به نام او...

پروژه‌ی سوم درس یادگیری ماشین آماری دسته‌بندی (مدل‌های احتمالاتی گرافی)

شرح مجموعه‌ی دادگان

برای این پروژه، مجموعه داده‌ی مربوط به «بیماری قلبی» است که از بین مجموعه دادگان سایت UCI انتخاب شده‌است. در این مجموعه داده، اطلاعات از ۴ مکان مختلف جمع‌آوری شده است که در مجموع اطلاعات مربوط به ۹۲۰ بیمار وجود دارد. برای هر بیمار ۷۶ ویژگی در نظر گرفته شده‌است که از این بین ۱۴ ویژگی به عنوان ویژگی اصلی انتخاب شده‌اند. متغیر هدف دارای ۵ مقدار مختلف است که مقدار ۰ به معنای عدم وجود بیماری است و چهار مقدار دیگر به معنای نوعی از بیماری قلبی هستند.

این مجموعه داده را می‌توانید از آرشیو سایت UCI دریافت کنید.

http://archive.ics.uci.edu/ml/datasets/Heart+Disease

هدف پروژه

هدف از این پروژه پیش‌بینی وجود یا عدم وجود بیماری (نوع بیماری مد نظر نیست.) یک فرد با توجه به ویژگی‌های سلامتی اوست (منظور ۱۳ ویژگی اصلی موجود در مجموعه داده است). برای مدل‌سازی مساله لازم است تا از مدل‌های گرافی احتمالاتی مختلف استفاده شود.

شرح پروژه

الف) ابتدا لازم است با بررسی مجموعه داده، پیش‌پردازش‌های مختلفی که ممکن است مورد نیاز باشد را انجام دهید. برای مثال پیش‌پردازش‌هایی مانند حذف مقادیر نامشخص[1] و گسسته‌سازی متغیرهای پیوسته ممکن است مفید باشند. در گزارش خود پیش‌پردازش‌های صورت گرفته را توضیح دهید.

ب) با رسم نمودار Scatter مربوط به هر یک از ویژگی‌های اصلی، به طور شهودی در مورد میزان ارتباط هر ویژگی با متغیر هدف و میزان جداکنندگی آن‌ها بحث کنید.

ج) مدل بیز ساده[2] را با در نظر گرفتن تمام متغیرهای اصلی پیاده‌سازی کنید. با استفاده از روش Leave one out cross validation دقت مدل خود را گزارش کنید.

د) با بررسی زیرمجموعه‌های مختلف، به انتخاب خودتان چند زیر مجموعه (حداقل دو زیرمجموعه) از ویژگی‌ها را انتخاب کنید و مدل بیز ساده را با استفاده از این ویژگی‌ها ایجاد کنید. (دلیل انتخاب این زیرمجموعه‌ها را بیان

[1] Missing Values
[2] Naïve Bayes

کنید.) مدل خود را با استفاده از روش بیان شده در قسمت قبل ارزیابی کنید و نتایج قسمت قبل را با نتایج مدل‌های این قسمت مقایسه کنید.

ه) با بررسی ویژگی‌های موجود در مجموعه داده و استفاده از دانش خبره (می‌توانید از دانش خودتان استفاده کنید!!)، چند مدل گرافی پیشنهاد دهید (حداقل ۲ مدل). مدل‌های احتمالاتی خود را در گزارش رسم کنید و در مورد علت انتخاب مدل توضیح دهید. لیست احتمالات شرطی لازم برای هر یک از مدل‌ها را در گزارش بیان کنید و در مورد نحوه‌ی تصمیم‌گیری مدل برای یک داده‌ی آزمایشی توضیح دهید. مدل‌های خود را پیاده‌سازی کنید و مدل‌ها را با استفاده از روش بیان شده در قسمت قبل ارزیابی کنید. تمامی نتایج به دست آمده در پروژه را با یکدیگر مقایسه کنید.

پ ن. نیازی به استفاده از تمامی متغیرهای اصلی در مدل‌های پیشنهادی نیست. می‌توانید برای راحتی کار از تعداد کمتری از آنها در مدل‌های ارائه شده استفاده کنید.

فرمت گزارش:

گزارش بایستی به زبان فارسی و در قالب فایل PDF باشد. در گزارش تحلیل و نتیجه‌گیری خود را در رابطه با هر بخش به شکل مختصر بیان فرمایید.

فایل گزارش خود را به شکل « Project3_StdNum.pdf» نامگذاری کنید. (مانند Project3_8931064.pdf)

فرمت کدها:

برای پروژه بایستی پیاده‌سازی در یکی از محیطهای MATLAB، R یا Python تهیه شود.

تمامی کدهای پیاده‌سازی پروژه باید دارای شرح کامل درون کد باشد. (Well Commented)

نحوه تحویل:

فایلهای کد و گزارش خود را که طبق فرمتهای فوق تهیه شده‌اند، در قالب یک فایل فشرده در سایت درس بارگذاری نمایید.

فایل فشرده را به شکل «P3_StdNum» نامگذاری کنید. (مانند P3_8931064)

مهلت ارسال تمرین ساعت ۲۳:۵۵ دقیقه‌ی روز چهارشنبه مورخ ۲۰ بهمن ماه می‌باشد.

پروژه شامل تحویل حضوری نیز خواهد بود.

هر گونه سوال در مورد پروژه را می‌توانید از طریق ایمیل به آدرس AUT.SMLf16@gmail.com بیان نمایید.