تمرین سوم یادگیری ماشین

توجه:

۱. **استفاده از کتابخانه‌های از پیش آماده مجاز است.**

۲. **لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.**

سوال اول

داده‌های دو کلاس در فضای دو بعدی به صورت زیر داده شده است. با استفاده از روش LDA بهترین برداری را پیدا کنید که داده‌های دو کلاس بعد از نگاشت بر روی آن، بیشترین فاصله را مطابق معیار LDA داشته باشند. معیار LDA، پارامترهای بهینه مدل و نتایج نهایی بدست آمده را گزارش دهید. تمام روند بایستی به صورت تشریحی انجام شود.

$$C_0 = \left\{ \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix} \right\}$$
$$C_1 = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right\}$$

سوال دوم

۱. ماتریس پراکندگی درون کلاسی و بین کلاسی را برای یک مجموعه داده با بیشتر از دو کلاس تعریف کنید.

۲. با توجه به ماتریس‌های تعریف شده در بخش قبل، یک تابع هزینه برای LDA با بیشتر از دو کلاس ارائه دهید.

۳. حال با کمینه کردن تابع هزینه بخش قبل و نوشتن محاسبات مربوطه، بردار جهت نهایی را بیابید.

سوال سوم

۱. اهداف احتمالی از انجام LDA چیست؟ به بیان دیگر LDA پاسخگوی چه مشکل یا مشکلاتی است و برای کدام دسته از مجموعه داده‌ها نتایج مثبت یا منفی خواهد داشت؟

۲. آیا روش LDA با تعریف یادگیری بدون ناظر همخوانی دارد و می‌توان از آن برای خوشه بندی [1] استفاده کرد؟

[1] Clustering

سوال چهارم

مجموعه داده‌ای برای مسئله دسته‌بندی با کمک LDA در اختیار شما قرار گرفته شده است. دادگان به دو بخش آموزشی و تست در پوشه مجموعه داده تقسیم شده‌اند. ۲ ستون اول متغیرهای مستقل و ستون آخر متغیر وابسته است. معیار ارزیابی در این مسئله دقت [۲] است.
توجه: اگر چه الگوریتم LDA توانایی کاهش بعد را دارد، اما به خودی خود نمی‌تواند فرآیند دسته‌بندی را انجام دهد. به این منظور، در کتابخانه scikit-learn به صورت پیش‌فرض از دسته‌بند Bayes استفاده می‌شود. در صورتی که قصد پیاده‌سازی الگوریتم را دارید، می توانید از دسته‌بند Perceptron استفاده کنید.

۱. دادگان آموزشی را با درنظر گرفتن کلاس هر داده به کمک Scatter Plot رسم کنید. با توجه به شکل، چه عملکردی از LDA انتظار دارید؟ توضیح دهید.

۲. داده‌ها را بر اساس مجموعه داده آموزش استاندارد سازی کنید (از مجموعه داده تست استفاده نکنید).

$$X_{standard} = \frac{X - E[X]}{\sqrt{Var(X)}}$$

۳. استخراج ویژگی چند جمله‌ای با درجه‌های ۵، ۱۰، ۱۵، ۲۰ و ۲۵ را بر روی داده‌ها انجام دهید. سپس به کمک LDA در هر کدام از این حالات، داده‌ها را به فضایی با **۲ بعد** کاهش دهید و سپس از دسته بند برای جداسازی داده‌ها استفاده کنید. همچنین بایستی خطا آموزشی و ارزیابی را با استفاده از روش Repeated 5 Fold Cross-Validation با تکرار ۱۰ گزارش دهید. سپس نتایج را در Boxplot رسم کنید (از مجموعه داده تست استفاده نکنید).

۴. با توجه به Boxplot، مدل‌ها را از دیدگاه Overfit و Underfit بررسی کنید. همچنین با توجه به نتایج، آیا می‌توان تنظیمات بهتری ارائه داد؟ توضیح دهید.

۵. تنظیم بهنیه را تعیین کنید و مدل با تنظیم بهینه را بر روی داده تست اجرا کنید و نتیجه را در فایلی با نام prediction.csv ذخیره کنید. توجه کنید که هر سطر بایستی پیش‌بینی سطر متناظر با داده تست باشد.

۶. (**امتیازی**) نتایج بدست آمده در قسمت ۳ را به کمک Scatter Plot رسم کنید. توجه کنید که علاوه بر مشخص کردن نمونه‌های مرتبط با هر کلاس، نمونه‌هایی را که به اشتباه تشخیص داده شده‌اند نمایش دهید. به نظر شما دلیل پیش‌بینی اشتباه این نمونه‌ها چیست؟

[2] Accuracy