

درس محاسبات نرم

تکلیف شماره یک ( رگرسیون خطی ) - ۲ نمره

داده‌های بیماری دیابت برای ۴۴۲ بیمار در کتابخانه sklearn موجود است. جزییات این داده‌ها در آدرس زیر را مطالعه کنید:

[https://scikit-learn.org/stable/datasets/toy\\_dataset.html](https://scikit-learn.org/stable/datasets/toy_dataset.html)

هدف رگرسیون Ridge با ثابت تنظیم  $\alpha$  برای پیش‌بینی میزان پیشرفت بیماری طی یک سال می‌باشد. برای یادگیری ضرایب از روش SGD استفاده نمایید. برای رگرسیون داده‌ها را به سه دسته تقسیم کنید. از 70% داده‌ها برای یادگیری (Learning) و 15% برای اعتبارسنجی (Validation) و از الباقی داده‌ها برای آزمایش (Testing) استفاده کنید.

داده‌های اعتبارسنجی برای تنظیم ابرپارامترها در یک مساله استفاده می‌شود. در این مساله ابرپارامترهای ثابت تنظیم  $\alpha$  (Regularization Constant) و نرخ یادگیری  $\eta$  وجود دارد. جهت تنظیم این پارامترها باید مقدار میانگین مربع خطا (Mean Square Error) را برای داده‌های یادگیری و داده‌ی اعتبارسنجی بازای مقادیر مختلف ابرپارامترها مثلاً  $\alpha$  رسم نمایید. منحنی برای داده‌های یادگیری همواره صعودی و یا همواره نزولی است اما برای داده‌های اعتبارسنجی دارای کمینه‌ی خطا خواهد بود که پارامتر مذکور را برای آن نقطه می‌توان تنظیم نمود.

در نهایت بعد از تنظیم ابرپارامترها و بدست آوردن ضرایب رگرسیون، مقدار خطا را در داده‌های آزمایش محاسبه نمایید.