درس یادگیری ماشین

مدرس: دکتر سامان هراتی‌زاده

نیم‌سال اول ۱۴۰۱-۱۴۰۰

تمرین شماره‌ی ۱

مهلت ارسال تمرین: ۱۴۰۰/۹/۳

الگوریتم: oneR, Prism, ID3

تحویل آنلاین: اعلام خواهد شد

هدف: ارائه‌ی یک مدل پیش‌بینی برای تشخیص بیماری کرونا

داده‌ها : مجموعه داده‌ی پیوست شده شامل علائم و ویژگی‌هایی است که در احتمال ابتلا به بیماری کرونا موثر هستند. در این مجموعه داده ۵ علامت سرفه، تب، گلودرد، تنگی نفس و سردرد و ویژگی‌هایی نظیر جنسیت، سن بالای ۶۰ و دلیل تست (تماس با فرد بیمار، سفر خارجی و غیره) برای ۴۰۰۰ نمونه گزارش شده است. هدف ما پیش‌بینی نتیجه‌ی تست کرونا (Corona Result) است. تمامی ویژگی‌های این مجموعه داده اسمی هستند.

۱.الگوریتم oneR را برای پیش‌بینی ابتلا به بیماری کرونا پیاده‌سازی کنید. قواعد ایجاد شده توسط این مدل را همراه با support و confidence آنها در داده آموزش نمایش دهید. نتایج ارزیابی مدل را برای بهترین قاعده‌ی استخراج شده گزارش کنید.

۲. با پیاده‌سازی الگوریتم Prism، مدلی برای تشخیص ابتلا به بیماری کرونا ارائه دهید. برای الگوریتم شرط توقف بر اساس تعداد ویژگی اضافه کنید و عدد بهینه‌ی تعداد ویژگی را برای حل مشکل بیش‌برازش را پیدا کنید. نمودار دقت بر حسب تعداد ویژگی ها را رسم کنید. قواعد ایجاد شده توسط این مدل را نمایش دهید. نتایج ارزیابی مدل را گزارش کنید.

۳. الگوریتم ID3 را برای تشخیص بیماری کرونا پیاده سازی کنید. درخت را حداکثر تا ارتفاع ۴ ادامه دهید. قواعد استخراج توسط این الگوریتم را را به همراه نتایج ارزیابی آن را گزارش کنید. (نیازی به ترسیم درخت خروجی به صورت اتوماتیک و گرافیکی توسط سیستم نیست. اما باید امکان پیماش درخت از ریشه تا برگ وجود داشته باشد.)

۴.(امتیازی) با توجه به اینکه با یک مسئله‌ی طبقه‌بندی Imbalanced رو به رو هستیم، به نظر شما Imbalanced بودن دیتا چه مشکلاتی را برای یک مدل پیش‌بینی ایجاد می‌کند. راه حل شما برای حل این مسئله چیست؟ راه حل خود را با روش‌های موجود برای حل این معضل مقایسه کنید.

۵.(امتیازی) الگوریتم درخت تصمیم را با کتابخانه‌ی scikit-learn پیاده‌سازی کنید و درخت حاصل از آن را ترسیم نمایید و نتایج آن‌را با مدل پیاده‌سازی شده توسط خودتان مقایسه کنید.

آ) تمامی مدل‌ها را با مترهای accuracy، precision، recall و f-measure ارزیابی و میانگین آن‌ها را برحسب 5-Fold-CrossValidation اعلام کنید. [1]

- Precision = TruePositives / (TruePositives + FalsePositives)
- Recall = TruePositives / (TruePositives + FalseNegatives)
- F-measure = (2 * Precision * Recall) / (Precision + Recall)

ب) نتایج بدست آمده از پیاده‌سازی این الگوریتم‌ها را در گزارش خود شرح دهید. گزارش شما باید شامل توضیح الگوریتم و شرح مختصری از قسمت های مختلف کد شما باشد. دقت داشته باشید که بخشی از نمره شما به گزارش شما تعلق می گیرد، بنابراین زمان کافی به آن اختصاص دهید.

- برای پیاده سازی می‌توانید از زبان پایتون و کتابخانه‌های pandas، NumPy، SciPy، math و یا کتابخانه های مشابه استفاده کنید. دقت داشته‌باشید که استفاده از کتابخانه‌ی آماده برای سوالاتی که بایستی الگوریتم را پیاده‌سازی کنید مجاز نیست.

- الگوریتم‌ها را خودتان پیاده‌سازی کنید. به کدهای کپی‌شده از گیت‌هاب و دوستانتان نمره‌ای تعلق نمی‌گیرد. درصورت مشاهده‌ی کدهای مشابه بین چند نفر به هیچ یک از افراد نمره‌ای تعلق نخواهد گرفت.

- کد خود را کامنت گذاری کنید. برای کامنت گذاری مناسب و تمیز بودن کد امتیاز مثبت در نظر گرفته خواهد شد.

- در صورت تاخیر تا ٢٤ ساعت ٢٠% تا ۴۸ ساعت ۴۵% و بیش از ۴۸ ساعت تمامی نمره تمرین را از دست خواهید داد.

- برای تحویل فایل های مربوطه را در یک فایل زیپ با نام ML_StudentID_FullName_HW #1 قرار دهید.

- در صورت وجود هرگونه سوال یا ابهام می‌توانید با دستیاران آموزشی از طریق Piazza یا گروه WhatsApp در ارتباط باشید.

[1] پیشنهاد می شود برای K-Fold-CrossValidation تابعی بنویسید که برای تمرین های بعدی هم قابل استفاده باشد.