تمرین سوم هوش مصنوعی و یادگیری ماشین (دسته بندی چندگانه)

اردیبهشت ۱۴۰۲

هدف این تمرین آشنایی با شیوهٔ بکارگیری روش‌های مختلف دسته بندی چندگانه است.

بخش اول: دسته‌بندی گیاهان بر پایهٔ ویژگی‌های ظاهری آن‌ها

فناوری‌های گلخانهٔ هوشمند پیش از هر چیز نیازمند تشخیص خودکار نوع گیاهان مورد نظر است. این تشخیص معمولا به کمک ویژگی‌های ظاهری گیاهان صورت می‌گیرد. دادگان پیوست (فایل iris.csv) شامل پنج ستون است که چهار ستون اول ورودی‌ها (ویژگی‌ها)ی گیاه شامل طول و عرض کاسبرگ (sepal) و طول و عرض گلبرگ (petal) و ستون پنجم خروجی یا دسته (class) گیاه را نشان می‌دهد.

الف) ابتدا ۸۰٪ داده‌ها (شامل ۸۰٪ از داده‌های هر دسته) را برای آموزش و ۲۰٪ باقی مانده (شامل ۲۰٪ از داده‌های هر دسته) را برای آزمایش در نظر بگیرید. سپس به کمک الگوریتم KNN (K نزدیک ترین همسایه) با K= 5 مدل را برای دسته‌بندی گیاهان تربیت کنید و با محاسبهٔ خروجی‌های مدل برای داده‌های آزمایش و تشکیل ماتریس سردرگمی، امتیازهای accuracy, recall, precision و jaccard را محاسبه کنید. (برای آشنایی با امتیاز jaccard عبارت "multiclass jaccard similarity score" را جستجو کنید).

ب) با اعمال یک نرمال سازی استاندارد روی داده‌ها، خواسته‌های بند الف را (با همان روش و با همان مقدار K) مجددا بدست آورید. آیا امتیاز دقت (accuracy) تغییر می‌کند؟ چرا؟

توجه: از آن جا که در حالت‌های چندکلاسه، امتیازهای recall، precision و jaccard از روش‌های میانگیری مختلفی محاسبه می‌شوند، برای این مسئله از روش میانگیری macro (اهمیت برابر دقت در هر کلاس) استفاده کنید.

بخش دوم: پیش‌بینیِ گونه (genre) و میزان محبوبیت آهنگ‌های جدید

یک سایت موسیقی برای افزایش مخاطبان خود قصد دارد آهنگ‌های منتشر شده بین سال‌های ۲۰۰۰ تا ۲۰۲۰ را از نظر گونه (genre) و میزان محبوبیت رتبه‌بندی کند. برای این کار از ویژگی‌هایی همچون نام آهنگ، نام خواننده، سال انتشار، انرژی، ریتم و ضرباهنگ استفاده می‌شود. داده‌های مورد نظر در فایل musics.csv ارایه شده است. این فایل شامل ۱۸ ستون است که دو ستون محبوبیت (popularity) و گونه (genre) خروجی‌های مدل و ۱۶ ستون دیگر ورودی‌ها (ویژگی‌ها) را نشان می‌دهند.

الف) ابتدا ستون‌های نام آهنگ، سال انتشار و نام خواننده را از دادگان حذف کنید (علت را توضیح دهید). در ستون محبوبیت، میزان محبوبیت هر آهنگ با عددی بین ۰ (کمترین محبوبیت) و ۱۰۰ (بیشترین محبوبیت) نشان داده شده است. برای سادگی، میزان محبوبیت را به پنج دسته تقسیم کنید، به این صورت که اعداد ۰ تا ۲۰ را با ۱ (نامحبوب)، اعداد ۲۰ تا ۴۰ را با ۲ (محبوبیت کم) و . . . جایگزین کنید.

ب) ورودی‌ها و خروجی‌های کیفی را به مقادیر کمی تبدیل کنید. تنها ستون‌های مربوط به ویژگیِ صریح بودن (explicit) و خروجیِ گونه (genre) کیفی هستند. ویژگی صریح بودن تنها دارای دو مقدار درست و نادرست است که می‌توانید آن‌ها را به ترتیب به ۰ و ۱ تبدیل کنید. خروجیِ گونه نیز می‌تواند به روش زیر کمی‌سازی شود:

- یک ستون جدید در دادگان به نام new_genre بسازید.
- هر گونه با یک ویرگول از باقی گونه‌ها جدا شده است. سعی کنید همهٔ گونه‌های موجود را پیدا کنید.

- تعداد کل گونه‌های موجود را محاسبه کنید و برای هر موسیقی در ستون new_genre یک بردار صفر به اندازهٔ تعداد کل گونه‌های موجود بسازید که هر درایهٔ آن نشان دهندهٔ یک گونهٔ خاص است.
- برای هر موسیقی، درایه‌های مربوط به آن گونه را برابر ۱ قرار دهید.

(برای توضیحات بیشتر در این مورد می‌توانید عبارت one hot encoding را در اینترنت جستجو کنید.)

ج) توزیع دادگان را برای تمام ۱۳ ویژگی (همهٔ ستون‌ها بجز ستون‌های محبوبیت و گونه و سه ویژگی حذف شده) بدست آورید. داده‌های با درجه محبوبیت متفاوت را با رنگ های متفاوت نمایش دهید. پراکندگی داده‌ها را بر حسب ویژگی‌های مختلف نشان دهید. بر پایهٔ این پراکندگی‌ها برداشت خود را از ماهیت داده‌ها بیان کنید (راهنمایی: برای نمایش توزیع داده‌ها میتوانید از نمودارهای نقشه گرمایی، گسسته، هیستوگرام یا جعبه‌ای استفاده کنید.)

د) ستون گونه را کنار گذاشته و در صورت نیاز تمام ستون ها به جز محبوبیت را نرمال‌سازی کنید. سپس رابطه و تاثیرگذاری هر ویژگی را بر میزان محبوبیت آهنگ پیدا کنید. (راهنمایی: می‌توانید از معیارهای آماری مانند همبستگی یا correlation استفاده کنید تا میزان تاثیر داده‌ها بر روی خروجی و حتی رابطهٔ آن‌ها با یکدیگر را مشاهده کنید. برای نمایش هم می‌توانید از pair-plot در کتابخانهٔ seaborn یا از نقشهٔ گرمایی استفاده کنید.)

ه) داده ها را به سه بخش آموزش (training)، آزمایش (test) و ارزیابی (validation) تقسیم کنید. پیشنهاد می‌شود ٪۸۰ کل داده‌ها به آموزش، ٪۱۰ به آزمایش و ٪۱۰ به ارزیابی اختصاص داده شود.

و) با استفاده از داده‌های آموزش و به کمک الگوریتم‌های درخت تصمیم‌گیری، جنگل تصادفی، k نزدیک‌ترین همسایه و بردار پشتیبان، مدل خود را برای پیش‌بینی محبوبیت آهنگ‌ها تربیت کنید. پس از آن، عملکرد مدل را بر روی داده‌های آزمایش (با استفاده از ماتریس سردرگمی) بررسی کنید. بهترین مدل را انتخاب کرده و دلیل برتری آن را بر مدل‌های (الگوریتم‌های) دیگر شرح دهید. آیا دقت بهترین مدل را مناسب می‌دانید یا خیر؟ پیشنهادهای خود را برای افزایش دقت ارائه دهید.

ز) **امتیازی**: بعد از آموزش، مواردی را که به غلط توسط مدل پیش‌بینی شده جدا کنید. با استفاده از تحلیل آماری یا تفسیر بصری علت این پیش‌بینی غلط را توضیح دهید و با ذکر دلیل، در جهت بهبود دقت مدل تلاش کنید.

ح) این بار ستونِ محبوبیت را نادیده گرفته و بجای آن ستونِ گونه را در نظر بگیرید. در این ستون پر تکرارترین و کم تکرارترین گونه را پیدا کنید. طبیعتا هر آهنگ میتواند همزمان دارای گونه‌های مختلفی باشد. بیشترین تعداد گونه مربوط به یک آهنگ را پیدا کنید. اگر آهنگی برای مثال در دو گونهٔ pop و hip hop به صورت همزمان قرار گرفته باشد، درصد احتمال قرارگیری در ژانر R&B را محاسبه کنید. (راهنمایی: از قانون بیز استفاده کنید.)

ت) **امتیازی:** برای پیش‌بینی گونهٔ موسیقی، دادگان را مانند قسمت (ه) به ۳ بخش تقسیم کنید. پس از تغییر و تبدیل دادگان، با استفاده از الگوریتم درخت تصمیم‌گیری مدل خود را تربیت کنید و در پایان ماتریس سردرگمی را تشکیل داده و تحلیل کنید.

چند توضیح:

- برای یادگیری مفاهیمی که در تمرین مطرح شده و احتمالا تدریس نشده‌اند از منابع موجود در اینترنت استفاده کنید.
- برای انجام بخش‌های مختلف تمرین می‌توانید از کتابخانه‌های آماده‌ای مانند numpy، matplotlib ،pandas ، sklearn و seaborn استفاده کنید.

- تحویل گزارش این تمرین ضروری است و به تمرین بدون گزارش نمره‌ای تعلق نمی‌گیرد. حجم گزارش معیاری برای ارزیابی نخواهد بود و لزومی به توضیح جزئیات کد نیست؛ اما از آنجا که برای این تمرین از کتابخانه‌های موجود استفاده می‌کنید لطفا تمامی پارامترهای تنظیم‌شده در هر قسمت از کد را گزارش کرده و فرض‌هایی را که برای پیاده‌سازی‌ها و محاسبات خود به کار برده‌اید ذکر کنید. از ارائهٔ توضیحات کلیشه‌ای و همانند بَرداری از منابع موجود بپرهیزید.
- در فرایند ارزیابی گزارش، کدهای شما لزوما اجرا نخواهد شد. بنابراین همهٔ نتایج و تحلیل‌های خود را به‌طور کامل ارائه کنید.
- شباهت بیش از حد گزارش و کدها باعث از دست دادن نمرهٔ تمرین خواهد شد. همچنین گزارش‌هایی که در آنها از کدهای آماده استفاده شده باشد پذیرفته نخواهند شد.
- گزارش شما باید به صورت تایپ شده و با فرمت pdf ارائه شود و کدهایی که به همراه گزارش تحویل می‌دهید باید قابل اجرا باشند. در انتها تمامی فایل‌های لازم را در یک فایل zip یا rar بارگذاری و ارسال کنید.
- در صورت استفاده از گیت هاب جهت ارائهٔ گزارش و کد، نمره امتیازی به شخص تعلق میگیرد.
- پرسش‌های خود را از طریق ایمیل یا تلگرام از دستیاران آموزشی مربوطه بپرسید:

	ایمیل	تلگرام
محمد مهدی عابدی	mohammadabedi@ut.ac.ir	@mohammadabedi1179
پارسا شفیعی	parsa.shafiei@ut.ac.ir	@blind_side