به نام خدا

دانشگاه صنعتی شریف
دانشکده مهندسی برق

تمرین سری سوم یادگیری آماری
پاییز ۱۴۰۲
دکتر هدی محمدزاده

۱- فرض کنید سه کلاس با توزیع های زیر داریم:

$$P(x|y_1) = N\left(\begin{bmatrix}0.5\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) \qquad P(y_1) = 0.3$$

$$P(x|y_2) = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2 & 0.5\\0.5 & 2\end{bmatrix}\right) \qquad P(y_2) = 0.3$$

$$P(x|y_3) = N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}5 & 4\\4 & 5\end{bmatrix}\right) \qquad P(y_3) = 0.4$$

احتمال قرار گرفتن نمونه $x^{(0)} = \begin{bmatrix}0.5\\0.5\end{bmatrix}$ در سه توزیع را محاسبه کنید و کلاس این نمونه را مشخص کنید.

۲- تابع توزیع احتمال زیر را در نظر بگیرید:

$$P_\theta(x) = 2\theta x e^{-\theta x^2}$$

که θ پارامتر مسئله و x یک عدد ثابت مثبت است. فرض کنید m نمونه ی x_i که $i.i.d.$ هستند را از این توزیع در اختیار دارید. تخمین $Maximum\ Likelihood$ را برای θ بر اساس نمونه ها بدست آورید.

۳- فرض کنید دیتاستی را جمع آوری کرده‌ایم که در آن بر اساس میزان ساعت مطالعه (X_1) و معدل (X_2)، شانس گرفتن نمره A (Y) را تخمین می‌زنیم. برای رسیدن به این تخمین از یک مدل $logistic\ regression$ استفاده می‌کنیم. پس از کردن fit مدل بر روی داده ها، ضرایب این مدل به صورت $\hat{\beta}_0 = -6,\ \hat{\beta}_1 = 0.05,\ \ \hat{\beta}_2 = 1$ هستند.

الف: احتمال اینکه دانش آموزی با ۴۰ ساعت مطالعه و معدل ۳,۵ بتواند نمره A بگیرد، را محاسبه کنید.

ب: این دانش آموز با معدل ۳,۵، برای اینکه50% شانس دریافت نمره A را داشته باشد، باید حداقل چند ساعت مطالعه داشته باشد؟

۴- یک مسئله طبقه بندی دو کلاسه را با نمونه های شامل یک ویژگی در نظر بگیرید. نمونه های کلاس ۱ به طور یکنواخت در بازه ی $[-5,5]$ و نمونه های کلاس صفر به طور یکنواخت در بازه $[a,b]$ قرار دارند. می‌خواهیم مرز طبقه بندی با روش تحلیل افتراقی خطی (LDA) پیدا کنیم. توجه داشته باشید که در این روش، واریانس نمونه های دو کلاس، برابر فرض می‌شوند.

اگر فرض کنیم که احتمال های اولیه دو کلاس با هم برابر باشند، آیا در این مسئله مقدار واریانس نمونه ها، در محل قرارگیری مرز طبقه بندی موثر است؟ جواب خود را اثبات کنید.

۵- یک مسئله طبقه بندی سه کلاسه را با دو ویژگی (X_2, X_1) در نظر بگیرید. فرض کنید $Discriminant\ Function$ هر کلاس را مطابق زیر بدست آورده ایم:

$$\delta_1(x) = x_1 + x_2 - 1$$

$$\delta_2(x) = 2x_1 + 3$$

$$\delta_3(x) = x_2$$

نواحی را در فضای $\mathbb{R}^2$ که هر کلاس به آن تعلق دارد را رسم کنید.

۶ - فرض کنید π_A , π_B نشان دهنده $P(Y=B)$ و $P(Y=A)$ باشند و (μ_B, Σ_B) و (μ_A, Σ_A) میانگین و ماتریس کواریانس کلاس های A و B را مشخص کنند.

$$N(\mu_A, \Sigma_A) = \left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2 & 0.5\\0.5 & 2\end{bmatrix}\right), \pi_A = 0.6$$

$$N(\mu_B, \Sigma_B) = \left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}5 & 4\\4 & 5\end{bmatrix}\right), \pi_B = 0.4$$

اگر از QDA برای طبقه بندی استفاده کنیم، مرز تصمیم گیری را بدست آورید.

۷- سوال پایتون

در این سوال قصد داریم به صورت عمیق تر با مفاهیم PCA و همچنین روش های تحلیل افتراقی بیشتر آشنا شویم.

آشنایی با تبدیل PCA

- در ابتدا با استفاده از کتابخانه های ساده ی پایتون PCA را خودتان پیاده سازی کنید.
- سپس با استفاده از کتابخانه ی skylearn دیتاست IRIS را بخوانید PCA را روی آن طوری اعمال کنید که بتوانید کل dataset را در یک صفحه ی دو بعدی بکشید. (در واقع داده ها را دو بعدی کنید و توجه کنید که کلاس های مختلف باید رنگ تفکیک پذیر داشته باشند)

آشنایی با طبقه بند LDA

در ادامه ی این تمرین نیازی نیست از PCA ای که خودتان پیاده سازی کردید استفاده کنید بلکه میتوانید از توابع آماده برای این کار استفاده کنید.

- در ابتدا طبقه بند LDA را خودتان پیاده سازی کنید برای اینکار کلاس LDA را بسازید که از کلاس زیر ارث بری میکند و تابع های آنرا کامل کنید.

```python
class Classifier:
    def predict(self):
        raise NotImplementedError("Predict method must be implemented in subclasses.")

    def score(self):
        raise NotImplementedError("score method must be implemented in subclasses.")
```

سپس مقادیر زیر را کامل کنید:

```
class LDA(Classifier):
    def __init__(self):
        self.class_means = None
        self.covariance_matrix = None

    def train(self, X_train, y_train):
        pass

    def predict(self, X_test):
        pass

        return np.array(predictions)

    def score(self,y_test):
        pass
```

- در ادامه باید از فایل زیپی که در اختیارتان قرار داده شده است استفاده کنید، این فایل شامل ۴۰ پوشه است که هر پوشه شامل ۱۰ تصویر از صورت یک فرد از زاویه ی مختلف می باشد که ما در این تمرین میخواهیم به کمک طبقه بند LDA این افراد را شناسایی کنیم.
- در ابتدا درباره ی eigenface و را بطه ی آن با PCA تحقیق کنید و توضیح دهید.
- فایل را بخوانید و ماتریس train و test خود را بسازید، (برای هر فرد ۵ تصویر اول را برای train و پنج تصویر دوم را برای test بگیرید) و دقت کنید label ها را هم باید خودتان بسازید.
- تبدیل PCA را روی داده اعمال کنید و بعد آنرا کم کنید.(مقداری بین ۴۰ تا ۵۰ مناسب است)
- سپس طبقه بند LDA که خودتان طراحی کرده اید را آموزش دهید و دقت آنرا بر روی داده های تست گزارش کنید.
- حالا با کمک تابع های آماده ی python، مرحله ی قبل را انجام دهید و نتیجه تان را گزارش کنید و با هم مقایسه کنید.
- همه ی مراحل بالا را برای ابعاد مختلف PCA امتحان کنید و نظرتان را در این باره کاملا شرح دهید. به نظرتان چه زمانی باید کاهش بعد را تمام کنیم.