مینی‌پروژهٔ شمارهٔ سه

۱ پرسش یک: مفاهیم پایهٔ درخت تصمیم

در این پرسش می‌خواهیم با مفهوم کلی درخت تصمیم و اجزای آن آشنا شویم.

- توضیح دهید درخت تصمیم چیست و چرا می‌تواند یک ساختار مناسب برای یادگیری داده‌ها باشد.
- نقش گرهٔ میانی، شاخه و برگ را در درخت تصمیم توضیح دهید.
- یک مثال واقعی از دنیای واقعی (مثلاً سیستم اعتباردهی بانکی، تشخیص بیماری، پیش‌بینی قبولی در کنکور و ...) بزنید که درخت تصمیم برای حل آن مسئله مناسب باشد و توضیح دهید ورودی‌ها و خروجی‌های درخت چه خواهند بود.

۲ پرسش دو: محاسبهٔ آنتروپی و Information Gain

جدول دادهٔ زیر را در نظر بگیرید:

Play	Wind	Humidity	Temp	Outlook
No	Weak	High	Hot	Sunny
No	Strong	High	Mild	Sunny
Yes	Weak	Normal	Cool	Overcast
Yes	Weak	High	Mild	Rain
Yes	Strong	Normal	Cool	Rain

آ) آنتروپی کل دادگان را نسبت به متغیر Play محاسبه کنید.

ب) مقدار Information Gain ویژگی Outlook را برای این داده محاسبه کنید.

ج) توضیح دهید چرا در الگوریتم ID3 همواره ویژگی‌ای انتخاب می‌شود که بیشترین Information Gain را دارد؛ از منظر کاهش عدم‌قطعیت و بهبود تفکیک کلاس‌ها تحلیل کنید.

۳ پرسش سه: مزایا و معایب درخت تصمیم

درخت تصمیم یک مدل قدرتمند اما با محدودیت‌هایی است.

آ) چهار مورد از مشکلات و معایب درخت تصمیم را توضیح دهید. (به‌عنوان مثال: بیش‌برازش، ناپایداری نسبت به تغییرات کوچک داده، تمایل به ویژگی‌های با تعداد مقادیر زیاد، پیچیده‌شدن بیش از حد برای مرزهای غیرخطی و ...)

ب) چهار مورد از مزایا و نقاط قوت درخت تصمیم را بیان کنید. (مثلاً تفسیرپذیری، کار با داده‌های عددی و طبقه‌ای، نیاز کم به پیش‌پردازش، سرعت مناسب و ...)

۴ پرسش چهار: هرس کردن درخت تصمیم

آ) هرس پیش‌گیرانه (Pre-pruning) و هرس پس از ساخت (Post-pruning) را تعریف کرده و تفاوت‌های اصلی آن‌ها را توضیح دهید.

ب) یک مثال مفهومی ارائه دهید که نشان دهد چرا هرس پس از ساخت معمولاً می‌تواند به مدلی با تعمیم‌پذیری بهتر نسبت به هرس پیش‌گیرانه منجر شود.

۵ پرسش پنج: پیاده‌سازی درخت تصمیم روی دادگان **Titanic** (بخش اول)

در این پرسش، قصد داریم یک مدل درخت تصمیم را روی دادگان Titanic از وب‌گاه Kaggle پیاده‌سازی کنیم.

الف) بارگذاری و پیش‌پردازش داده‌ها

- فایل train.csv مربوط به مسئلهٔ Titanic – Machine Learning from Disaster را دانلود کرده و در محیط Python (۱OCZPZTqYox۵GpVkS۴ffOXR۰C۶eF-۴I۹R) بارگذاری کنید.
- ستون‌های دارای مقادیر گمشده (مانند Age، Fare و Embarked) را با روش مناسب (میانه، میانگین یا مد) پر کنید.
- ستون‌های غیرقابل استفاده‌ای مانند Name، Ticket و Cabin را از دادگان حذف کنید.
- ویژگی‌های طبقه‌ای (مانند Sex و Embarked) را با استفاده از One-Hot Encoding یا روشی مشابه به ویژگی‌های عددی تبدیل کنید.
- دادگان را به دو بخش Train و Test (با نسبت ۸۰/۲۰) تقسیم کنید.

ب) آموزش درخت کم‌عمق با $\mathbf{max\_depth = 4}$

- یک مدل DecisionTreeClassifier با $\text{max\_depth} = 4$ آموزش دهید.
- دقت مدل روی دادهٔ آموزش و دادهٔ آزمون را گزارش کنید.
- تعداد گره‌ها و عمق درخت را چاپ کنید.

۶ پرسش شش: پیاده‌سازی درخت تصمیم روی دادگان **Titanic** (بخش دوم: عمق و هرس)

در این پرسش ادامهٔ کار پرسش قبل را انجام می‌دهید.

الف) آموزش درخت کامل (بدون محدودیت عمق)

- یک مدل DecisionTreeClassifier با $\text{max\_depth} = \text{None}$ (بدون محدودیت عمق) آموزش دهید.
- دقت مدل روی دادهٔ آموزش و آزمون را گزارش کنید.
- تعداد گره‌ها و عمق درخت را چاپ کنید.
- تفاوت این مدل را با مدل کم‌عمقِ پرسش پنج از نظر Overfitting و پیچیدگی مدل تحلیل کنید.

ب) بررسی عمق‌های مختلف درخت

- مدل درخت تصمیم را با مقادیر مختلف max_depth برابر با ۳، ۵، ۱۰ و None آموزش دهید.
- برای هر مقدار، دقت آموزش و آزمون را گزارش کنید.
- یک نمودار از تغییر دقت (روی آموزش و آزمون) نسبت به max_depth رسم کرده و توضیح دهید در کدام بازه‌ها مدل دچار Underfitting و در کدام بازه‌ها دچار Overfitting می‌شود.

ج) هرس کردن درخت با **Cost-Complexity Pruning**

- با استفاده از تابع cost_complexity_pruning_path مجموعه‌ای از مقادیر ccp_alpha را به‌دست آورید.
- تنها هایalpha مثبت را در نظر بگیرید (مقدار صفر را حذف کنید) و برای هر ccp_alpha یک درخت تصمیم آموزش دهید.
- دقت آموزش و آزمون را برای این مدل‌ها محاسبه کرده و دقت‌ها را بر حسب ccp_alpha (در مقیاس لگاریتمی) رسم کنید.
- یک مقدار ccp_alpha مثبت انتخاب کنید که در آن:
  - تعداد گره‌ها و عمق درخت نسبت به درخت کامل کاهش معنی‌دار داشته باشد،
  - و دقت روی دادهٔ آزمون تقریباً ثابت یا با کاهش بسیار اندک باشد.
- برای مدل انتخاب‌شده، تعداد گره‌ها، عمق درخت و دقت آموزش و آزمون را گزارش کرده و آن را با درخت کامل و درخت کم‌عمق مقایسه کنید.
- در نهایت، سه مدل زیر را از نظر پیچیدگی، تعمیم‌پذیری و تفسیرپذیری مقایسه و تحلیل کنید:
  (i) درخت کم‌عمق با $\text{max\_depth} = 4$،
  (ii) درخت کامل بدون هرس،
  (iii) درخت هرس‌شده با ccp_alpha انتخاب‌شده.

۷ پرسش یک: مفاهیم پایهٔ Learning Ensemble

آ) توضیح دهید Learning Ensemble چیست و چرا ترکیب چند مدل می‌تواند عملکرد بهتری نسبت به یک مدل تکی داشته باشد.

ب) تفاوت دو مفهوم Diversification و Aggregation را در هاEnsemble توضیح دهید.

ج) یک مثال واقعی از کاربرد Ensemble (مانند تشخیص تقلب بانکی، سیستم‌های توصیه‌گر یا تحلیل پزشکی) ارائه دهید و ورودی‌ها و خروجی‌های آن را توضیح دهید.

۸ پرسش دو: Bagging و Forest Random

داده‌های زیر را در نظر بگیرید:

x	y
۲	۰
۳	۰
۸	۱
۹	۱

آ) توضیح دهید در روش Bagging چگونه Bootstrap Sampling انجام می‌شود.

ب) اگر سه مجموعهٔ Bootstrap ساخته شود، یک نمونهٔ احتمالی از سه Bootstrap (مثلاً B1, B2, B3) بنویسید.

ج) توضیح دهید چرا Random Forest معمولاً عملکرد بهتری نسبت به یک درخت تصمیم تکی دارد.

۹ پرسش سه: Boosting

۱. مفهوم Boosting را توضیح دهید و بیان کنید تفاوت آن با Bagging چیست.

۲. نقش وزن نمونه‌ها در الگوریتم AdaBoost چیست؟

۳. توضیح دهید چرا Boosting می‌تواند مدل‌های بسیار قوی تولید کند اما در عین حال نسبت به نویز در داده‌ها حساس است.

۱۰ پرسش چهار: Stacking

آ) Stacking چیست و چگونه از یک مدل سطح دوم (Meta Learner) استفاده می‌کند؟

ب) چرا برای آموزش Learner Meta باید از Cross-Validation استفاده شود؟

ج) یک مثال از معماری یک سیستم Stacking ارائه دهید.

۱۱ پرسش پنج: مزایا و معایب Ensemble

آ) چهار مورد از مزایای Learning Ensemble را بیان کنید.

ب) چهار مورد از معایب Learning Ensemble را توضیح دهید.

۱۲ پرسش شش: پیاده‌سازی Ensemble روی داده‌های **Titanic**

در این تمرین قصد داریم دو مدل Ensemble روی داده‌های Titanic بسازیم.

۱. داده را پیش‌پردازش کنید: حذف ستون‌های غیرضروری، پرکردن مقادیر گمشده و انجام One-Hot Encoding.

۲. یک RandomForestClassifier با $n\_estimators = 200$ آموزش دهید و دقت Train/Test را گزارش کنید.

۳. مدل GradientBoostingClassifier را آموزش دهید و نتایج آن را مقایسه کنید.

۴. تحلیل کنید کدام مدل دچار Overfitting یا Underfitting شده است و دلیل آن چیست.

در انجام این تمرین حتماً به نکات زیر توجه کنید:

- موعد تحویل این تمرین، ساعت ۱۸:۰۰ روز ۳۰ آذرماه ۱۴۰۴ است.
- برای گزارش لازم است که پاسخ هر سؤال و زیربخش‌هایش به‌ترتیب و به‌صورت مشخص نوشته شده باشند. بخش زیادی از نمره به توضیحات دقیق و تحلیل‌های کافی شما روی نتایج بستگی خواهد داشت.
- لازم است که در صفحهٔ اول گزارشِ خود لینک مخزن گیت‌هاب و گوگل‌کولبِ مربوط به مینی‌پروژهٔ خود را درج کنید. درخصوص گیت‌هاب، یک مخزنِ خصوصی درست کنید و آی‌دی‌های `MJAHMADEE`، `ParisaGhorbani`، `AliBagheriNejad` و `erfanY2AJ` را به‌عنوان Collaborator به مخزن اضافه کنید. پروژه‌های گیت‌هاب می‌بایست در انتهای ترم پابلیک شوند. در مقابل، لینک گوگل‌کولب را در حالتی که دسترسی عمومی دارد به اشتراک بگذارید. دفترچهٔ گوگل‌کولب باید به‌صورت منظم و با بخش‌بندیِ مشخص تنظیم شده باشد و خروجیِ سلول‌های اجراشده قابل مشاهده باشد. در گیت‌هاب نیز یک مخزن برای درس و یک پوشهٔ مجزا برای هر مینی‌پروژه ایجاد کنید.
(آموزش پرایوت‌کردن مخزن گیتهاب و آموزش افزودن Collaborator به مخزن گیت‌هاب)
- چرا از دفترچهٔ گوگل‌کولب؟ شما نباید به فراخوانیِ فایلی خارج از محیط نیاز داشته باشید؛ مطابق آموزش‌های ارائه‌شده ملزم هستید از دستور gdown استفاده کنید و مسیرهای فایل‌ها را طوری تنظیم کنید که صرفاً با اجرای سلول‌ها، امکان فراخوانی و خواندن فایل‌ها توسط هر کاربری وجود داشته باشد.
- در تمامی مراحلِ تعریف داده و مدل و هر جای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقمِ آخرِ شمارهٔ دانشجوییِ خود در نظر بگیرید.
- استفاده از ابزارهای هوشمند (مانند ChatGPT) در کمک‌گرفتن برای بهبود کدها مجاز است؛ امّا لازم است تمام جزئیات مواردی که در خروجی‌های مختلفِ گزارش خود عنوان می‌کنید را به‌خوبی خوانده، درک و تحلیل کرده باشید. استفاده از این ابزارهای هوشمند در نوشتنِ گزارش و تحلیل‌ها ممنوع است.
- در جاهایی که با توجه به دو رقمِ آخر شمارهٔ دانشجوییِ خود محدود به انتخاب عددی معین یا داده‌ای خاص شده‌اید، برای تست‌های اضافه‌تر و نمایشِ بهبود در نتایجِ خود، مجاز هستید از مقادیر دیگر هم استفاده کنید.
- رعایت نکات بالا به حرفه‌ای‌تر شدن شما کمک خواهد کرد و اهمیتی معادلِ مطالبِ درسیِ فراگرفته‌شده دارد؛ بنابراین، در صورت عدم رعایت یک یا چند مورد از این نکات، از نمرهٔ تمرین شما کاسته خواهد شد.
- آی‌دیِ پرسشِ هرگونه سؤال درخصوص مینی‌پروژهٔ شمارهٔ ۳

منابع

[1] Titanic – Machine Learning from Disaster