به نام خدا

دانشگاه صنعتی شریف
دانشکده مهندسی برق

تمرین سری پنجم یادگیری آماری
پاییز ۱۴۰۲
دکتر هدی محمدزاده

توجه: سوالات تئوری سری پنجم **امتیازی** هستند.

۱- فرض کنید دیتاست زیر برای طبقه بندی در اختیار داریم و می خواهیم از درخت تصمیم گیری برای طبقه بندی استفاده کنیم:

Weight	Eye Color	Num. Eyes	Output
N	A	2	L
N	V	2	L
N	V	2	L
U	V	3	L
U	V	3	L
U	A	4	D
N	A	4	D
N	V	4	D
U	A	3	D
U	A	3	D

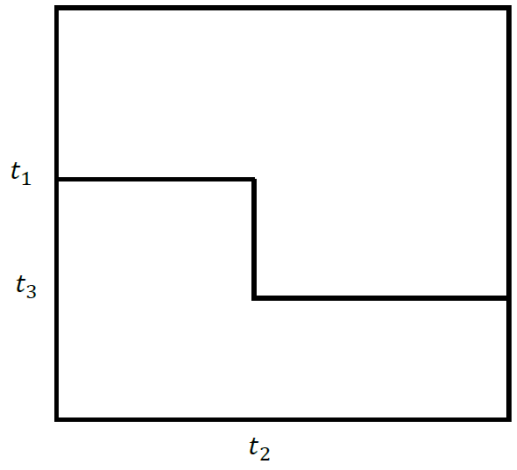
درخت تصمیم گیری کامل را برای طبقه بندی این دیتاست رسم کنید.

۲- با توجه به داده‌های موجود در جدول زیر، اولین ویژگی‌ای که در درخت تصمیم‌گیری انتخاب می‌شود چه خواهد بود؟ چرا؟

x_1	x_2	x_3	x_4	class
0	0	0	0	-
0	0	1	1	+
0	1	0	1	-
0	1	1	0	+
1	0	0	0	+
1	0	1	1	+
1	1	1	0	+
1	1	1	1	+

۳- الف: یک درخت تصمیم گیری برای مرز نشان داده شده در شکل زیر با حداقل برگ پیدا کنید.

ب: آیا این مرز می تواند در نتیجه ترکیب چند درخت به عمق یک بدست آمده باشد؟ توضیح دهید.(توجه داشته باشید که در ترکیب درخت های تصمیم گیری از رای اکثریت درخت ها برای مشخص کردن برچسب یک ناحیه استفاده می کنیم.)

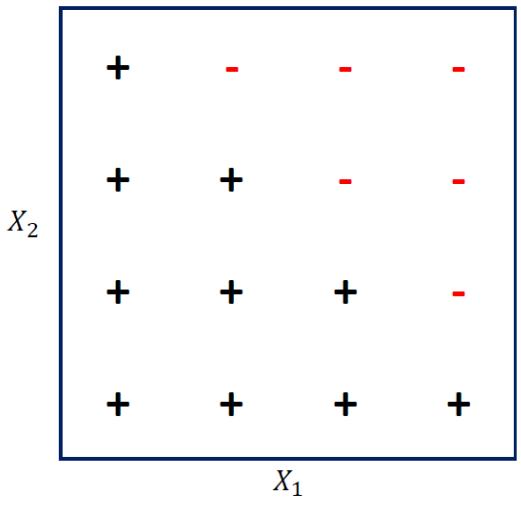


۴- در درخت تصمیم چند متغیره، در هر انشعاب، تابعی از ویژگی‌ها مورد بررسی قرار می گیرند. مثلا ممکن است در یک انشعاب، اگر $3X_2 + 6X_1 + 8 \geq 0$ به شاخه راست و اگر $3X_2 + 6X_1 + 8 < 0$ به شاخه چپ برویم.

داده های شکل زیر را در نظر بگیرید:

الف: درخت تصمیم تک متغیره با خطای آموزش صفر بسازید.

ب: نشان دهید درخت تصمیم چند متغیره با عمق یک وجود دارد که خطای آموزش آن صفر است.



سوال پایتون

تحویل این سوال اجباری می باشد. برای بخش های پیاده سازی نمیتوانید از منابع اینترنتی استفاده کنید

بخش اول

۱. در ابتدای برای آشنایی با الگوریتم SVM نوت بوک svm.ipynb پر کنید.
۲. در این بخش می خواهیم از روش بردار های پشتیبان استفاده کنیم که اعداد فارسی را دسته بندی کنیم. (برای این سوال از ۴۰۰۰ تصویر اول hoda dataset استفاده می کنیم.)

الف) در دیتایی که داده شده است سایز تصاویر با هم برابر نیست؟ الگوریتمی پیاده سازی کنید که این مشکل را برطرف کند و آنرا به صورت کامل توضیح دهید. (توجه شود که نمیتوانید از resizing استفاده کنید. چرا؟)

ب) پس از آن که مشکل برابر نبودن سایز تصاویر را برطرف کردید، ماتریس تصاویر را یک بعدی کنید. (پیکسل های تصاویر ویژگی های مورد نظر ما هستند)

ج) داده ها رو به دو بخش train و test تقسیم کنید.(۳۰ درصد داده ها برای train باشد)

د) در اینجا تعداد کلاس های ما ۲ نمی باشد، به نظرتان چگونه می توان الگوریتم SVM را برای بیش از دو کلاس استفاده کنیم؟ به خوبی الگوریتم را توضیح دهید.

ه) حالا الگوریتم یادگیری SVM را به کمک کتابخانه ی sklearn روی داده ها اجرا کنید. (کرنل های مختلف را امتحان کنید، کدام یک نتیجه ی بهتری می دهد و چرا؟)

خ) پس از اینکه دقت آموزش را گزارش کردید، حالا میخواهیم ببینیم که مدل ما تا چه حد به نویز مقاوم است، پس به دیتای test نویز فلفل نمکی اضافه کنید و دقت را گزارش کنید. (احتمال های مختلف را امتحان کنید مانند ۰,۰۵،۰,۱،۰,۱۵ و...)

ت) حالا می خواهیم میزان مقاومت مدل را به rotation چک کنیم. داده های تست را ۳۰ درجه rotate کنید و دقت مدل را گزارش کنید. نظرتان چیست؟ (توجه کنید باید عکس دو بعدی rotate شود و برای rotate کردن می توانید از توابع آماده استفاده کنید – کتابخانه ی cv2 تابع های آماده ای برای اینکار آماده کرده است که می توانید استفاده کنید)

توجه : اگر در learn کردن داده ها مشکل داشتید می توانید با استفاده از PCA بعد داده ها را کم کنید.

بخش دوم:

در این بخش می خواهیم با مفهوم Decision Tree بهتر آشنا شویم.

Data set ای که برای این سوال از آن استفاده میکنیم، data set تایتانیک است که از این لینک می توانید آنرا دانلود کنید.

هدف ما در این سوال این است که با توجه به اطلاعاتی که از مسافران وجود دارد بتوانیم مرگ یا زندگی آنهارا پیشبینی کنیم.

۱. در ابتدا پس از خواندن فایل X را از y جدا کرده.
۲. برخی ویژگی ها در فایل وجود دارند که به هدف ما کاملا بی ربط می باشند آنهارا شناسایی کنید و از X حذف کنید. (همچنین توضیح دهید که چرا این موارد بی ربط می باشند)
۳. سپس داده ها را به test و train تقسیم بندی کنید (۰,۳ از داده ها برای test) و با استفاده از decision tree آنهارا کلاسبندی کنید و بهترین پارامتر های درخت را بیابید و دقت آنرا گزارش کنید. (برای اینکار می توانید از grid search استفاده کنید)
۴. در این بخش هدف و نحوه ی انجام k-fold cross validation و leave one out را توضیح دهید .
۵. پس از انتخاب روش مناسب validation در بالا آنرا را روی داده های train اعمال کرده و بهترین پارامتر ها را پیدا و آنرا روی داده های test آزمایش کرده و گزارش کنید. (دقت کنید که اگر k-fold را انتخاب کردید $k = 10$ قرار بدهید)