



PH #4: Deep Neural Network

پیشبینی تقاضای خرید بلیط هواپیما

تخصیص بهینه و متناسب با نیاز منابع و تجهیزات در صنایع و تجارت‌های مختلف بسیار حائز اهمیت است. داشتن یک پیشبینی دقیق و درست نسبت به نیازهای پیشرو میتواند یک گام بسیار مهم برای تخصیص بهینه منابع باشد. در همین راستا قصد داریم تا بتوانیم با استفاده از ابزارها و راهلهای نوین داده کاوی و یادگیری ماشین (شبکه عصبی)، تعداد درخواست مسافران برای خرید بلیط هواپیما بین شهرهای مختلف را در روزهای مختلف سال پیشبینی کنیم.

سناریو

داده های مربوط به حدود 2 میلیون خرید بلیط هواپیما مربوط به خطوط مختلف هوایی موجود است. تاریخ خرید بلیط مربوط به سالهای 1395 و 1396 میباشد. هر سطر داده شامل اطلاعاتی از جمله «تاریخ و زمان خرید بلیط»، «تاریخ و زمان پرواز»، «قیمت بلیط»، «شماره فرودگاه مبدا»، «شماره فرودگاه مقصد» و «شماره ایرلاین» میباشد. با استفاده از این داده ها میتوان تعداد خرید بلیط در هر روز را برای هر زوج «شماره فرودگاه مبدا» و «شماره فرودگاه مقصد» محاسبه کرد. در اینجا، هدف پیشبینی تعداد خرید بلیط برای زوج های مختلفی از «شماره فرودگاه مبدا» و «شماره فرودگاه مقصد» در روزهای آتی است. از شما انتظار میرود با استفاده از تکنیک شبکه عصبی مدل مناسبی بر اساس داده های 1.5 سال گذشته طراحی کنید و از آن مدل برای پیشبینی دقیق متغیر هدف در روزهای مشخصی در 6 ماه دوم سال 1396 استفاده کنید. فایل آزمون شامل حدود 35000 سطر است که شما باید با توجه به «تاریخ خرید بلیط»، «شماره فرودگاه مبدا» و «شماره فرودگاه مقصد» مربوطه، پیشبینی خود را از تعداد خرید بلیط در آن تاریخ و بین شهر مبدا و مقصد ارائه کنید.

داده ها

داده های ارائه شده مربوط به خریدهای بلیط از سایت علی بابا است. در ادامه چند نکته در مورد فایل های ضمیمه شده آورده شده است:

- هر سطر از داده ها نشانگر یک درخواست خرید بلیط است.
- خط اول هر فایل داده حاوی نام ویژگی ها است.
- در تمام سطرها، علامت "،" ستونها را از هم جدا میکند.
- علامت "." نشانگر جدا کننده اعشار است.

دقت کنید که امکان وجود داده ی پرت وجود دارد و تیم ها بنا به صلاحدید خود میتوانند هر پردازی روی آنها انجام دهند.



نام ستون	توضیحات	محدوده‌ی مقادیر
Log_Date	تاریخ درخواست خرید	YYYY/MM/DD
FROM	مبدأ مسیر	اعداد صحیح مثبت
TO	مقصد مسیر	اعداد صحیح مثبت
Sales	تعداد درخواست خرید	اعداد صحیح مثبت

فایل نهایی باید دقیقاً با فرمت بالا باشد. مثال زیر یک بارگذاری صحیح را نشان می‌دهد:

Log_Date,FROM,TO,Sales

1397/01/01,3,30,10

1397/01/01,3,49,10

1397/01/01,3,66,10

...

برای فایل نهایی از فرمت csv استفاده شود.

معیار ارزیابی

برای هر ترکیب "Log_Date, FROM, TO" که در فایل "test.csv" داده شده است، مقدار "Sales" باید پیشبینی شود. ارزیابی با معیار MAPE (Mean Absolute Percentage Error) انجام خواهد پذیرفت:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{A_i - P_i}{A_i} \right|$$

که A_i نشان دهنده‌ی مقدار واقعی و P_i پیشبینی شده هستند.

گزارش و کد

هر تیم باید اقدام به تهیه‌ی گزارشی در مورد روش حل مسئله و همچنین دید بیزینسی مسئله کند. بررسی دید بیزینسی از اهمیت بیشتری برخوردار است.