به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین چهارم

طراحان	حسین سیفی Hosein.Seifi@ut.ac.ir
تاریخ بارگذاری	۱۴۰۳/۲/۱۵
مهلت ارسال	۱۴۰۳/۲/۲۸

فهرست

بخش تشریحی

مجموعه داده زیر را در نظر بگیرید:

	چشم انداز	دما	رطوبت	باد	امکان برگزاری
۱	آفتابی	گرم	زیاد	ضعیف	خیر
۲	آفتابی	گرم	زیاد	قوی	خیر
۳	ابری	گرم	زیاد	ضعیف	بله
۴	بارانی	معتدل	زیاد	ضعیف	بله
۵	بارانی	خنک	عادی	ضعیف	بله
۶	بارانی	خنک	عادی	قوی	خیر
۷	ابری	خنک	عادی	قوی	بله
۸	آفتابی	معتدل	زیاد	ضعیف	خیر
۹	آفتابی	خنک	عادی	ضعیف	بله
۱۰	بارانی	معتدل	عادی	ضعیف	بله
۱۱	آفتابی	معتدل	عادی	قوی	بله
۱۲	ابری	معتدل	زیاد	قوی	بله
۱۳	ابری	گرم	عادی	ضعیف	بله
۱۴	بارانی	معتدل	زیاد	قوی	خیر

این مجموعه داده، اطلاعات مربوط به آب و هوای ۱۴ روز و امکان برگزاری مسابقه تنیس را نشان می‌دهد. اطلاعات مربوط به هر روز شامل وضعیت باد، رطوبت هوا، دما و چشم‌انداز کلی آب و هوا است و برای هر نمونه نیز برچسبی با عنوان امکان برگزاری در نظر گرفته شده است. با توجه به مجموعه داده فوق به سوالات پاسخ دهید.

سوالات

۱. آنتروپی امکان برگزاری مسابقه را محاسبه کنید.
۲. مقدار Information Gain ویژگی دما را به دست آورید.
۳. محاسبه کنید که الگوریتم ID3 کدام ویژگی را به عنوان ریشه درخت انتخاب می‌کند.
۴. درخت تصمیم تشخیص امکان برگزاری مسابقه تنیس را با توجه به مجموعه داده فوق و استفاده از الگوریتم ID3(استفاده از Information Gain) را حداکثر تا عمق ۲ به دست آورید و رسم کنید.
۵. برچسب نمونه‌های زیر را به کمک درخت تصمیم ایجاد شده مشخص کنید.

	چشم انداز	دما	رطوبت	باد
۱	بارانی	گرم	زیاد	ضعیف
۲	آفتابی	خنک	زیاد	قوی

۶. با فرض اینکه برچسب اصلی هر دو نمونه فوق برابر با بله باشد، صحت[1] درخت تصمیم ایجاد شده بر روی نمونه‌های فوق را به دست آورید.
۷. از درخت تصمیم ایجاد شده، چند قانون می‌توان استخراج کرد؟ تمامی این قوانین را به دست آورید و در گزارش ذکر کنید.

[1] Accuracy

بخش عملی

پیش‌نیازها

برای پاسخ به این تمرین عملی باید از زبان برنامه‌نویسی **Python** استفاده‌کنید و نیاز است که پیش از شروع، یک سرور **Jupyter** بر روی سیستم نصب و راه‌اندازی شود تا بتوانید بر روی یک فایل **ipynb.** کدهای خود را اجرا کنید، همچنین راه حل جایگزین آن استفاده از **Google Colab** است.

استفاده از کتابخانه‌های **Pandas**، **Numpy**، **Datetime** و Scikit-learn می‌تواند گزینه‌ی مناسبی برای حل مسائل پیشرو باشد.

شرح دادگان

مجموعه داده بخش اول شامل تعدادی ویژگی در مورد قارچ‌ها است و متغیر هدف این مجموعه داده، خوراکی یا سمی بودن هر یک از آن‌ها است. این مجموعه داده شامل ۲۰ ویژگی و ۱ متغیر هدف است. نام هر یک از ویژگی‌ها در مجموعه داده گویای محتوای آن است.

برای بارگذاری این مجموعه داده امکان استفاده از فایل آپلود شده به همراه تمرین و همچنین بارگذاری به کمک کتابخانه ucimlrepo وجود دارد و دستورات بارگذاری به کمک کتابخانه معرفی شده به شکل زیر است:

```
from ucimlrepo import fetch_ucirepo
df = fetch_ucirepo(id=848)
Features = df.data.features
label = df.data.targets
```

در قطعه کد فوق، دیتافریم Features شامل ویژگی‌های هر نمونه و دیتافریم Label شامل برچسب هر نمونه است.

در بخش دوم سوالات تشریحی، یک مجموعه داده شامل حداقل دمای ثبت شده به صورت روزانه در شهر ملبورن استرالیا بین سال‌های ۱۹۸۱ تا ۱۹۹۰ در کنار فایل تمرین در اختیار شما قرار گرفته است. این مجموعه داده تاریخ و حداقل دمای ثبت شده در آن تاریخ مشخص را شامل می‌شود.

همچنین می‌توانید مجموعه داده مورد نظر را با استفاده از لینک به صورت مستقیم در محیط برنامه‌نویسی پایتون و با استفاده از کتابخانه Pandas بارگذاری کنید.

بخش اول

اگرچه اولین راهی که برای آموزش یک مدل یادگیری ماشین با خروجی مناسب به ذهن بسیاری از افراد می‌رسد، استفاده از مدل‌ها و الگوریتم‌های پیچیده است، اما پایه و اساس مدل‌های موفق یادگیری ماشین، بر روی کیفیت داده‌هایی است که برای آموزش این مدل‌ها مورد استفاده می‌گیرند. پیش‌پردازش داده‌ها مرحله مهمی در روند آموزش یک مدل دسته‌بند است که داده‌های خام را به ویژگی‌هایی معنادار تبدیل می‌کند.

در این بخش قصد داریم بررسی و پیش‌پردازش مجموعه دادگان مورد استفاده، انتخاب، آموزش و ارزیابی مدل دسته‌بندی که تا حد ممکن دارای توانایی تفکیک دادگان باشد را گام به گام انجام دهیم و با روند انجام یک پروژه دسته‌بندی دادگان آشنا شویم.

- در تمرین اول درس داده کاوی با پیش‌پردازش‌های متفاوت دادگان آشنا شدید. در این تمرین نیاز دارید تا از تکنیک‌هایی که در تمرین اول فراگرفته‌اید استفاده کنید.

أ. بررسی کنید که در هر ویژگی مجموعه داده چه نوع مقادیری وجود دارند و هر یک از ویژگی‌ها دارای چه تعداد مقدار از دست رفته[1] هستند.

ب. بسیاری از مدل‌ها مقادیر از دست رفته(nan) را به عنوان ورودی نمی‌پذیرند. به منظور استفاده از این مدل‌ها و پاکسازی مجموعه داده از مقادیر از دست رفته، بین حذف ویژگی‌های شامل مقادیر از دسته رفته یا حذف نمونه‌های شامل این مقادیر، سیاست مناسب را اتخاذ کنید.

ج. همانطور که می‌دانید برخی انواع مدل‌های دسته‌بند تنها قادر به استفاده از ویژگی‌های عددی هستند. ویژگی‌های غیرعددی و برچسب‌های موجود در این مجموعه داده را به مقادیر عددی تبدیل کنید.

د. موضوع پراهمیت دیگر، توجه بیشتر مدل به ویژگی‌هایی است که دارای مقادیر بزرگتری هستند، بنابراین نیاز دارید بر روی داده‌های موجود، نرمالسازی انجام دهید.(انتخاب روش نرمالسازی مناسب بر عهده‌ی شماست)

ه. تعداد داده‌هایی با هر یک از برچسب‌ها را به دست آورید. آیا اختلاف تعداد هر یک از برچسب‌ها به گونه‌ای هست که مجموعه داده را نامتوازن[2] بدانیم؟ در صورتی که پاسخ شما به این سوال مثبت است، با استفاده از دانش خود، راهکار مناسب را برای متوازن کردن تعداد نمونه‌های هر کلاس به کار گیرید. (روش‌هایی مانند نمونه‌گیری[3]، تکرار داده‌ها یا افزودن داده‌ها[4])

[1] Missing Values
[2] Imbalanced
[3] Sampling
[4] Data Augmentation

- پس از انجام پیش‌پردازش‌های مورد نیاز، و تبدیل داده‌های خام به یک مجموعه داده قابل استفاده در مدل‌های یادگیری ماشین، شما باید مدل‌های دسته‌بند خواسته شده را آموزش دهید، ارزیابی کنید و با یکدیگر مقایسه کنید.

أ. داده‌ها پیش‌پردازش شده را با نسبت ۲۰/۸۰ به مجموعه داده‌های آموزش و تست تقسیم کنید.

ب. یک مدل درخت تصمیم(بدون در نظر گرفتن حداکثر عمق) با استفاده از کتابخانه Scikit-learn و با هایپرپارامترهای پیش‌فرض را ایجاد کنید و بر روی دادگان مربوطه آموزش دهید و برچسب متناظر با هر یک از نمونه‌های تست را به کمک مدل آموزش داده شده پیش‌بینی کنید. به کمک توابع موجود، زمان انجام هر یک از مراحل آموزش و پیش‌بینی را محاسبه کنید و گزارش دهید.

ج. عملیات خواسته شده در بخش b سوال را با استفاده از الگوریتم K نزدیک‌ترین همسایه[1] به ازای K برابر با ۹ انجام دهید.

د. زمان اجرای مراحل آموزش و پیش‌بینی را در هر کدام از مدل‌های آموزش دیده را مقایسه کنید. تفاوت در زمان اجرای هر کدام از مراحل این مدل‌ها نشان‌دهنده کدام خصوصیت این الگوریتم‌ها است؟ هر یک از این الگوریتم‌ها برای چه کاربردهایی مناسب هستند؟

ه. مقدار معیارهای ارزیابی صحت[2]، دقت[3]، بازخوانی[4] و F1 را برای هر یک از مدل‌ها به دست آورید و در قالب یک نمودار میله‌ای[5] در کنار یکدیگر نمایش دهید. در پیش‌بینی برچسب‌های نمونه‌های تست این مجموعه داده، برتری با کدام یک از مدل‌های آموزش دیده است؟

و. ماتریس آشفتگی[6] را برای هر یک از مدل‌های آموزش داده شده رسم کنید و آن را تفسیر کنید.

ز. درخت تصمیم به دست آمده در بخش ب را رسم کنید و در گزارش الصاق کنید. آیا امکان تفسیر این مدل وجود دارد؟ برای بهبود تفسیرپذیری درخت تصمیم چه روش‌هایی را پیشنهاد می‌کنید؟ این امکان در ازای چه هزینه‌ای امکان پذیر است؟

[1] K Nearest Neighbor (KNN)
[2] Accuracy
[3] Precision
[4] Recall
[5] Bar Chart
[6] Confusion Matrix

بخش دوم

یکی از کاربردهای مهم مدل‌های دسته‌بند، دسته‌بندی توالی داده‌هاست. در این سوال قصد داریم با استفاده از مجموعه داده موجود و ارائه اطلاعات مربوط به آب و هوای تعداد روزی مشخص، به پیش‌بینی اطلاعاتی در مورد آب و هوای روز بعدی بپردازیم.

أ. ابتدا نیاز داریم تا مجموعه داده را به تعدادی نمونه دارای توالی زمانی تبدیل کنیم. برای انجام این کار یک پنجره زمانی با طول ۱۰ روز در نظر می‌گیریم که دمای ده روز به عنوان ویژگی هر نمونه و دمای روز یازدهم به عنوان برچسب نمونه در نظر گرفته می‌شود. برای مثال دمای روز اول تا دهم به عنوان ویژگی و دمای روز یازدهم به عنوان برچسب نمونه اول، دمای روز دوم تا یازدهم به عنوان ویژگی و دمای روز دوازدهم به عنوان برچسب نمونه دوم و به طور کلی دمای روزهای N تا N+9 به عنوان ویژگی و دمای روز N+10ام به عنوان برچسب آن نمونه استفاده می‌شود.(توجه داشته باشید که مجموعه داده ابتدایی شامل ۳۶۵۰ نمونه است و پس از انجام پیش‌پردازش، باید شما ۳۶۴۰ سری زمانی برچسب‌دار ایجاد کرده باشید.)

ب. داده‌ها را با نسبت ۸۰٫۲۰ به دو بخش تست و آموزش تقسیم کنید.

ج. یک مدل رگرسیون خطی[1] ایجاد کنید و با استفاده از داده‌های مربوطه، آن را آموزش دهید.

د. دمای روز بعدی در برای هر یک از سری‌های زمانی موجود در بخش تست پیش‌بینی کنید، سپس معیارهای جذر میانگین مربعات خطا[2] و میانگین مطلق خطا[3] را برای مقادیر به دست آمده محاسبه کنید و گزارش کنید.

ه. مقادیر پیش‌بینی شده و مقادیر واقعی موجود در سری‌های زمانی ایجاد شده را به شکل دو منحنی را برای داده‌های تست در یک نمودار رسم کنید و عملکرد مدل را با توجه به نمودار به دست آمده ارزیابی کنید.

و. برچسب سری‌های زمانی را به نحوی تغییر دهید که نشان‌دهنده افزایش یا کاهش دمای روز بعدی نسبت به دمای آخرین روز موجود در هر نمونه باشد. برای مثال در نمونه زیر دمای روز بعدی دارای مقدار ۱۳٫۳ است و با توجه به کاهش دمای روز بعدی نسبت به آخرین روز موجود در این سری زمانی(۱۶٫۲)، برچسب جدیدی با مقدار ۰ دریافت می‌کند و در صورتی که مقدار برچسب قبلی از دمای آخرین روز نمونه متناظر با آن بیشتر باشد، برچسب جدید با مقدار ۱ را دریافت

17.9	18.8	14.6	15.8	15.8	15.8	17.4	21.8	20.0	16.2

[1] Linear Regression
[2] Root Mean Squared Error
[3] Mean Absolute Error

می‌کند.

ز. یک مدل ماشین بردار پشتیبان[1] ایجاد کنید و آن را با استفاده از بخش آموزش مجموعه داده با برچسب‌های دودویی ایجاد شده در بخش "و" آموزش دهید.

ح. نمونه‌های موجود در بخش تست را با استفاده از مدل آموزش داده شده برچسب بزنید و مقدار معیارهای ارزیابی صحت، دقت، بازخوانی و F1 را برای مدل ایجاد شده به دست آورید.

ط. (امتیازی) بررسی کنید که آیا مقادیر پیش بینی شده برای داده‌های تست در دو مدل بخش "ج" و "ز" ، تناظری با یکدیگر دارند یا خیر؟ آیا می‌توان حد آستانه‌ای[2] برای پیش‌بینی‌های مدل "ج" تعریف کرد که مقادیر بیش از آن متعلق به یک کلاس پیش‌بینی شده مدل "ز" و مقادیر کمتر از آستانه متعلق به کلاس دیگر دسته‌بند "ز" باشد؟

[1] Support Vector Machine (SVM)
[2] Threshold

ملاحظات

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA4_StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفا تمام موارد مورد نیاز را در گزارش ذکر کنید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- برای تحویل تمارین از چارچوب قرارداده شده در سامانه و کانال تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفا پیش از ارسال با دستیارمسئول تمرین هماهنگ کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:hosein.seifi@ut.ac.ir

مهلت تحویل: ۲۸ اردی‌بهشت ۱۴۰۳

مهلت تحویل با تاخیر: 4 خرداد ۱۴۰۳