سوال 1:

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

هدف رگرسیون Ridge با ثابت تنظیم α 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 α (Regularization Constant) و نرخ یادگیری η 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 α 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 پایان بعد از تنظیم ابرپارامترها و بدست آوردن ضرایب رگرسیون، مقدار خطا را در داده‌های آزمایش محاسبه نمایید.

سوال 2:

براساس لینک زیر PCA را پیاده سازی کنید

/https://dataminingbook.info/projects/proj_pca

و دو بردار اصلی با مقدار ویژه بزرگتر را برای داده‌های سوال 1 محاسبه کنید و پاسخ خود را با نتیجه‌ی استفاده از متد PCA از کتابخانه‌ی sklearn مقایسه نمایید.

سوال 3:

حال تنها با استفاده از دو بردار اصلی بدست آمده در سوال 2، سوال 1 را تکرار نمایید.

دقت فرمایید که تنها مجاز به استفاده از کتابخانه‌های numpy و pandas و mathplotlib و برای پیش‌پردازش بخش preprocessing و datasets در sklearn هستید. البته برای مقایسه پاسخ خود از هر کتابخانه‌ای می‌توانید استفاده کنید.

لینکهای راهنما:

https://scikit-learn.org/stable/modules/preprocessing.html

https://numpy.org/doc/stable/reference/generated/numpy.linalg.eigh.html