بخش تئوری

۱. (تعاریف پایه) به هریک از سوالات زیر پاسخ دهید.

الف) رگرسیون خطی چیست و چگونه در مدل پیش‌بینی استفاده می‌شود؟ با ذکر مثالی در داده‌های دنیای واقعی (مانند طبیعت و ...) این مدل را بیان کنید.

ب) مفروضات و ویژگی‌های مورد نظر در رگرسیون خطی را بیان کنید.

ج) رگرسیون لجستیک چیست و چه تفاوتی با رگرسیون خطی دارد و مفروضات مطرح شده در این مدل را بیان کنید.

د) براورد بیشینه درست نمایی یا به اختصار MLE را توضیح دهید و نحوه‌ی استفاده از آن را در رگرسیون لجستیک بیان کنید.

۲. مجموعه داده زیر را در نظر بگیرید.

X	Y
5	2
0	1
2	1
1	1
2	0

رگرسیون تک متغیره Y بر روی X را بدست آورید.

۳. (اثبات ناریبی برآوردگر رگرسیون یک متغیره) مدل رگرسیون خطی یک متغیره را در نظر بگیرید. فرض کنید n داده داریم که شامل متغیرهای پیشگو (متغیرهای مستقل) و متغیرهای پاسخ (متغیرهای وابسته) به‌صورت

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

فرض کنید مقدار واقعی متغیر پاسخ برابر است با

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

مقدار پیش‌بینی شده به‌ازای هر متغیر x_i برابر است با

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

و مقدار خطا به ازای هر داده، برابر است با تفاضل مقدار متغیر پاسخ و مقدار پیش‌بینی شده. یعنی

$$e_i = y_i - \hat{y}_i$$

و در حالت نمایش برداری به ازای تمام داده‌ها، مقدار پیش‌بینی شده و خطا را بصورت زیر نمایش می‌دهیم

$$\mathrm{y} = \beta_0 + \beta_1 x + \mathrm{e}, \quad \mathrm{e} = \mathrm{y} - \hat{y}$$

$$\hat{y} = \theta_0 + \theta_1 x$$

بطوری که

$$\mathrm{y} = (y_1, y_2, \ldots, y_n), \mathrm{x} = (x_1, x_2, \ldots, x_n), \mathrm{e} = (e_1, e_2, \ldots, e_n), \hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$$

حال ثابت کنید برآوردگر $\hat{y}$ ناریب است.

(راهنمایی: در ابتدا می‌توانید نشان دهید که $\mathrm{E}[\mathrm{e}] = 0$ و سپس نشان دهید $\mathrm{E}[\theta_0] = \beta_0$ و $\mathrm{E}[\theta_1] = \beta_1$. و سپس به‌راحتی می‌توان برررسی کرد که $\mathrm{E}[\hat{y}|x] = \beta_0 + \beta_1 x$. برای این‌کار لازم است از $\mathrm{MSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ بهترین مقادیر θ_0 و θ_1 که MSE را مینیمم می‌کند، بدست آورید و سپس موارد بالا را در آن بررسی کنید. به شرط صادق بودن این موارد، گوییم برآوردگر ناریب است و امیدواریم مقدار پیش‌بینی شده، با مقدار مجهولی که از آن اطلاعاتی نداریم، برابر شود.)

بخش پیاده سازی

۱. قصد داریم رگرسیون خطی چند متغیره را پیاده سازی کنیم.

الف) با استفاده از کتابخانه sklearn دیتاست boston را لود کنید. ۸۰ درصد از نمونه‌های دیتاست را برای داده‌های train و ۲۰ درصد را برای داده‌های test با استفاده از train_test_split جدا کنید و داده‌ها را نرمال‌سازی کنید. در فرآیند آموزش از Mini-Batch Gradient Descent استفاده کنید و تابع خطا را Mean Squared error در نظر بگیرید.

ب) با استفاده از نرم ۲ ($L_2 - norm$) تابع هزینه در بخش الف را تغییر دهید و پایدارسازی (regularization) را اعمال کنید.

ج) نتایج بخش الف و ب را با یکدیگر مقایسه کنید.

۲. طبقه‌بند لجستیک (logistic regression) را اعمال کنید.

الف) مانند سوال ۱، دیتاست breast-cancer را از sklearn لود کنید. ۸۰ درصد نمونه‌ها را برای train و ۲۰ درصد را برای test با استفاده از train_test_split جدا کنید و داده‌ها را نرمال‌سازی کنید.

ب) با استفاده از نرم ۲ ($L_2 - norm$) تابع هزینه در بخش الف را تغییر دهید و پایدارسازی (regularization) را اعمال کنید.

ج) نتایج الف و ب را با یکدیگر مقایسه کنید.

۳. (استفاده از کتابخانه) سوال ۱ را با استفاده از کتابخانه sklearn پیاده‌سازی کنید. نتایج و پارامترهای بدست آمده را با نتایج و پارامترهای سوال ۱ مقایسه کنید.

۴. (استفاده از کتابخانه) سوال ۲ را با استفاده از کتابخانه sklearn پیاده‌سازی کنید. نتایج و پارامترهای بدست آمده را با نتایج و پارامترهای سوال ۲ مقایسه کنید.

۵. بصورت دلخواه، یک مثال ساده در معادله نرمال پیاده‌سازی کنید.

نکات

- تمرین‌ها را در سامانه ایلرن بارگزاری کنید.
- تمام تمرین‌های تئوری، باید بصورت دست‌نویس و خوانا باشند.
- لطفا گزارش را به زبان فارسی بنویسید و تمامی نکات، فرض‌ها و فرمول‌ها در آن ذکر شود. گزارش در روند تصحیح تمارین، از اهمیت ویژه‌ای برخوردار است.
- **کپی کردن کدهای آماده موجود در اینترنت و یا استفاده از تکالیف همکلاسی‌ها تقلب محسوب می‌شود.**
- **درصورت مشاهده تقلب، نمرات تمامی افراد شرکت کننده در آن، صفر لحاظ می‌شود.**
- بجز مواردی که ذکر شده از کتابخانه sklearn استفاده شود، در دیگر موراد فقط از توابع پایتون و کتابخانه numpy استفاده شود.