امیرحسین پاشایی‌هیر ۹۷۳۱۰۱۳

آماده‌سازی داده:

بخش مربوط به آماده‌سازی داده و categorizer کردن داده‌ها در فایلcategorizer.py پیاده سازی شده‌اند. برای ستون‌های عددی (چه اعشاری چه صحیح) اعداد مربوط به این ستون در هر جفت dataset و unknown dataset استخراج شده و مرتبط شده و صدک ۳۳ و ۶۶ آنها به عنوان مرز تقسیم کردن انتخاب شده است. مقدار صدک اول A، مقدار صدک دوم B و مقدار صدک سوم C قرار داده شده است.

درخت تصمیم‌گیری:

این روش در فایل decision_tree.py پیاده سازی شده است.

برای پیاده سازی درخت تصمیم گیری ابتدا تابع خلوص پرسیده شده است. سپس با استفاده از تابع خلوص گرفته شده درخت ساخته شده است. روند ساخت درخت به این صورت است که بین تمام خاصیت‌ها غیر از خاصیت هدف، به ترتیب برای هر راس درخت، gain بدست آمده از اعمال این خاصیت بر این راس درخت محاسبه شده و بیشینه gain به عنوان خصوصیت مورد سوال مربوط به این راس انتخاب شده است.

همچنین برای افزایش دقت، رئوس با مجموعه داده‌ی کمتر از ۱۰ سطر یا ارتفاع بیشتر از ۳ (با شروع از ۰) هرس شده‌اند. (به روش pre pruning)

درخت ساخته شده در فایل DT.pdf قابل مشاهده می‌باشد. (دقت کنید قابل زوم می‌باشد عکس و تمام رئوس درخت پس از زوم مشخص می‌شوند)

دقت این روش به طور متوسط ۸۰ درصد است.

خروجی ستون جواب برای داده‌های تست به ترتیب فایل Dataset1_Unknown در فایل DT.txt نوشته شده است.

روش knn:

این روش در فایل knn.py پیاده سازی شده است.

برای پیاده سازی این روش تابع فاصله برای هر سطر مورد بررسی ساخته شده و بین k نزدیک‌ترین حالات ممکن، ۵ بار به صورت اتفاقی یک نقطه بین این k نقطه انتخاب شده و بین آن‌ها جواب اکثریت انتخاب شده است، به این صورت احتمال جواب براساس تعداد بار تکرار آن در k نقطه برتر به صورت تصاعدی بالا می‌رود.

دقت برای k=1,5,10 برابر با تقریبا ۹۹ درصد می‌باشد.

دقت برای k=100 برابر با ۹۸ درصد می‌باشد.

خروجی ستون جواب برای داده‌های تست به ترتیب فایل Dataset2_Unknown در فایل KNN{K}.txt نوشته شده است.

روش نایو بیز:

این روش در فایل naive_bayes.py پیاده سازی شده است.

برای پیاده سازی این روش احتمال هر مقدار در هر ستون بر اساس کلاس آن در ستون جواب به صورت on demand محاسبه شده و در یک دیکشنری cache شده است که سرعت پردازش بالا برود.

دقت این روش به صورت میانگین حدود ۸۰ درصد می‌باشد.

خروجی ستون جواب برای داده‌های تست به ترتیب فایل Dataset3_Unknown در فایل NB.txt نوشته شده است.