به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری پنجم یادگیری

ماشین

دانشگاه تهران

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.
4. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.
5. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW5_StudentNumber داشته باشد.
6. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**10 نمره امتیازی**) می توانید کسب کنید.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل elahe.bvakili97@gmail.com، سوال خود را مطرح کنید.

سوال اول (15 نمره)

1. آیا model Selection و model assessment یکی هستند؟

2. یکی از روش هایی را که با استفاده از آن میتوانیم با کمترین خطای generalization مدل خود را انتخاب کنیم را توضیح دهید.

3. در مواقعی که تعداد دادههای کمی در دیتاست داریم، برای ارزیابی و انتخاب مدل از چه روش هایی میتوانیم استفاده کنیم؟توضیح دهید.

4. یک متریک برای ارزیابی مدل خود نام برده و آن را توضیح دهید.

سوال دوم (15 نمره)

1. هدف از انجام feature selection چیست؟ چرا تمام ویژگی ها را به الگوریتم یادگیری ماشین خود نمی‌دهیم تا تصمیم بگیرد کدام ویژگی ها را انتخاب کند؟
2. روش fisher’s score را برای هدف انتخاب ویژگی ها توضیح دهید.

سوال سوم (15 نمره)

در این سوال قرار است با دیتاست wine که فایل آن قرار داده شده است، کار کنیم. این دیتاست شامل 12 عدد ویژگیست و قرار است طبقه بندی را روی 2 گروه انجام دهد. حال با توجه به این دیتاست، مراحل خواسته شده ی زیر را انجام دهید. دقت فرمایید که در انجام این سوال برای اجرای فقط الگوریتم ها قادر به استفاده از کتابخانه های آماده نیستید.

1. ابتدا الگوریتم Sequential Forward Selection را شرح داده سپس برای داده های ذکرشده پیاده سازی نمایید. این روش را تا جایی ادامه دهید که تنها دو ویژگی باقی بماند. سپس این دو ویژگی را با لیبل های متناظر برای هر wine رسم کنید.

2. در ادامه روش Recursive Feature Elimination را توضیح داده و سپس آن را بر روی این دیتاست پیاده سازی کنید. مانند قسمت بالا، این روش را نیز تا جایی ادامه دهید که تنها دو ویژگی باقی بماند. سپس این دو ویژگی را با لیبل های متناظر برای هر wine رسم کنید.

3. برای هر دو الگوریتم بالا بهترین ویژگی ها و همچنین نمودار خطا بر حسب تعداد ویژگی ها را رسم نمایید. سپس دقت و سرعت دو الگوریتم را مقایسه نمایید و توضیحات لازم را ارائه دهید.

سوال چهارم (20 نمره)

کاهش ابعاد با استفاده از PCA تکنیک متداولی برای فشرده کردن تصاویر است. توضیح دهید این روش چگونه میتواند دقت مارا برای تسک های متفاوت در کاربرد عکس ها افزایش دهد. مثال بزنید.

1. دیتاست FACES داده شده را در نظر بگیرید. مقادیر ویژه از PCA را به ترتیب کاهشی رسم نمایید و بیان نمایید که چگونه میتوان تعداد کامپوننت مناسب را در فرآیند فشرده سازی تشخیص داد؟

2. 4 مقدار ویژه اول و 4 مقدار ویژه نهایی را برای یک کلاس دلخواه نشان دهید و تحلیل کنید که این تصاویر بیانگر چه میباشند؟

3. حال طبقه بند KNN را با $k = 1,2$ را یک بار برا داده های کاهش بعد یافته و یک بار بر داده های خالص اعمال کنید و CCR و ماتریس کانفیوژن را گزارش نمایید و مقایسه نمایید.

4. اکنون مقدار کامپوننت تابع PCA را متغیر گرفته و CCR مربوط به طبقه بند نزدیکترین همسایه را برحسب تعداد کامپوننت PCA رسم و تحلیل کنید.

سوال پنجم (20 نمره)

1. ماتریس پراکندگی درون کلاسی و بین کلاسی را توضیح داده و سپس برای دیتاست داده شده محاسبه کرده و روش LDA را پیاده سازی نمایید.

2. ماتریس جدایی پذیری $(S_w^{-1}S_B)$ را حساب کرده و مقادیر ویژه ی آن را در قالب نزولی رسم نمایید.

3. در یک نمودار مقدار $trace(S_w^{-1}S_B)$ را نسبت به تعداد ویژگی ها رسم نمایید و در مورد تاثیر تعداد ویژگی ها بر آن بحث کنید.

4. بر اساس دو بخش پیشین تعداد بهینه ویژگی ها را بیابید. بعد از تصویر کردن داده به زیرفضای جدید، طبقه بند بهینه Bayes Naïve را با تخمین پارامتری گوسی اعمال کنید و مقدار CCR و ماتریس کانفیوژن را گزارش کنید.

5. طبقه بند ذکر شده را بدون LDA روی داده اعمال کنید و نتایج را با یکدیگر مقایسه کنید.

سوال ششم (15 نمره)

یکی از ایده های اصلی در clustering استفاده از فاصله ی بین نقاط است. آیا این روش همیشه جواب می دهد؟ در چه شرایطی این روش میتواند نتیجه ی منفی دهد؟

الگوریتم DBSCAN را توضیح دهید. و همچنین توضیح دهید در کدام دسته از الگوریتم های clustering قرار میگیرد. در ادامه تفاوت آن را با الگوریتم OPTICS شرح دهید.

سوال هفتم (10 نمره)

در داده های زیر که شکل داده ها توصیف کننده دسته های مختلف است، مولفه اول LDA و مولفه اول PCA را رسم کرده و چرایی انتخاب خود را توصیف کنید.