پروژه دوم درس یادگیری ماشین آماری (اختیاری)

هدف از این پروژه آشنایی با مدل های احتمالاتی گرافی و نحوه ی استفاده از آن در کاربرد دسته بندی می باشد. آزمایش های خواسته شده را با دقت انجام داده و نتایج هر آزمایش را به همراه نتیجه گیری و استدلال خود در گزارش ارائه نمایید.

مجموعه داده

در این پروژه از مجموعه داده‌ی پیش‌بینی درآمد استفاده می‌شود. درآمد یک فرد به صورت دو کلاس بیش‌تر یا کم‌تر مساوی با ۵۰ هزار دلار در نظر گرفته می‌شود. تعداد ۱۴ ویژگی عددی و غیر عددی از افراد در مجموعه داده موجود می‌باشد. این مجموعه داده از لینک زیر قابل دریافت می‌باشد.

http://archive.ics.uci.edu/ml/datasets/Adult

شرح پروژه

الف) پیش‌پردازش‌های مورد نیاز مانند پر کردن مقادیر نامشخص و گسسته‌سازی ویژگی‌های پیوسته را انجام داده و در گزارش خود توضیح دهید.

ب) نمودار Scatter مربوط به هر یک از ویژگی‌ها (با توجه به کلاس مورد نظر) را رسم کرده و در مورد قدرت جداکنندگی هرکدام توضیح دهید.

ج) مدل بیز ساده را با در نظر گرفتن تمام متغیرهای اصلی آموزش داده و دقت مدل را با استفاده از 10 Fold Cross Validation گزارش کنید.

د) حداقل سه زیرمجموعه از ویژگی‌ها را انتخاب کرده (با ذکر دلیل) و مدل بیز ساده را آموزش داده و همانند قسمت ج دقت حاصل را گزارش کرده و با آن مقایسه نمایید.

ه) ویژگی‌های مجموعه داده را بررسی کرده و حداقل دو مدل گرافی را با کمک گرفتن از دانش خبره یا تحلیل خودتان ساخته و دلایل انتخاب هر مدل را نیز توضیح دهید. دقت مدل‌ها را همانند بخش‌های قبل گزارش کرده و نتایج را با یکدیگر و با بخش‌های قبل مقایسه نمایید. مدل‌های ساخته شده را در گزارش خود رسم کرده و احتمال‌های شرطی لازم برای هرکدام را بیان کنید. در هر مدل حداقل از ۷ ویژگی استفاده کنید.

و) با استفاده از معیار BIC و یک روش جست و جو ساختارهای مختلف مدل گرافی را جست و جو کنید. پس از یافتن ساختار مناسب توسط معیار امتیاز BIC، دقت مدل را همانند بخش‌های قبل به دست آورده و گزارش نمایید. نتایج را تحلیل کرده و با قسمت ه مقایسه کنید.

این پروژه اختیاری است و حداکثر یک و نیم نمره برای آن منظورمی گردد.

نکته : کلیه روابط ریاضی استفاده شده در هر یک از بخش های مختلف پروژه را به شکل دقیق ذکر نمایید.

فرمت گزارش:

گزارش باید به زبان فارسی و در قالب PDF باشد. در گزارش تحلیل و نتیجه‌گیری خود را در رابطه با هر بخش بیان کنید.

فایل گزارش خود را به شکل Project2_StdNum.pdf نامگذاری کنید.

از کتابخانه‌های آماده مانند pgmpy می‌توانید استفاده کنید.

فرمت کدها:

از یکی از محیط‌های برنامه‌نویسی Matlab، R و یا Python برای پیاده سازی پروژه استفاده نمایید.

کدهای خود را به تفکیک بخش های مختلف آزمایش بنویسید و در کلیه بخش ها کامنت کافی قرار دهید.