

ابزارها و کتابخانه های مهم

Pandas

پانداس یک کتابخانه پیشرفته برای کار کردن با داده هاست. در این پروژه خیلی راحت تر (و حرفه ای تر) هست که برای خواندن داده ها و کار با آن ها از پانداس و دیتافریم هایش استفاده کنید.

**نکته: استفاده از پانداس اجباری نیست اما در صورت استفاده کامل و مناسب از آن نمره امتیازی می گیرید.*
**نکته ۲: در صورتی که از پانداس استفاده نمی کنید، برای خواندن ورودی باید از کتابخانه مناسبی مانند csv استفاده کنید.*

Numpy

با این کتابخانه آشنا شده و برای تولید اعداد قسمت 'خروجی نهایی' از آن استفاده کنید. همین طور در قسمت هیستوگرام نیز از این کتابخانه برای گرفتن خروجی استفاده کنید.

Matplotlib

برای کشیدن داده ها - مانند تصویر شماره ۱ - از کتابخانه matplotlib استفاده کنید. درواقع وقتی حجم زیادی داده دارید بهتر هست که آن ها را مانند تصویر شماره ۱ بکشید و تابع خطی که به دست می آوردید را هم روی آن رسم کنید(visualization) تا شهود بهتری داشته باشید که شیب و عرض از مبدا خط را به چه صورتی باید تغییر دهید که خطا کمتر شود.

**نکته: در تمام پروژه ها تصویر سازی و ضمیمه تصاویر در گزارش کار اهمیت بالایی دارد.*

Jupyter Notebook

استفاده از نوتبوک خیلی از مواقع – مانند این پروژه - باعث راحت تر شدن انجام آن می شود. همین طور گزارش کار گرفتن از نوت بوک بسیار ساده تر است. در این جا برای شروع از نوت بوک استفاده کنید اما در مابقی پروژه ها استفاده از آن بر عهده خودتان خواهد بود.

sklearn/ scikit-learn

در این پروژه و هیچ کدام از پروژه های بعدی از این کتابخانه نباید استفاده کنید مگر این که دقیقا ذکر شده باشد.

گزارش کار

گزارش کار در همه پروژه ها باید کامل باشد و تصحیح از روی آن انجام می شود. نمودارها و تحلیل هایی که در هر مرحله به دست می آید، در آن ضمیمه شده باشد. در این پروژه نیز نمودارها، تصویر خط و جواب نهایی را حتما در گزارش قرار دهید. فرمت نهایی گزارش کار باید pdf یا html باشد.

**نکته: گزارش کار نمره جداگانه ندارد ولی تمامی تصحیح ها از روی آن انجام می شود. به عبارتی ضعیف بودن گزارش کار به طور مستقیم باعث کسر نمره نمی شود اما نوع انعکاس فعالیت هایی که در پروژه انجام داده تاثیر جدی رو نمره نهایی شما دارد.*

- در صورتی که سوال داشتید در فروم درس مطرح کنید و اگر ابهامی داشتید هم می تونید حضورا با من صحبت کنید.

موفق باشید!

- آرمین -