تمرین و پروژه شماره ۲ درس داده کاوی، پروژه تحویلی ۲

تکین جزایری، ۹۸۱۳۰۰۶ – امیرحسین رجبی، ۹۸۱۳۰۱۳

برای این پروژه از یکی از دیتاست های UCI به نام Bank Marketing Data Set استفاده شده است که فایل csv آن با نام bank-additional.csv ارسال شده است. این دیتاست مرتبط با یکی از بانک های پرتغال است و هدف آن پیشبینی آن است که آیا مشتری ای که اطلاعات آن در هر نمونه موجود است، حاضر به افتتاح حساب سپرده مدت دار هست یا نه. این دیتاست در مجموع دارای ۱۷ خصیصه (Attribute) می باشد.

الگوریتم IBk:

این الگوریتم یک پیاده سازی از الگوریتم k–نزدیکترین همسایه (k-nearest neighbor) است و می توانیم از بخش classifier.lazy در نرم افزار WEKA به این الگوریتم دسترسی داشته باشیم. برای داشتن k = 1 (IB1 – نزدیکترین همسایه)، با کلیک چپ بر روی فیلد Classifier، در فیلد مقابل KNN مقدار 1 را قرار می دهیم. نتیجه اجرای IB1 روی دیتاست فوق الذکر با روش 10-cross-validation در فایل IB1_model.model موجود است. نتیجه این اجرا توانست با دقت نزدیک به ۸۷٪ لیبل مناسب را تشخیص دهد. (البته دقت این الگوریتم در تشخیص موارد "yes" بسیار کمتر و تقریباً برابر با ۲۹٪ می باشد.)

الگوریتم LWL:

این الگوریتم در مقایسه با بیشتر الگوریتم های موجود که یک مدل سراسری (global) آموزش می دهند، کل دیتاست را به مجموعه های کوچک تر تقسیم می کند و توابع محلی (local functions) روی هر یک از این زیرمجموعه ها آموزش می دهد. این الگوریتم هم از بخش classifier.lazy قابل دسترسی است. اگر الگوریتم LWL را روی دیتاست فوق با روش 10-cross-validation اجرا کنیم، نتیجه ای با دقت ۸۹٪ بدست می آید که مدل تولید شدهٔ آن در فایل LWL_model.model موجود است. با این وجود، نتیجه حاصل از این الگوریتم قابل اطمینان نیست زیرا تقریباً به ازای تمام موارد، لیبل "no" به نمونه تعلق می گیرد و تشخیص موارد "yes" با دقتی ناچیزی انجام می شود.