تمرین ۳-الف – آمار و احتمال در فرآیند یادگیری

مجموعه دادگان مربوط به **۱۰** سهم (سری زمانی)، "Stock.csv" در اختیار شما قرار داده شده است. دادگان جمع‌آوری شده، ارزش هر سهم در یک بازه زمانی مشخص را نشان می‌دهد.

پیش‌پردازش دادگان

۱. ابتدا این فایل را با استفاده از کتابخانه pandas تبدیل به یک دیتافریم کنید.

۲. با توجه به این که در بخش‌هایی از این مجموعه داده، مقدار از دست رفته[1] داریم باید در مورد این مقادیر تصمیم بگیریم. برای این کار هر مقداری که وجود نداشت را برابر میانگین دو روز قبل و دو روز بعد قرار دهید. (در تمام بخش‌های این تمرین در برخورد با داده‌های از دست رفته به همین صورت عمل کنید.)

۳. تابعی بنویسید که (بدون استفاده از کتابخانه) مقادیر هر سهم را با استفاده از روش نرمال‌سازی Min-Max[2] به عددی بین **۰** و **۱** تغییر مقیاس دهد.

٤. استانداردسازی یک تکنیک مقیاس‌پذیری مجدد است که به متمرکز کردن توزیع داده‌ها بر روی مقدار میانگین **۰** و انحراف معیار به مقدار **۱** اشاره دارد. میانگین و انحراف معیار با هم می‌توانند برای خلاصه کردن یک توزیع نرمال استفاده شوند که توزیع گاوسی (نرمال) نیز نامیده می‌شود. برای این کار باید مقادیر میانگین و انحراف معیار اعضای هر ستون (سهم) را محاسبه کنیم.

انحراف معیار از طریق فرمول زیر محاسبه می‌شود:

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(value_i - mean)^2}{count(values) - 1}}$$

- دو تابع (بدون استفاده از کتابخانه) بنویسید که دو مقدار میانگین و انحراف معیار را برای هر ستون تخمین بزند.
- مقدار میانگین چه درکی از مقادیر یک ستون به ما می‌دهد؟
- انحراف معیار چه درکی از مقادیر یک ستون به ما می‌دهد؟
- با استفاده از این دو معیار (میانگین و انحراف معیار) مجموعه‌دادگان را استاندارد کنید.

استانداردسازی از طریق فرمول زیر محاسبه می‌شود:

$$\text{standardized value} = \frac{value_i - mean}{stdev}$$

[1] Missing data
[2] Feature scaling

تخمین

جهت انجام محاسبات این بخش، از مجموعه‌دادگان اولیه (بدون اعمال تغییرات بخش قبل) استفاده کنید.

۱. برای هر سهم یک توزیع نرمال با ***۱۰*** نمونه، که میانگین این توزیع برابر مقدار هر سهم در روز ***۱۴ نوامبر ۲۰۲۲*** باشد، تولید کنید. (می‌توانید برای تولید این توزیع از کتابخانه numpy استفاده کنید.)

۲. تابعی بنویسید که (بدون استفاده از کتابخانه) مقدار مجذور میانگین مربعات خطا[3] را برای دو ورودی که، یکی مقدار واقعی و دیگری مقدار پیش‌بینی شده است حساب کند.

۳. از توزیع نرمال تولید شده برای هر کدام از سهم‌ها، سه مقدار(Min, Max, Mean) را انتخاب کرده و هر کدام از این مقادیر را به عنوان مقدار پیش‌بینی یک سهم در ***روز ۱۵ نوامبر ۲۰۲۲*** استفاده کنید و مقدار RMSE را گزارش کنید.

٤. حال با توجه به مقدار RMSE سعی کنید با تغییر پارامترهای توزیع نرمال (σ, n) و استفاده از سه مقدار (Min, Max, Mean) به بهترین مقدار RMSE دست یابید. مقادیر حاصل از محاسبه RMSE برای هر کدام از این سه مقدار را جهت استفاده در بخش بعدی نگه‌دارید. (برای ارزیابی این بخش شما باید حتما حداقل بیست مرتبه مقادیر مختلف یعنی حداقل بیست توزیع مختلف برای هر سهم را آزمایش کرده باشید.)
فرمول محاسبه مجذور میانگین مربعات خطا (RMSE) به صورت زیر است:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{totalpredictions}}$$

٥. از مقادیر به‌دست آمده از قسمت ۴، سه نمودار برای هر سهم، به طوریکه هر نمودار شامل دو منحی از مقدار واقعی سهم و مقدار انتخاب شده (Min, Max, Mean) باشد، رسم کنید. این نمودارها را از نظر برازش تحلیل کنید.

تخمین بیشینه شباهت

در این بخش می‌خواهیم وضعیت جابجایی یک سهم (صعودی و نزولی) را تخمین بزنیم. وضعیت هر روز در مقایسه با روز قبلی به صورت باینری تعیین می‌شود به طور مثال اگر دیروز مقدار سهم اپل برابر ۱۰۰ دلار بود

[3] RMSE

و امروز مقدار سهم ۱۱۰ دلار باشد وضعیت جابجایی سهم اپل در امروز ۱ می‌شود و در صورتی که امروز مقدار سهم برابر ۹۹ دلار باشد وضعیت جابجایی برابر ۰ خواهد بود. (اولین سطر مجموعه‌دادگان را به عنوان مرجع تغییر ندهید و در محاسبات بعدی استفاده نکنید.)

۱. با توجه به توضیحات ابتدای این بخش، با حفظ دادگان کنونی، یک نمونه باینری از این دادگان، ایجاد کنید.

۲. تخمین بیشینه شباهت را می‌توان به راحتی به حالتی تعمیم داد که هدف، تخمین احتمال شرطی $P(y|X;\theta)$ به منظور پیش‌بینی y براساس X است. از آن جایی که مسئله جاری، باینری است می‌توان برچسب نمونه‌ها را با یک خط جدا کرد. اگر توزیع احتمال نمونه‌های خود را دو جمله‌ایی فرض کنیم، برچسب هر نمونه، نتیجه‌ی آزمایش برنولی خواهد بود. توزیع برنولی دارای یک پارامتر (P) که نشان‌دهنده‌ی احتمال برد یا باخت است. برای محاسبه تخمین بیشینه شباهت در توزیع برنولی به صورت زیر عمل می‌کنیم:

$$\text{likelihood} = \hat{y} \times y + (1-\hat{y}) \times (1-y)$$

شش سهم AMN,TSLA,META,NVDA,WMT,MSFT را از هر دو مجموعه‌دادگان قسمت قبل جدا کنید، سپس از مقادیر سهام AMN در مجموعه‌دادگان اولیه یک کپی بگیرید و تمام داده‌های موجود در آن را حذف کنید. با استفاده از فرمول زیر برای وضعیت هر روز سهام AMN، در مجموعه‌دادگان اولیه، از مقادیر پنج سهم دیگر استفاده کنید.

$$\hat{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_m \cdot x_m$$

در این فرمول X ارزش پنج سهم هستند و β :

$$\beta = [0.25, 0.71, 0.33, 0.48, 0.47, 0.5]$$

برای نرمال‌سازی اعداد واقعی به مقادیر احتمالی باید از فرمول زیر استفاده کرد:

$$\hat{y} = \frac{1}{1 + e^{-(x\beta)}}$$

۳. حال مقادیر سهم AMN را با استفاده از فرمول فوق به صورت احتمالی درآورید؛ این مقادیر، وضعیت پیش‌بینی سهم هستند، با توجه به این مقادیر و ستون داده مربوط به AMN در مجموعه‌دادگان باینری به عنوان برچسب واقعی، تخمین بیشینه شباهت را در سال ۲۰۱۹ محاسبه کنید.

٤. این‌بار مقادیر β را با توجه به تحلیل خروجی تابع تخمین بیشینه شباهت آپدیت کنید، سپس تحلیل خود را در مورد تاثیرگذاری سهام‌های پنج‌گانه بر روی وضعیت سهام AMN گزارش کنید. (**امتیازی**)

۵. با استفاده از توابع بهینه‌ساز وضعیت جابه‌جایی سهم AMN را پیش‌بینی کنید و با مقادیری که خودتان بدست آورده‌ید مقایسه کنید، تحلیل نتایج خود را گزارش کنید؛ استفاده از کتابخانه‌ها صرفا جهت بهینه‌سازی در این قسمت مجاز است. (**امتیازی**)