دانشگاه تهران

به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری سوم یادگیری ماشین

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.

۳. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.

۴. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۵. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.

۶. از بین سوالات **شبیه سازی** حتما به هر سه مورد پاسخ داده شود.

۷. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.

۸. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.

۹. در صورت داشتن سوال، از طریق ایمیل های زیر سوال خود را مطرح کنید.

سوالات ۱و ۲ و ۳ و ۴ : taabansoleymani@gmail.com
سوالات ۵ و ۶ : s.m.moosavi000@ut.ac.ir

سوال ۱: (۱۵ نمره)

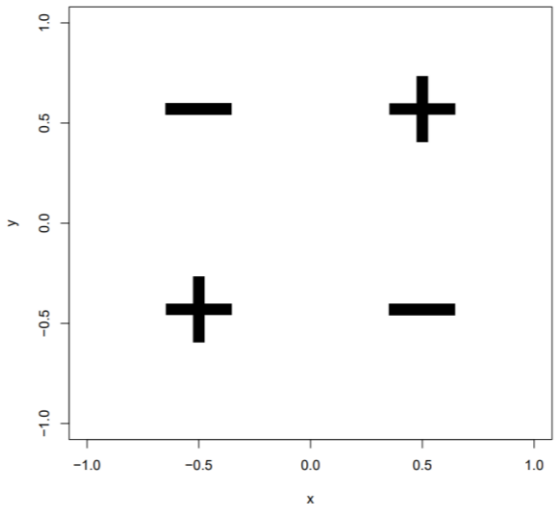
در یک دادگاه N نفر از اعضای هیئت منصفه حضور دارند و هریک از آنها مستقل از یکدیگر، با احتمال 70% می‌توانند جرم متهم را به درستی تشخیص دهند. در صورتیکه تصمیم نهایی با Majority Vote صورت گیرد (بدین ترتیب کلاسی انتخاب می‌شود که حداقل $\frac{N+1}{2}$ نفر از اعضا به آن رای دهند)، احتمال اینکه هیئت منصفه مشترکاً به رأی صحیح برسند را برای هریک از حالات زیر به دست آورید.

۱. $N = 5$
۲. $N = 9$
۳. $N \to \infty$. آیا در واقعیت با زیادکردن تعداد اعضا می‌توانیم به این دقت برسیم؟
۴. حالت 1 را دوباره برای زمانی که دقت هریک از اعضا 50% باشد، تکرار کنید. چه نتیجه‌ای می‌گیرید؟

سوال ۲: (۲۵ نمره)

الف)

۱. فرض کنید مجموعه داده دو بعدی مطابق با شکل ۱ در اختیار دارید. آیا به کمک AdaBoost (با استفاده از طبقه‌بندهای Decision Stump به عنوان طبقه‌بند ضعیف) می توان به دقت بیشتر از 50% دست یافت؟ (به طور خلاصه پاسخ خود را توجیه کنید.)

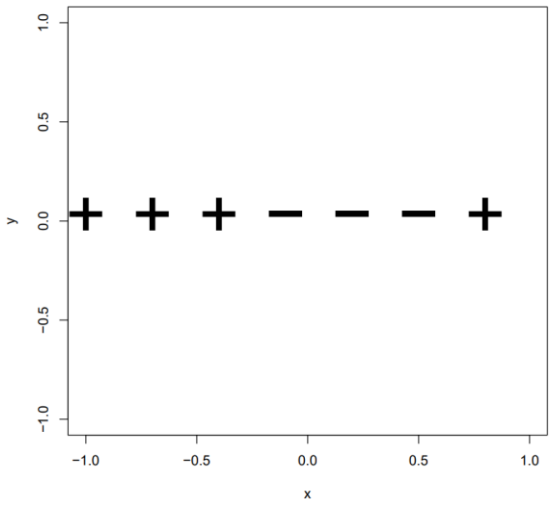


شکل ۱

۲. فرض کنید این بار طبقه‌بند AdaBoost را بر روی m داده‌ی آموزشی اجرا کرده‌ایم و در هر round، حداکثر مقدار weighted training error، ϵ_t، برابر با مقدار γ در بازه تعریف $0 < \gamma < 0 \cdot 5$ می‌باشد. پس از چه تعداد تکرار، به خطای یادگیری 0 می توان دست یافت؟

ب)

برای سوالات ۳-۸، مجموعه داده تک بعدی نمایش داده شده در شکل ۲ را در نظر بگیرید. و بار دیگر فرض کنید که از طبقه بند AdaBoost با طبقه‌بندهای Decision Stump به عنوان طبقه‌بند ضعیف استفاده کرده‌ایم.



شکل ۲

۳. مرز تصمیم اولین طبقه‌بند h_1 را رسم کنید و مشخص نمایید که چه قسمتی از فضا به عنوان کلاس + طبقه‌بندی می‌شود.

۴. مقادیر α_1 و ϵ_1 را محاسبه کرده و دقت طبقه‌بند را تا این مرحله گزارش نمایید.

۵. مقادیر جدید وزن هر یک از نقاط را به دست آورید.

۶. مرز تصمیم دومین طبقه‌بند h_2 را رسم کنید و بار دیگر مشخص نمایید که چه قسمتی به عنوان کلاس + طبقه‌بندی می‌شود.

۷. در صورت توقف AdaBoost در تکرار دوم، نقاطی که کمترین وزن را دارند مشخص کنید.

۸. آیا دقت طبقه بند AdaBoost بین تکرار اول و دوم بهبود می یابد؟ به طور خلاصه علت را بیان نمایید.

سوال ۳: (شبیه سازی، ۲۰ نمره)

در این قسمت انتظار می‌رود که با استفاده از مجموعه داده credit scoring sample پیش بینی نمایید که آیا مشتری در طول ۹۰ روز بدهی خود را بازپرداخت می‌نماید یا خیر؛ در واقع طبقه‌بند باینری به دست آمده مشتریان را به دو دسته خوش حساب و بد حساب تقسیم می‌نماید.
در ادامه توضیحاتی در ارتباط با ستون‌های دیتاست ارائه گردیده است.

feature	Description
Age	سن مشتری
DebtRatio	مجموع پرداختی وام/درصد درآمد ماهانه کل
NumberOfTime30-59DaysPastDueNotWorse	تعداد مواردی که باز پرداخت مشتری در طول ۲ سال گذشته ۳۰ تا ۵۹ روز (نه بیشتر) عقب افتاده است.
NumberOfTimes90DaysLate	تعداد مواردی که باز پرداخت مشتری ۹۰ روز یا بیشتر عقب افتاده است.
NumberOfTime60-89DaysPastDueNotWorse	تعداد مواردی که باز پرداخت مشتری در طول ۲ سال گذشته ۶۰ تا ۸۹ روز (نه بیشتر) عقب افتاده است.
NumberOfDependents	تعداد افراد وابسته به مشتری
SeriousDlqin2yrs	مشتری بدهی را ظرف مدت ۹۰ روز پرداخت نکرده است.

در ابتدا توزیع ستون هدف (SeriousDlqin2yrs) رسم نمایید، سپس مقادیر Nan در هریک از ستون‌های ورودی را با مقدار میانه مربوط به مقادیر همان ستون جایگزین نمایید.

Bootstrapping

۱- با نمونه برداری به روش bootstrapping، بازه‌ای را که میانگین سنی مشتریان بدحساب در آن قرار می‌گیرد، با سطح اطمینان ۹۰ درصد برآورد کنید.

Random Forest

۲- هدف این قسمت یافتن طبقه‌بند Random Forest بهینه با تعداد 100 درخت از میان مجموعه پارامترهای داده شده به صورت

```
parameters = {'max_features': [1, 2, 4], 'min_samples_leaf': [3, 5, 7, 9],
'max_depth': [5,10,15]}
```

به روش Grid Search می‌باشد. ابتدا لازمست تا برای جبران عدم توازن در دیتاست، مقدار پارامتر

```
class_weight='balanced'
```

ست کنید. برای یافتن مقدار بهینه هریک از پارامترها از مقیاس ارزیابی stratified 5-fold validation استفاده کنید. مقدار $ROC\ AUC$ را برای طبقه‌بند بهینه گزارش کنید.

۳- کدام ویژگی ضعیف‌ترین تأثیر بر مدل Random Forest را دارد؟

Bagging

۴- هدف این قسمت یافتن طبقه‌بند Bagging بهینه از میان مجموعه پارامترهای داده شده به صورت

```
parameters = {'max_features': [2, 3, 4], 'max_samples': [0.5, 0.7, 0.9],
              'base_estimator__C': [0.0001, 0.001, 0.01, 1, 10, 100]}
```

به روش Randomized Search می‌باشد. برای این کار از طبقه‌بند Logistic Regression به عنوان طبقه‌بند پایه استفاده کنید. (تعداد طبقه‌بندها را مانند مرحله قبل 100 در نظر بگیرید.) برای صرفه جویی در زمان تعداد تکرارهای RandomizedSearchCV را برابر با 20 قرار دهید و مانند قبل از مقیاس ارزیابی stratified 5-fold validation استفاده نمایید و مقدار $ROC\ AUC$ را برای طبقه‌بند بهینه گزارش کنید.

۵- بهترین مقادیر به دست آمده برای پارامترهای قسمت قبل را توجیه نمایید.

سوال ۴: (شبیه سازی، ۲۰ نمره)

هدف این بخش پیاده سازی AdaBoostClassifier برای طبقه‌بندی مجموعه داده iris می‌باشد که بدین منظور باید آن را از کتابخانه sklearn لود نمایید و از Decision Tree Classifier با عمق 1 به عنوان طبقه‌بند پایه استفاده کنید و تعداد این نوع طبقه‌بندها را 50 در نظر بگیرید. بعد از تقسیم کردن داده به دو قسمت train و test با نسبت 70 – 30، مدل را ارزیابی کرده و دقت و ماتریس آشفتگی را گزارش کنید.

برای پیاده‌سازی می‌توانید کدی که در اختیار شما قرار گرفته است را تکمیل نمایید، به علاوه اینکه مجاز به استفاده از کتابخانه های آماده به جز ماژول `sklearn.ensemble.AdaBoostClassifier` هستید.

سوال ۵: (۱۵ نمره)

جدول ۱ اطلاعات تعدادی بیمار را نشان می دهد :

جدول ۱ : اطلاعات بیماران و وضعیت ابتلا به بیماری انسداد شرایین برای دادگان آموزش

شماره	فشار خون	سطح کلسترول	مصرف سیگار	وزن	انسداد شرایین
۱	بله	نرمال	نه	اضافه وزن	بله
۲	نه	نرمال	بله	نرمال	نه
۳	نه	بحرانی	نه	اضافه وزن	بله
۴	نه	بالا	بله	اضافه وزن	بله
۵	بله	بحرانی	بله	چاق	بله
۶	بله	بالا	بله	نرمال	بله
۷	نه	بالا	نه	چاق	نه
۸	بله	نرمال	بله	نرمال	بله
۹	بله	بحرانی	نه	چاق	بله
۱۰	نه	نرمال	نه	اضافه وزن	نه
۱۱	نه	بحرانی	بله	نرمال	بله
۱۲	بله	بالا	نه	اضافه وزن	نه
۱۳	بله	نرمال	بله	اضافه وزن	بله
۱۴	بله	بالا	نه	چاق	نه

الف) با استفاده از معیار information gain یک درخت تصمیم برای ویژگی انسداد شرایین آموزش دهید.

ب) با کمک درخت به دست آمده در قسمت قبل دادگان جدول ۲ را پیش بینی کرده و عملکرد مدل را به کمک ماتریس آشفتگی توضیح دهید.

جدول ۲: اطلاعات بیماران و وضعیت ابتلا به بیماری انسداد شرایین برای دادگان آزمون

شماره	فشار خون	سطح کلسترول	مصرف سیگار	وزن	انسداد شرایین
۱۵	بله	نرمال	بله	چاق	بله
۱۶	بله	بالا	بله	چاق	بله
۱۷	بله	بالا	نه	نرمال	نه
۱۸	بله	نرمال	نه	نرمال	نه
۱۹	نه	نرمال	بله	اضافه وزن	بله

ج) چرا طبقه بندهای درخت تصمیم در برابر بیش برازش مقاوم نیستند؟ دو روش برای جلوگیری از این مشکل ارائه دهید.

سوال ۶: (**شبیه سازی**، ۱۵ نمره)

در این سوال با استفاده از پیاده سازی درخت تصمیم براساس الگوریتم ID۳، قصد داریم داده های دادگان dataset_prison را طبقه بندی کنیم.

ویژگی هدف ما 'بازگشت به زندان' خواهد بود و میخواهیم براساس ویژگی های دیگر تصمیم گیری را انجام دهیم. پس از تقسیم کردن داده به دو قسمت آموزش و آزمون با نسبت ۸۰-۲۰ , درخت تصمیمی با حداکثر عمق ۳ آموزش دهید ودقت و همچنین ماتریس آشفتگی را گزارش کنید.(توجه کنید که در این سوال مجاز به استفاده از توابع کتابخانه ای برای پیاده سازی درخت تصمیم نیستید.)