بنام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره **2**

آبان ۱۴۰۱

فهرست سوالات

سوال ۱ – درخت تصمیم (تحلیلی)

در این سوال بررسی می‌کنیم که چگونه دو نوع جاندار را، براساس رنگ، تعداد پا، قد و محل زندگی، می‌توان از هم تشخیص داد. ستون جاندار، برچسبی را که می‌خواهیم پیش بینی کنیم، نشان می‌دهد.

جدول 1-1: اطلاعات مورد نیاز برای طبقه‌بندی دو نوع جاندار (داده‌های آموزش)

شماره	رنگ	تعداد پا	قد	محل زندگی	جاندار
۱	قهوه‌ای	2	بلند	خشکی	A
۲	قهوه‌ای	3	کوتاه	خشکی	B
۳	سبز	2	بلند	آب	B
۴	سبز	3	بلند	آب	B
۵	قهوه‌ای	2	کوتاه	آب	A
۶	قهوه‌ای	2	بلند	آب	A
۷	قهوه‌ای	2	کوتاه	خشکی	B
۸	سبز	2	کوتاه	آب	A
۹	سبز	3	بلند	آب	B
۱۰	قهوه‌ای	2	بلند	خشکی	A

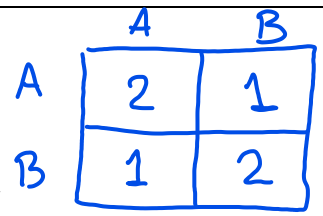
الف) طراحی طبقه‌بند

با استفاده از جدول ۱-۱، یک طبقه بند درخت تصمیم[1] برای تشخیص نوع جاندار (A یا B)، بر مبنای بهره‌ی اطلاعات [2] و با الگوریتم ID3 را آموزش دهید.

ب) آزمون طبقه‌بند

با استفاده از طبقه‌بند قسمت الف، نوع جاندار هر یک از نمونه‌های زیر را (جدول ۱-۲) مشخص کرده و عملکرد مدل را به کمک ماتریس آشفتگی[3] بررسی کنید.

[1] Decision Tree
[2] Information Gain
[3] Confusion Matrix



جدول 1-2: اطلاعات مورد نیاز برای طبقه‌بندی دو نوع جاندار (داده‌های آزمون)

شماره	رنگ	تعداد پا	قد	محل زندگی	جاندار
۱	قهوه‌ای	3	بلند	خشکی	B
۲	سبز	2	بلند	خشکی	A
۳	سبز	2	کوتاه	خشکی	A
۴	قهوه‌ای	2	کوتاه	آب	B
۵	قهوه‌ای	2	بلند	خشکی	A

ج) رویکرد حریصانه[4] 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.

د) افزایش قوام طبقه‌بند

چرا طبقه‌بندهای درخت تصمیم در برابر بیش‌برازش[5] مقاوم[6] نیستند؟ دو روش برای جلوگیری از این مشکل ارائه دهید.

[4] Greedy approach
[5] Overfitting
[6] Robust

سوال ۲ – پیاده‌سازی الگوریتم درخت تصمیم

در این بخش، با استفاده از داده‌های پیوست شده هدف آن است که الگوریتم درخت تصمیم را بدون استفاده از کتابخانه های آماده، پیاده‌سازی کنیم.

الف) پیاده‌سازی مدل درخت

در ابتدا، به داده های مورد نیاز دسترسی پیدا می کنیم:

```
Import pandas as pd
Train = pd.read_csv("titanic-train.csv")
Test = pd.read_csv("titanic-test.csv")
```

هدف آن است که بتوانیم با استفاده از ویژگی‌های افرادی که سوار بر کشتی تایتانیک بودند، نجات یافتن یا نیافتن هر کدام را پیشبینی کنیم. به نظر شما، چه ویژگی هایی می توانند در این امر تأثیرگذار باشند؟ (مشخصا، اسم فرد بی‌تأثیر خواهد بود، اما سن او می تواند اطلاع مفیدی باشد).

با توجه به این دید، در ابتدا، پیش پردازش های لازم را بر روی این داده‌ها انجام دهید و هر کدام را توضیح دهید.

(** برای پیش‌پردازش های لازم، می توانیم از این لینک، کمک بگیرید)

عمق درخت را (تعداد ویژگی های استفاده شده برای پیشبینی) در ابتدا ۳ در نظر بگیرید.

(**دقت شود که در مراحل گوناگون، می توان از یک ویژگی چند مرتبه استفاده کرد)

معیار انتخاب ویژگی برتر را به دلخواه، از میان معیارهای معرفی شده در کلاس انتخاب نمایید.

در هر مرحله، ماتریس آشفتگی[۷] را نمایش دهید و همچنین دقت مدل خود را گزارش کنید. با افزایش عمق درخت، دقت طبقه‌بند چه تغییری می‌کند؟ توضیح دهید.

ب) بهبود بخشی الگوریتم درخت تصمیم

ایرادات الگوریتم درخت تصمیم را بیان نمایید و توضیح دهید چگونه می‌توان عملکرد این الگوریتم را با استفاده از روش های Bagging و جنگل تصادفی[۸] بهبود دهیم.

ج) استفاده از جنگل تصادفی

حال، با استفاده از الگوریتم جنگل تصادفی، سعی بر آن است که دقت طبقه‌بند افزایش یابد. در این راستا، تعدادی درخت تصمیم(حداقل 3 درخت) که هر کدام بر اساس ویژگی های تصادفی بر روی تعدادی داده‌های تصادفی پیاده می شوند. با استفاده از رأی اکثریت[۹] بایستی پیشبینی مورد نظر را اعلام کنیم. حال دقت و ماتریس آشفتگی را گزارش دهید.

---

[7] Confusion Matrix
[8] Random Forest
[9] Majority Voting

سوال ۳ – یادگیری بر اساس معیار

در این سوال قرار است که تاثیر روش های یادگیری بر اساس معیار را بر طبقه بند k همسایه نزدیک بررسی کنیم. در ابتدا به کمک دستور زیر، دیتاست "wine" را به کمک کتابخانه Scikit-learn بخوانید:

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

(در تمامی مراحل سوال نیاز به استانداردسازی[10] دادگان نمی‌باشد)

طبقه بند k همسایه نزدیک

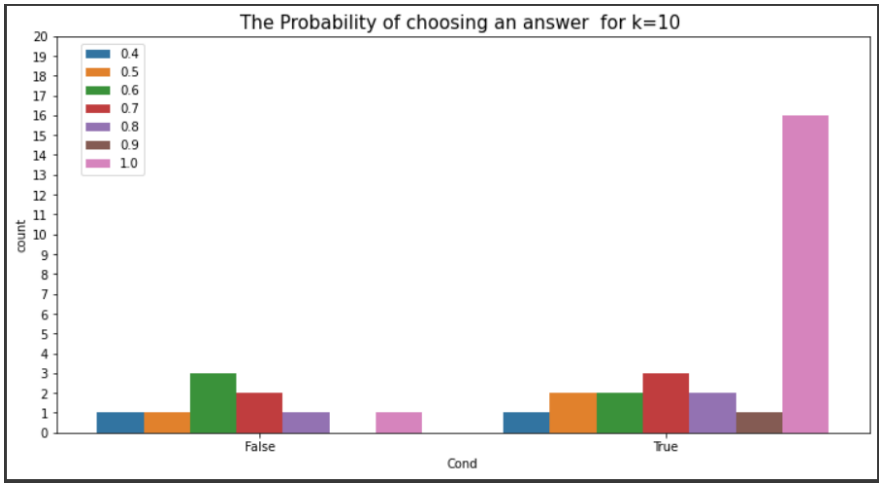
(دراین قسمت مجاز به استفاده از کتابخانه Scikit-learn نیستید و باید تمامی پیاده سازی را به کمک کتابخانه NumPy انجام دهید)

الف) طراحی طبقه بند

در ابتدا 20 درصد دادگان را به دادگان آزمون و بقیه را به دادگان آموزش اختصاص دهید. به ازای تعداد همسایه های 1،5،10 و 20 دقت طبقه بند و ماترس آشفتگی را بر روی دادگان آزمون گزارش دهید.

ب) محاسبه توزیع احتمال تعلق به هر کلاس

برای تعداد همسایه‌های ذکر شده در بخش قبل، احتمال تعلق دادگان آزمون را به هر کلاس مشابه شکل آمده در زیر رسم کنید. با تغییر تعداد همسایه ها، 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.



شکل ۳-۱ : نمودار توزیع احتمالی تعلق به هر کلاس

[10] Normalization

یادگیری بر اساس معیار

(در این قسمت می‌توانید از کتابخانه‌های آماده Scikit-learn و metric-learn استفاده کنید)

به کمک دو یادگیری بر اساس معیار LMNN[11] و LFDA[12] :

الف) بررسی کارکرد روش یادگیری

هدف تعریف هر یک از یادگیری‌ها، قیود[۱۳] تعریف شده در فرآیند یادگیری وعلت وجود هر یک از قید ها را توضیح دهید.

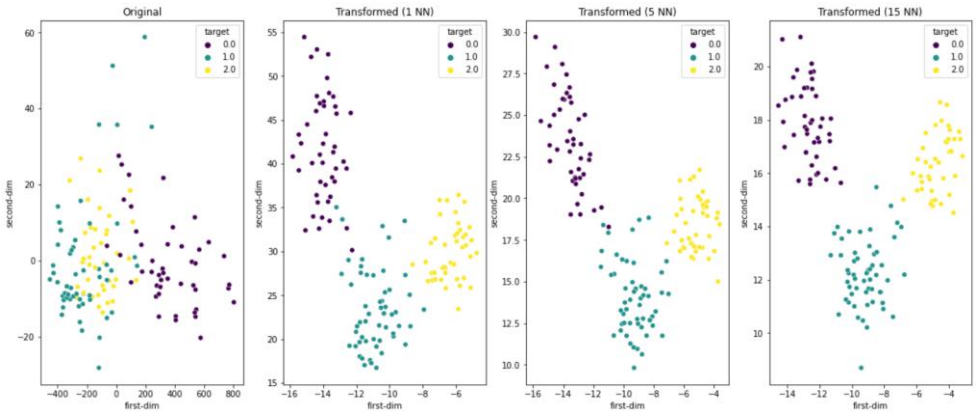
ب) ترسیم دادگان انتقال یافته در فضای جدید

1. 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 ؟
2. در این قسمت می‌خواهیم تاثیر یادگیری بر اساس معیار در افراز داده ها در 2 بعد ببینیم. از آنجایی داده اولیه ما دارای 13 بعد می‌باشد، نمایش آن در فضای دو بعدی امکان پذیر نیست. بدین منظور دو راه کار در پیش دارید:

- به کمک کتابخانه metric-learning (راه کار پیشنهادی) : پارامتری را در فراخوان توابع مربوطه پیدا کنید که به کمک آن بتوان دادگان را به فضای با بعد پایین تر انتقال داد.
- به کمک روش های کاهش بعد در کتابخانه Scikit-learn: میتوانید بعد از انتقال دادگان به فضای جدید به کمک روش های کاهش بعد مانند PCA[14]، دادگان را به فضای با بعد پایین تر انتقال داد.

بعد از کاهش بعد دادگان اصلی و انتقال یافته در فضای جدید، به ازای 3 مقدار مختلف برای پارامتر 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 تفکیک پذیری بیشتری دارند ؟ چرا ؟ **(نمودار مشابه در شکل زیر آورده شده است.)**

[11] Largest Margin Nearest Neighbor
[12] Local Fisher Discriminant analysis
[13] Constraints
[14] Principal analysis component



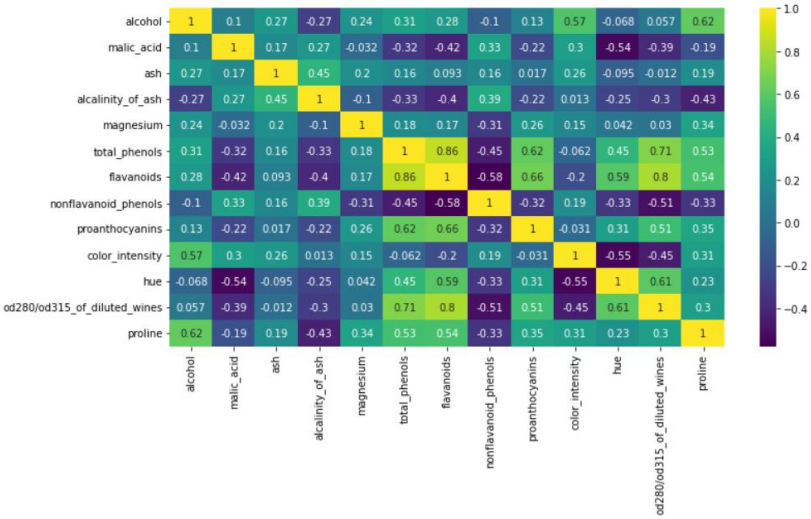
شکل ۳-۲ : افراز دادگان اصلی و انتقال یافته به ازای مقادیر مختلف k

ج) مقایسه عمکلرد طبقه بند

برای بهترین مقدارk بدست آمده از قسمت قبل، دقت و ماتریس آشفتگی طبقه بند را این بار برای دادگان انتقال یافته در فضای جدید به ازای تعداد همسایه مشابه بخش الف در قسمت طبقه بند k همسایه بدست آورده و مقایسه کنید.

د) ضریب همبستگی

یکی از اطلاعات مفیدی که میتوان همواره از دادگان استخراج کرد، همبستگی بین ستون های ویژگی [۱۵] 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 تاثیرمتقابل 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 آورید و سپس رسم کنید که نمونه‌ای از آن در پایین آورده شده است :



شکل ۳-۳ : نمایش ماتریس همبستگی یک دیتاسیت دلخواه

[15] Feature set

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

ه) GMML

در این مقاله، روش GMML به عنوان روشی جدید در یادگیری بر اساس معیار معرفی شده است. مدل سازی مساله را توضیح دهید. روش ارائه شده چه تفاوتی با روش LMNN مطرح شده در ابتدا دارد ؟

نکات تحویل:

- مهلت تحویل این تمرین 6 آذر میباشد.
- انجام این تمرین به صورت یک نفره است.
- برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.
- در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن -۱۰۰ لحاظ میشود.
- 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

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

  مسئول تمرین سوال۱: عاطفه ملاباقر (ut.ac.ir@ati.mollabagher)

  مسئول تمرین سوال۲: دریا افضلی (darya.afzali@ut.ac.ir)

  مسئول تمرین سوال۳: شایان واصف (sh.vassef@ut.ac.ir)