پروژه درس مباحث ویژه ۱ تاریخ تحویل: ۱۴۰۰/۱۰/۳۰ ساعت ۱۵

اطلاعات مربوط به فایل iris.csv بخوانید و عملیات زیر را پیاده سازی نمائید؟ لازم به ذکر است که در این پروژه باید کلاس های زیر را طراحی و استفاده نمائید لازم به ذکر است پیاده سازی این کلاس‌ها برعهده دانشجو می‌باشد:

ClearMissData برای قسمت (۱) که وظیفه حذف داده های اشتباه را دارد

SelectData برای قسمت 2.a که وظیفه ایجاد مجموعه آموزش و آزمون را دارد.

NormalizeData برای قسمت 2.c وظیفه نرمال سازی هر ستون را دارد.

CreateNewFeature برای قسمت 2.d و 2.g.i و 2.g.ii وظیفه ساخت ویژگی جدید با ترکیب های خواسته شده را دارد.

ShowInfo برای قسمت 2.e، 2.f، 2.h و ۳ نمایش گرافیک خروجی می باشد.

CalcSimilarites برای 2.h وظیفه محاسبه شباهت بین مجموعه آزمون و آموزش را دارد.

۱) اطلاعات مربوط به فایل را نمایش دهید در نمایش اگر سطری اطلاعات اشتباهی داشته آنها را نادیده بگیرید.
۲) در ادامه مراحل زیر را برای ۱۰ مرتبه تکرار کنید

a. در هر تکرار ۱۰٪ را به عنوان مجموعه آزمون و ۹۰٪ از نمونه ها را به عنوان مجموعه آموزش در نظر بگیرید. داده ها به صورت تصادفی انتخاب نمائید. محل نمونه های انتخاب شده را به صورت تصادفی تغییر دهید. (توجه کنید در هر تکرار باید مجموعه آزمون و آموزش متفاوت باشند دقت نمائید انتخاب باید بر اساس برچسب ها ستون Name باشد یعنی در هر تکرار از هر دسته به صورت تصادفی برای آزمون ۱۰٪ و برای آموزش ۹۰٪ انتخاب شود. برای مثال اگر برچسب Iris-setosa شامل ۳۰ نمونه باشد ۳ نمونه برای آزمون و بقیه برای آموزش انتخاب شوند)
b. مقادیر ستون Name را به عدد تبدیل نمائید. به ازای هر برچسب متفاوت یک عدد با شروع از ۰ در نظر بگیرید.

c. مقادیر ویژگی های مربوط به ۴ ستون اول را در بازه [-1,1] نرمال کنید. دقت نمائید در مجموعه آزمون از min و max مجموعه آموزش استفاده نمائید. (توضیح: برای محاسبه مجموعه min و max فقط از مجموعه آموزش استفاده نمائید و از مقادیر آنها برای نرمال سازی مجموعه آزمون استفاده کنید)

d. ویژگی‌های جدید از ترکیب های مختلف (به صورت زیر) ایجاد نمائید. لازم به ذکر است ویژگی ها جدید را نرمال نمائید با همان سناریو مطرح شده در بخش b)

   i. میانگین ۴ ویژگی اول با عنوان mean_4Features (ویژگی ۵)

   ii. ویژگی وزن دار از ویژگی‌های اول و سوم با عنوان Weight_2Features (0.7*SepalLength+0.12*PetalLength) (ویژگی ۶)

e. در مجموعه آموزش، ستون name را در نظر بگیرید مقادیر مربوط به دو ویژگی mean_4Features و Weight_2Features را برای ۳ دسته متفاوت از این ستون نمایش دهید. (برای این منظور از matplotlib) استفاده نمائید.

f. فراوانی مربوط به ویژگی SepalWidth و PetalWidth را نشان دهید (برای این منظور بعد از محاسبه فراوانی هر تیم از matplotlib برای نمایش استفاده نمائید).

g. به ازای هر نمونه از مجموعه آزمون، با ترکیبات زیر، نزدیکترین نمونه به مجموعه آموزش را انتخاب نمائید. مقدار معادل name را نمایش دهید. و محاسبه کنید برچسب انتخابی با برچسب واقعی تطبیق دارد یا نه؟ برای محاسبه می توانید از فاصله اقلیدیسی استفاده نمائید. هر نمونه فاصله کمتر یعنی نزدیکتر می باشد. اگر چند نمونه با کمترین فاصله برابر یافت شد یکی از نمونه ها را به صورت تصادفی انتخاب نمائید.

   i. ترکیب ویژگی ۱ (SepalLength) با ویژگی ۵ (mean_4Features) (ابتدا ویژگی ها نرمال شود)

   ii. ترکیب وزندار ویژگی ۲ (SepalWidth) با ویژگی ۶ (Weight_2Features)

((SepalWidth*0.6)/(varieance(SepalWidth))+(Weight_2Features*0.7)/(mean(Weight_2Features))

(ابتدا ویژگی ها نرمال شود) دقت نمائید در این بخش varieance فقط یک عدد می باشد که مربوط به کل ستون ویژگی SepalWidth برای مجموعه آموزش و mean نیز یک عدد می باشد که مربوط کل ستون ویژگی Weight_2Features برای مجموعه

آموزش می باشد. برای مجموعه آزمون از مقادیر varieance و mean مجموعه آموزش استفاده نمائید.

h. برای قسمت f میزان معیار های زیر را برای هر کلاس محاسبه و نمایش دهید. این قسمت پیاده سازی شود از ابزار آماده استفاده نکنید.

$$Ac(c_j) = \frac{TP(c_j) + TN(c_j)}{TP(c_j) + FP(c_j) + TN(c_j) + FN(c_j)}$$

$$\Pr(c_j) = \frac{TP(c_j)}{TP(c_j) + FP(c_j)}$$

$$\mathrm{Re}(c_j) = \frac{TP(c_j)}{TP(c_j) + FN(c_j)}$$

طبقه c_i		نتیجه واقعی: بلی	نتیجه واقعی: خیر
نتیجه طبقه‌بندی کننده	بلی	TP	FP
	خیر	FN	TN

TP نشان دهنده تعداد نمونه های که متعلق به کلاس c_j بوده اند و روش شما به درستی تشخیص داده است.

FP نشان دهنده تعداد نمونه های که متعلق به کلاس c_j بوده اند ولی روش شما به اشتباه تشخیص داده است.

FN نشان دهنده تعداد نمونه هایی است که متعلق به کلاس غیر از c_j بوده است ولی روش شما به اشتباه جز کلاس c تشخیص داده است.

TN نشان دهنده تعداد نمونه هایی که متعلق به کلاس غیر از c_j بوده است و روش شما به درستی تشخیص داده است.

۳) مقادیر مربوط به معیار های قسمت h را برای ده تکرار نمایش دهید (برای مجموعه آزمون، یک بار برای هر کلاس به صورت مجزا و یک بار هم برای میانگین کل کلاس) و مشخص نمائید که بهترین هر معیار در چه تکراری بهترین می باشد. برای نمایش از matplotlib استفاده نمائید.

موفق باشید

علی قنبری سرخی