به نام او...

تمرین دوم، یادگیری ماشین **اسفندماه ۱۳۹۴**

سوال اول

الف) آنتروپی این مجموعه را بر اساس متغیر هدف (برنامه آخر هفته) به دست آورید.

ب) بهره اطلاعات ویژگی‌ها را محاسبه کنید.

ج) بر اساس بهره اطلاعات، کدام ویژگی به عنوان ریشه برگزیده شده و چه تعداد نمونه توسط درخت حاصل به اشتباه دسته‌بندی می‌شود؟

هفته	آب و هوا	وضعیت مالی	کوییز	برنامه آخر هفته
۱	آفتابی	خوب	خیر	سینما
۲	آفتابی	خوب	بله	تنیس
۳	طوفانی	خوب	خیر	سینما
۴	بارانی	بد	خیر	سینما
۵	بارانی	خوب	بله	درس
۶	بارانی	بد	خیر	سینما
۷	طوفانی	بد	بله	سینما
۸	طوفانی	خوب	بله	خرید
۹	طوفانی	خوب	خیر	سینما
۱۰	آفتابی	خوب	بله	تنیس

سوال دوم

در این تمرین هدف آشنایی با دسته بندی کننده درخت تصمیم، با استفاده از ابزار وکا[1] می‌باشد.

آخرین نسخه این ابزار را می‌توانید از اینجا دانلود کنید.

بخش اول: آماده سازی مجموعه داده

در این قسمت لازم است تا مجموعه داده heart Disease را از این آدرس دانلود کنید. این مجموعه داده شامل ۲۷۰ نمونه از افرادی است که در دو دسته بیماران قلبی و سالم دسته‌بندی شده‌اند. برای استفاده از این مجموعه داده، لازم است تا آن را به فرمت arff.* که فرمت فایل‌های ورودی در وکا است تبدیل کنید.

برای تبدیل فایل داده به فرمت مورد نظر، لازم است header هایی که توصیف کننده نام و نوع هر ویژگی است به آن افزوده شود. در تعیین نوع ویژگی به اسمی[2] و عددی بودن ویژگی‌ها دقت کنید.

خروجی بخش اول: فایل arff متناظر با داده‌ها

بخش دوم: بارگذاری و بررسی داده‌ها

برای بارگذاری داده‌ها در نرم افزار وکا، پس از باز کردن قسمت Weka Explorer ، از طریق نوار ابزار preprocess داده‌ها را بارگذاری کنید. پس از بارگذاری مجموعه داده، از قسمت Visualize All ، نمودار پراکندگی دو کلاس به ازای هر ویژگی را نمایش دهید.

سه ویژگی ای را که از نظر شما بهتر می‌توانند برچسب کلاس را توصیف کنند، با ذکر دلیل مشخص کنید.

خروجی بخش دوم: نمایش نمودار پراکندگی کلاسی ویژگی ها ، تعیین و توصیف ویژگی‌های مناسب برای دسته‌بندی

بخش سوم: ساخت درخت تصمیم

برای ساخت درخت تصمیم با استفاده از داده‌های آموزشی، از نوار ابزار classify ، دسته‌بندی کننده j48 را که نوعی الگوریتم برای یادگیری درخت تصمیم می‌باشد انتخاب کنید. برای تنظیم پارامترهای الگوریتم انتخاب شده با کلیک بر روی عنوان آن می‌توانید پارامترهای آن را تغییر دهید.

---

[1] weka

[2] nominal

الف) با انتخاب داده‌های آموزشی از قسمت Test Options، confusion Matrix به دست آمده برای داده‌های آموزشی را نشان دهید و معیارهای مختلف ارزیابی را برای داده‌های آموزشی گزارش کنید. (Precision، TPR، ...)

ب) معیار F_measure (F_Score) که به صورت رابطه زیر تعریف می‌شود را در نظر بگیرید:

$$F_\beta = (1+\beta^2)\frac{precision.\text{recall}}{(\beta^2.precision)+recall}$$

با استفاده از confusion Matrix قسمت الف برای $\beta = 0.5,1,2$ معیار F_measure را به صورت دستی برای کلاس یک محاسبه کنید و تاثیرات پارامتر β را تحلیل کنید. مقدار F1_measure را با خروجی بدست آمده در وکا مقایسه کنید.

ج) با کلیک بر روی مدل ساخته شده، درخت حاصل را نمایش دهید.

د) با تغییر تعداد داده‌ها آموزشی از قسمت percentage split، ۹ درخت تصمیم با استفاده از ۱۰و۲۰ و ....۹۰ درصد داده‌های آموزشی بسازید و نمودار خطای تست را براساس تعداد نمونه‌های آموزشی رسم کرده و تحلیل کنید.

ه) با تغییر پارامتر unpruned اختلاف اندازه‌ی درخت‌های هرس شده و نشده بخش قبل را مشخص و نتیجه آن را مقایسه کنید.

و) با استفاده از ۶۰ درصد داده‌ها، درخت تصمیم را برای دو حالت هرس شده و هرس نشده آموزش دهید و خطای تست را مقایسه کنید. با استفاده از تمام داده‌های آموزشی درخت تصمیم هرس شده و نشده را آموزش داده و خطای داده‌های آموزشی را برای دو حالت مقایسه کنید. تحلیل نهایی خود را از این بخش بیان کنید.

ز) با تغییر متغیر هدف، دسته‌بند را برای پیش‌بینی جنسیت افراد آموزش داده و دقت دسته‌بندی را برای داده‌های آموزشی با روش 10-Fold Cross Validation به دست آورید.

شیوه‌ی تحویل تمرین: تا ساعت ۲۳:۵۵ روز جمعه ۲۸ اسفند فرصت دارید تا تمرین را در مودل بارگذاری کنید. (ضمنا در صورت داشتن تاخیر حداکثر ۲ هفته‌ای، نمره‌ی پروژه از ۸۰ درصد لحاظ خواهد شد.) فایل pdf مربوط به گزارش تمرین را به همراه فایل arff مجموعه داده، در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود قرار دهید. (برای مثال 93131130.rar)

در صورت وجود هر گونه سوال می‌توانید از طریق ایمیل با یکی از تدریس‌یاران درس در ارتباط باشید.

MR.Molavi@gmail.com , Marjan.Moodi@gmail.com , NavidFumani@gmail.com