به نام خدا

دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران

پروژه پایانی

درس: یادگیری ماشین **مهلت ارسال: ۱۳۹۷/۱۱/۰۲**

به منظور نگارش گزارش تمرین صرفا از فایل Template ارائه شده استفاده نمایید و کلیه نتایج را در آن تایپ کرده و از تغییر دادن فونت، اندازه حاشیه و ... خودداری کنید. همچنین ساختار جداول نتایج را از فایل صورت تمرین (فایلی که در حال مطالعه آن هستید) برداشته تا گزارش تمامی دانشجویان یکسان و قابل مقایسه باشد. گزارشاتی که موارد تذکر داده شده را رعایت نکنند مورد بررسی قرار نخواهند گرفت.

"برای ارسال گزارش صرفا از آدرس ایمیل کلاس استفاده نمایید"

هدف از این تمرین بررسی و مقایسه کارایی روش‌های مختلف دسته بندی در حل مسائل دو و چند کلاسه به همراه الگوریتم استخراج ویژگی PCA می‌باشد. روش‌های مورد بررسی عبارتند از:

- دسته بندی نزدیک‌ترین همسایگی (K-NN)
- دسته بندی نزدیک‌ترین همسایگی وزن دار (Distance Weighted K-NN)
- شبکه عصبی پایه شعاعی (RBF)
- درخت تصمیم با استفاده از نرم‌افزار WEKA (بدون استفاده از استخراج ویژگی)

کلیه مثال‌های معرفی شده از جمله مسائل محک بوده و در بسیاری از مقالات به منظور ارزیابی روش‌های مختلف مورد استفاده قرار گرفته‌اند. برای حل این تمرین چهار دسته داده مختلف مورد ارزیابی و سنجش قرار خواهند گرفت. این چهار مجموعه عبارتند از:

- مجموعه داده Glass
- مجموعه داده Pima Indians Diabetic
- مجموعه داده Statlog Heart
- مجموعه داده Breast Cancer

مجموعه داده Glass دارای ۲۱٤ نمونه با ۱۰ ویژگی و مشتمل بر ۷ دسته مختلف از انواع شیشه می‌باشد. خروجی این مجموعه داده در ستون ۱۱ با مقادیر ۱ الی ۷ مشخص شده است.

مجموعه داده Pima Indians Diabetic دارای ۷٦۸ نمونه و مشتمل بر دو کلاس بیمار و سالم می‌باشد. تعداد نمونه‌های بیمار در این مجموعه داده ۲٦۸ نمونه بوده و مابقی سالم هستند. این داده دارای ۸ ویژگی ورودی و یک خروجی می‌باشد.

مجموعه داده Statlog Heart دارای دو کلاس بیمار و سالم می‌باشد. تعداد نمونه‌های بیمار در این مجموعه داده ۱۲۰ نمونه و تعداد نمونه‌های سالم ۱۵۰ مورد می‌باشند. این داده دارای ۱۳ ویژگی ورودی و یک خروجی می‌باشد. در ستون آخر دو داده اخیر (ستون شماره ۹ و ستون شماره ۱٤) عدد ۰ نشان دهنده نمونه سالم و عدد ۱ نشان دهنده وضعیت بیمار می‌باشد. در جدول ۱ خصوصیات این چهار دسته داده آورده شده‌اند.

مجموعه داده Breast Cancer دارای دو دسته خوش‌خیم و بدخیم است. تعداد نمونه‌های خوش‌خیم در این مجموعه داده ۳۵۷ نمونه و تعداد نمونه‌های بدخیم ۲۱۲ مورد می‌باشند. این داده دارای ۳۰ ویژگی ورودی و یک خروجی می‌باشد که در ستون آخر (ستون شماره ۳۱) عدد ۰ نشان دهنده نمونه خوش‌خیم و عدد ۱ نشان دهنده وضعیت بدخیم می‌باشد. در جدول ۱ خصوصیات این چهار دسته داده آورده شده‌اند.

جدول ۱: ویژگی‌های مجموعه داده‌های استفاده شده

Data set	#Pattern	#Attributes	#Class
Glass	214	10	7
Pima Indians Diabetic	768	8	2
Statlog Heart	270	13	2
Breast Cancer	569	30	2

کلیه داده‌های مورد نیاز برای حل تمرین به پیوست ارسال شده‌اند. برای استفاده از این داده‌ها در نرم‌افزار MATLAB از دستورات زیر می‌توانید استفاده کنید:

load WDBC.dat یا load Heart.dat یا load Glass.dat یا load Pima.dat

در حل کلیه مسائل با استفاده از روش استخراج ویژگی PCA، حداکثر تعداد ویژگی‌های استخراج شده برای هر مجموعه داده را حداکثر ۵ ویژگی در نظر بگیرید. انتخاب تعداد کمتر ویژگی بر اساس شرایط کارکرد سیستم طبقه بندی مجاز است.

به منظور دسته بندی داده‌های فوق از روش 5-Fold Cross Validation استفاده نمایید، بدین ترتیب که هر مجموعه داده را به ۵ قسمت تقسیم کرده و در هر اجرا از ٤ قسمت به عنوان داده آموزش و از یک قسمت باقی مانده به عنوان داده آزمون استفاده نمایید و در کل این عمل را ۵ بار تکرار کنید تا هر کدام از دسته‌ها به عنوان داده آزمون مورد ارزیابی قرار گیرند. با توجه به این مساله که امکان تقسیم مجموعه داده‌های فوق به دسته‌های مساوی وجود ندارد، برای هر مجموعه داده ٤ دسته مساوی و یک دسته با اندازه بزرگتر یا کوچکتر استفاده نمایید.

برای هر چهار مجموعه داده روش‌های دسته بندی KNN و DWK-NN را مورد استفاده قرار داده و تعداد همسایگی را در دو حالت ۳ و ۵ مورد بررسی قرار دهید.

به منظور پیاده سازی شبکه RBF بر روی هر چهار دسته داده محک، از ۲ روش مختلف زیر استفاده نموده و به منظور حصول بهترین نتیجه تعداد نورون‌های لایه میانی را برای هر دسته داده تعیین کنید:

- استفاده از الگوریتم آموزش ۱ در اسلاید درس
- استفاده از Toolbox نرم افزار MATLAB با دستور newrb

به منظور پیاده سازی درخت تصمیم از نرم‌افزار WEKA استفاده نمایید. انتخاب نوع الگوریتم آموزش برای ایجاد درخت به دلخواه و بر اساس موارد موجود در نرم‌افزار انجام شود. نرم‌افزار WEKA را می‌توانید از آدرس اینترنتی http://www.cs.waikato.ac.nz/ml/weka دریافت کنید.

برای هر مجموعه داده با استفاده از هر روش، ۵ اجرای مستقل انجام داده و برای مجموعه داده Glass تنها مقادیر میانگین دقت دسته بندی، بهترین و بدترین پاسخ در ۵ اجرای مستقل به همراه انحراف معیار استاندارد و متوسط زمان اجرای هر روش (با استفاده از دو دستور tic در ابتدا و toc در انتهای برنامه) مطابق جدول ۲ گزارش شود.

برای مجموعه داده‌های Pima Indians Diabetic، Statlog Heart و Breast Cancer معیارهای مورد بررسی شامل میانگین دقت دسته بندی، انحراف معیار استاندارد، معیار MCC، معیار F1 Score، معیار Specificity ، معیار Sensitivity و متوسط زمان اجرا در ۵ اجرا خواهند بود که در جداولی مشابه جداول ۳ الی ۵ نمایش داده خواهند شد. بر اساس موارد ارائه شده دسته‌بندی کننده بهینه را معرفی نمایید.

جدول ۲ : نتایج حاصل از دسته بندی داده Glass با استفاده از روش‌های مختلف

Method	Mean	Best	Worst	STD	Time (s)
3NN	xxx	xxx	xxx	xxx	xxx
5NN	xxx	xxx	xxx	xxx	xxx
DW-3NN	xxx	xxx	xxx	xxx	xxx
DW-5NN	xxx	xxx	xxx	xxx	xxx
RBF-Code	xxx	xxx	xxx	xxx	xxx
RBF-Toolbox	xxx	xxx	xxx	xxx	xxx
Decision Tree	xxx	xxx	xxx	xxx	xxx

جدول ۳ : نتایج حاصل از دسته بندی داده Pima Indians Diabetic با استفاده از روش‌های مختلف

Method	Mean	STD	MCC	F1	Specificity	Sensitivity	Time (s)
3NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
5NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
DW-3NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
DW-5NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
RBF-Code	xxx	xxx	xxx	xxx	xxx	xxx	xxx
RBF-Toolbox	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Decision Tree	xxx	xxx	xxx	xxx	xxx	xxx	xxx

جدول ۴ : نتایج حاصل از دسته بندی داده Statlog Heart با استفاده از روش‌های مختلف

Method	Mean	STD	MCC	F1	Specificity	Sensitivity	Time (s)
3NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
5NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
DW-3NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
DW-5NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
RBF-Code	xxx	xxx	xxx	xxx	xxx	xxx	xxx
RBF-Toolbox	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Decision Tree	xxx	xxx	xxx	xxx	xxx	xxx	xxx

جدول ۵ : نتایج حاصل از دسته بندی داده Breast Cancer با استفاده از روش‌های مختلف

Method	Mean	STD	MCC	F1	Specificity	Sensitivity	Time (s)
3NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
5NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
DW-3NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
DW-5NN	xxx	xxx	xxx	xxx	xxx	xxx	xxx
RBF-Code	xxx	xxx	xxx	xxx	xxx	xxx	xxx
RBF-Toolbox	xxx	xxx	xxx	xxx	xxx	xxx	xxx
Decision Tree	xxx	xxx	xxx	xxx	xxx	xxx	xxx

موفق باشید