به نام خدا

**سیستم‌های هوشمند (دکتر رشاد حسینی)**

تمرین سری **دوم** (یادگیری ماشین نظارت‌شده)

نیم‌سال اول ۱۴۰۳-۱۴۰۲

سوال ۱: الگوریتم درخت تصمیم (تئوری) - (حمید سالمی)

سوال ۲: الگوریتم درخت تصمیم و جنگل تصادفی (پیاده‌سازی) - (محمدحسین وعیدی)

سوال ۳: الگوریتم kNN و یادگیری براساس معیار (پیاده‌سازی) - (عرفان پناهی)

---

## **سوال ۱:** الگوریتم درخت تصمیم (تئوری)

در این سوال قصد داریم با استفاده از الگوریتم درخت تصمیم طبقه بندی طراحی کنیم تا با استفاده از آن بتوان دو نوع جاندار را از هم تشخیص داد. ویژگی‌هایی که از جانداران داریم عبارتند از رنگ، تعداد پا، محل زندگی و قد.

جدول ۱-۱، داده‌های آموزش (ویژگی‌ها) دو جاندار برای طراحی طبقه بند درخت تصمیم را نشان می‌دهد.

| شماره | رنگ | تعداد پا | قد | محل زندگی | نوع جاندار |
|---|---|---|---|---|---|
| ۱ | قهوه‌ای | ۲ | بلند | خشکی | A |
| ۲ | قهوه‌ای | ۳ | کوتاه | خشکی | B |
| ۳ | سبز | ۲ | بلند | آب | B |
| ۴ | سبز | ۳ | بلند | آب | B |
| ۵ | قهوه‌ای | ۲ | کوتاه | آب | A |
| ۶ | قهوه‌ای | ۲ | بلند | آب | A |
| ۷ | قهوه‌ای | ۲ | کوتاه | خشکی | B |
| ۸ | سبز | ۲ | کوتاه | آب | A |
| ۹ | سبز | ۳ | بلند | آب | B |
| ۱۰ | قهوه‌ای | ۲ | بلند | خشکی | A |

**جدول ۱-۱:** اطلاعات مورد نیاز برای طراحی طبقه بند درخت تصمیم

### قسمت الف: طراحی طبقه‌بند

با استفاده از جدول ۱-۱، یک طبقه‌بند درخت تصمیم برای تشخیص نوع جاندار براساس بهرهٔ اطلاعات (Information gain) و با الگوریتم ID3 را آموزش دهید.

### قسمت ب: آزمون طبقه‌بند

با استفاده از طبقه‌بند قسمت الف، نوع جاندار هر یک از نمونه‌های جدول ۱-۲ را مشخص کرده و عملکرد مدل را به کمک ماتریس آشفتگی (Confusion matrix) بررسی کنید.

| شماره | رنگ | تعداد پا | قد | محل زندگی | نوع جاندار |
|---|---|---|---|---|---|
| ۱ | قهوه‌ای | ۲ | بلند | خشکی | A |
| ۲ | قهوه‌ای | ۳ | کوتاه | خشکی | B |
| ۳ | سبز | ۲ | بلند | آب | B |
| ۴ | سبز | ۳ | بلند | آب | B |
| ۵ | قهوه‌ای | ۲ | کوتاه | آب | A |

**جدول ۱-۲:** داده‌های آزمون طبقه بند درخت تصمیم

### قسمت ج: رویکرد حریصانه (Greedy) الگوریتم ID3

تمام شرایط لازم برای اینکه جانداری از نوع A باشد و یا از نوع B باشد را در نظر بگیرید. (برای مثال اگر تعداد پاها ۳ باشد، جاندار از نوع B است.) در هر کدام از این شروط، حداکثر از ۴ ویژگی استفاده شده است.

آیا می‌توانید، درخت تصمیم جدیدی طراحی کنید که فقط با استفاده از ۲ ویژگی بتواند نوع جاندار را تشخیص دهد و هم چنان باعث صفر شدن خطا در مجموعه آموزشی شود؟ (به این معنا که درخت تصمیم جدید هم چنان برای تمامی داده‌های آموزش صدق کند)

جواب خود را توجیه کنید.

### قسمت د: افزایش قوام طبقه‌بند

چرا طبقه‌بندهای درخت تصمیم در برابر بیش‌برازش (Overfitting) مقاوم نیستند؟ دو روش برای جلوگیری از این مشکل ارائه دهید.

## **سوال ۲:** الگوریتم درخت تصمیم و جنگل تصادفی (پیاده‌سازی)

در این سوال با استفاده از پیاده‌سازی درخت تصمیم بر اساس الگوریتم ID3، قصد داریم داده‌های دادگان "prison_dataset.csv" را طبقه‌بندی کنیم. ویژگی هدف ما، نرخ بازگشت به زندان( تکرار جرم) [1]خواهد بود و می‌خواهیم بر اساس ویژگی‌های دیگر تصمیم‌گیری را انجام دهیم.

### قسمت الف: طراحی طبقه‌بند

با نمونه‌برداری تصادفی[2] و به صورت ۸۰-۲۰ از دادگان داده شده، آن را به داده‌های آموزش و آزمون تقسیم کنید. با استفاده از الگوریتم ID3 درخت خود را پیاده‌سازی کنید و آن را با داده‌های آموزش، آموزش دهید. معیار انتخاب ویژگی برتر را بهره اطلاعات[3] درنظر گرفته و عمق درخت خود را ۳ درنظر بگیرید. در نهایت لازم است دقت طبقه‌بند برای داده‌های آزمایش و همچنین ماتریس آشفتگی را گزارش کنید. سپس عمق درخت را تغییر دهید و نتیجه‌گیری خود را بر اساس ماتریس آشفتگی توجیه کنید.

*** در این قسمت امکان استفاده از کتابخانه Scikit-Learn را ندارید.

### قسمت ب: استفاده از جنگل تصادفی

حال قصد داریم برای بهبود عملکرد طبقه‌بند، از الگوریتم جنگل تصادفی[4] استفاده کنیم. بدین منظور می‌توانید داده‌ها و ویژگی‌ها را تقسیم کرده و تعداد $K$ درخت (حداقل ۳ درخت) را آموزش دهید و درنهایت با استفاده از رای اکثریت[5] دقت طبقه‌بند برای داده‌های آزمون و همچنین ماتریس آشفتگی را گزارش کنید. آیا دقت طبقه‌بند افزایش پیدا کرد؟ چرا؟

### قسمت ج: استفاده از کتابخانه

در این قسمت با استفاده از کتابخانه Scikit-Learn الگوریتم جنگل تصادفی را با درنظر گرفتن موارد زیر پیاده‌سازی کنید و دقت طبقه‌بند برای داده‌های آزمایش و همچنین ماتریس آشفتگی را گزارش کرده و آن را با قسمت ب سوال مقایسه کنید.

$$\begin{cases} \max depth = 3 \\ random\ state = 0 \end{cases}$$

توجه: دقت کنید به دلیل اینکه جنگل تصادفی مربوط به کتابخانه Scikit-Learn مقادیر رشته برای ویژگی‌ها پشتیبانی نمی‌کند، لازم است که از رمزگذار برچسب[6] برای اینکار استفاده کنید؛ در اینترنت جستجو کنید و با استفاده از کتابخانه Scikit-Learn از روش مناسب برای رفع مشکل استفاده کنید.

[1] Recidivism – Return to Prison numeric
[2] Random Sampling
[3] Information Gain
[4] Random Forest
[5] Majority Voting
[6] Label Encoder

## **سوال ۳:** الگوریتم knn و یادگیری براساس معیار (metric learning) (پیاده‌سازی)

در این سوال قصد داریم در ابتدا الگوریتم k نزدیک‌ترین همسایه (kNN) را پیاده سازی کنیم. سپس با استفاده از روش‌های یادگیری براساس معیار می‌خواهیم این الگوریتم (kNN) را بهبود دهیم و تأثیر هر روش را مورد بررسی قرار دهیم.

به این منظور از دادگان "wine" از کتابخانه sklearn استفاده می‌کنیم. آن‌را به صورت زیر بخوانید:

```
from sklearn.datasets import load_wine
data = load_wine()
```

پس از اینکه دادگان را خواندید، ۲۰٪ آن‌را به دادگان آزمون و ۸۰٪ را به دادگان آموزش اختصاص دهید.

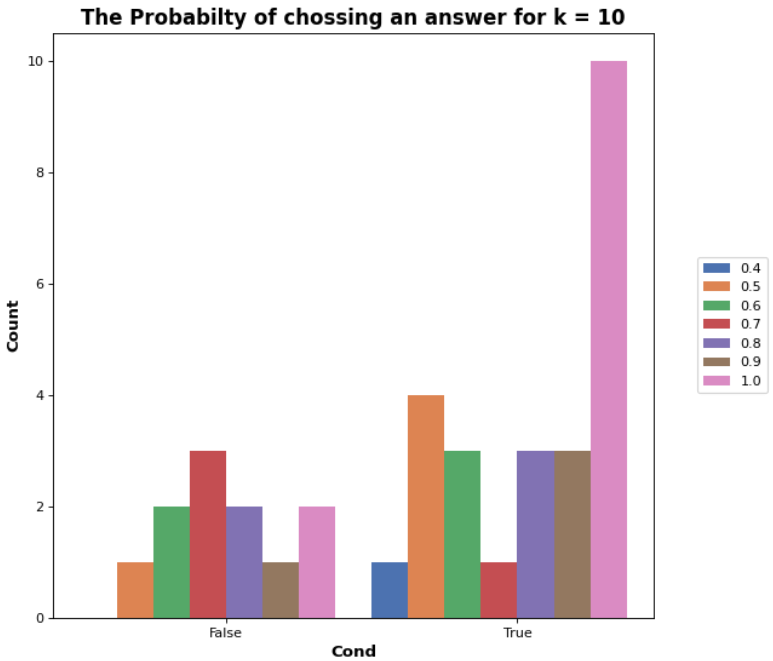
### **قسمت الف:** پیاده‌سازی طبقه‌بند kNN

مطابق با مطالب آموخته شده در درس الگوریتم k نزدیک‌ترین همسایه (kNN) را پیاده‌سازی نمایید و طبقه بندی دادگان تست "wine" را به ازای $k = 1, 5, 10, 20$ را انجام دهید. سپس دقت و ماتریس آشفتگی (confusion matrix) را برای هر حالت نمایش دهید.

*** در این قسمت امکان استفاده از کتابخانه sklearn را ندارید.

### **قسمت ب:** موضوع قسمت ب

برای $k = 5, 10, 20$، احتمال تعلق دادگان آزمون را به هر کلاس مشابه تصویر ۳-۱ رسم کنید. با تغییر تعداد همسایه ها (k)، نحوه تغییر در توزیع احتمال ها را بررسی کنید. برای کدام مقدار همسایه فکر می‌کنید مدل بهتر عمل کرده است؟ معیار خود را برای این انتخاب توضیح دهید.




**تصویر ۳-۱:** نمودار توزیع احتمالی تعلق به هر کلاس

## قسمت ج: یادگیری براساس معیار

*** در این قسمت می‌توانید از کتابخانه‌های آماده sklearn و metric-learn استفاده کنید.

در این قسمت می‌خواهیم دو روش یادگیری براساس معیار LMNN و LFDA و تأثیر آن‌ها روی دقت طبقه‌بندی kNN را بررسی کنیم.

۱- در هر یک از دو یادگیری بر اساس معیار، پارامتری بنام k وجود دارد. بنظر شما کارکرد این پارامتر چیست و چه تفاوتی با پارامتر k در طبقه بند k همسایه نزدیک دارد؟

۲- در این قسمت می‌خواهیم تاثیر یادگیری بر اساس معیار در افراز داده ها در ۲ بعد ببینیم. از آنجایی داده اولیه ما دارای ۱۳ بعد می‌باشد، نمایش آن در فضای دو بعدی امکان پذیر نیست. بدین منظور دو راه کار در پیش دارید:

- **با استفاده از کتابخانه metric-learning**: پارامتری را در فراخوان توابع مربوطه پیدا کنید که به کمک آن بتوان دادگان را به فضای با بعد پایین تر انتقال داد.
- **با استفاده از روش‌های کاهش بعد در کتابخانه sklearn**: می‌توانید بعد از انتقال دادگان به فضای جدید به کمک روش های کاهش بعد مانند PCA، دادگان را به فضای با بعد پایین‌تر انتقال دهید.

بعد از کاهش بعد دادگان اصلی و انتقال یافته در فضای جدید، به ازای $k = 1, 5, 15$ نحوه افراز دادگان هر کلاس با کلاس خود و کلاس های دیگر را رسم و تحلیل کنید. برای کدام مقدار $k$ دادگان در فضای جدید قابلیت تفکیک‌پذیری بیشتری دارند؟ چرا؟

۳- برای بهترین مقدار $k$ بدست آمده از سوال قبل، دقت و ماتریس آشفتگی طبقه‌بند را این بار برای دادگان انتقال یافته در فضای جدید به ازای تعداد همسایه مشابه بخش الف ($k = 1, 5, 10, 20$) در قسمت طبقه بند $k$ همسایه بدست آورده و مقایسه کنید.

## قسمت د: مقایسه ماتریس همبستگی

یکی از اطلاعات مفیدی که میتوان همواره از دادگان استخراج کرد، همبستگی بین ستون‌های ویژگی می‌باشد. بدین صورت که می‌توانیم ضریب همبستگی بین هر دو ستون ویژگی از دیتاست خود را داشته باشیم. این اطلاعات از این جهت سودمند است که می‌توانیم تاثیر متقابل ستون‌های ویژگی را در فرآیند یادگیری بیشتر درک کنیم. در کتابخانه Pandas می‌توانید به کمک دستور corr. همبستگی دو به دو بین ستون‌های ویژگی را در قالب یک آرایه دو بعدی بدست آورده و سپس رسم کنید.

*** برای رسم ماتریس همبستگی می‌توانید از کتابخانه seaborn استفاده کنید.

برای هر کدام از دو روش یادگیری بر اساس معیار، ماتریس‌های همبستگی را بررسی کنید. ستون‌های ویژگی در فضای انتقال یافته برای هر کدام از روش‌ها چه ویژگی متمایزی دارند. ستون‌های ویژگی بدست آمده در روش LMNN چه اطلاعات مهمی را در فضای جدید آشکار می‌کنند؟

## نکات تحویل

۱- مهلت تحویل این تمرین ۶ آذر ماه می‌باشد.

۲- انجام این تمرین به صورت یک‌نفره است.

۳- برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.

۴- در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن نمره صفر لحاظ می‌شود.

۵- در صورتی که از منبعی برای هر بخش استفاده می‌شود، حتماً لینک مربوط به آن در گزارش آورده شود. وجود شباهت بین منبع و پیاده‌سازی در صورت ذکر منبع بلامانع است. اما در صورت مشاهده شباهت با مطالب موجود در سایت‌های مرتبط نمره کسر می‌گردد.

۶- نتایج و تحلیل‌های شما در روند نمره‌دهی دستیاران آموزشی تأثیرگذار است.

۷- لطفا پاسخ تمرین خود را (به همراه کد/گزارش سوال کامپیوتری) به صورت زیر در صفحه درس آپلود نمایید:

HW[HW number]_[Last_name]_[Student number].zip

۸- در صورت وجود هر گونه ابهام یا مشکل می‌توانید از طریق ایمیل با طراحان تمرین در تماس باشید:

- عرفان پناهی (erfanpnhii@gmail.com)
- محمد حسین وعیدی (mohamadhoseinvaeedi@gmail.com)
- حمید سالمی (salemihamid77@gmail.com)