بسمه تعالی

تمرین سوم

دانشکده مهندسی صنایع، دانشگاه صنعتی امیرکبیر

درس: داده‌کاوی در سلامت، موعد تحویل: روز ۲۰ دی ۱۴۰۲

مجموعه داده pima-indian-diabetes با ۷۶۸ داده و ۸ ویژگی از مخزن داده‌های استاندارد دانشگاه کالیفرنیا (UCI Machine Learning Repository) مورد نظر است.

الف- ابتدا داده‌ها را پیش‌پردازش کرده و آنها را با روش مین‌ماکس استاندارد نمایید.

ب- با استفاده از روش «جستجوی روبه جلو»، بهترین ترکیب ویژگی‌ها را بدست آورید. از روش «نزدیکترین همسایگی» جهت دسته‌بندی استفاده کنید. بدین‌منظور ۶۵ درصد داده‌ها را برای آموزش و الباقی را برای تست به‌کار ببرید.

ج- حال با استفاده از «روش تحلیل مولفه اصلی»، ویژگی‌ها را کاهش دهید. مولفه‌های اصلی را بیان کنید. رویکرد خود را برای انتخاب تعداد مولفه‌های اصلی توضیح دهید. مجددا از روش «نزدیکترین همسایگی» جهت دسته‌بندی استفاده کنید. ۶۵ درصد داده‌ها را برای آموزش و الباقی را برای تست به‌کار ببرید. ماتریس در هم ریختگی را تشکیل داده و نتایج را بر اساس شاخص‌های مختلف (دقت، precision, recall, F-measure, AUC) ارائه دهید.

د- داده‌ها را با روش‌های «شبکه‌عصبی پرسپترون چند لایه» دسته‌بندی کنید. از روش اعتبارسنجی چهاربخشی استفاده کنید و نتایج را به تفکیک ویژگی‌های بندهای «ب» و «ج» بر اساس شاخص‌های مختلف (دقت، precision, recall, F-measure, AUC) ارائه دهید.

ه- حال داده‌ها را با «شبکه عصبی با تابع محرک شعاعی» دسته‌بندی کنید. ۵۰ درصد داده‌ها را برای آموزش و الباقی را برای تست به‌کار ببرید(تعداد نرونهای لایه پنهان را با روش شبکه عصبی خودسازمانده بیابید). ماتریس در هم ریختگی را تشکیل داده و نتایج را به تفکیک ویژگی‌های بندهای «ب» و «ج» بر اساس شاخص‌های مختلف (دقت، precision, recall, F-measure, AUC) ارائه دهید.

موفق باشید.