مبانی یادگیری ماشین

پاییز ۱۴۰۳

استاد: دکتر حامد ملک

دانشکده مهندسی و علوم کامپیوتر

تمرین شماره ۱

- قبل از شروع تمرین، فایل مربوط به قوانین حل و تحویل تمرین‌ها را مطالعه کنید.
- سؤالات و مشکلات خود را درباره این تمرین می‌توانید در گروه تلگرامی درس یا با طراحان این تمرین مطرح کنید.
- نوشتن گزارش کامل و با جزئیات و تفسیر نتایج اجباری است. جزئیاتی مانند روش‌های مورد استفاده، تأثیر هر روش در نتیجه نهایی و بهبود حاصل‌شده به همراه ارائه معیارهای ارزیابی در گزارش ضروری است. با هر تغییر و هر بهبود، تغییر مقادیر معیارهای ارزیابی نیز ذکر شود.
- پاسخ سؤالات تشریحی که در بخش سؤالات عملی مطرح شده‌اند را در گزارش خود ذکر کنید.
- برای سؤالات عملی می‌توانید از کتاب‌خانه‌هایی نظیر `pandas`، `numpy` و `matplotlib` برای کار با دادگان و تولید نمودارها استفاده کنید.
- در سؤال ۱ بخش عملی، استفاده از مدل‌های آماده آزاد است.
- برای دریافت دادگان‌های سؤالات عملی، از راهنمای دریافت دادگان‌ها که در سامانه درس‌افزار قرار داده شده است کمک بگیرید و دادگان‌ها را به‌صورت دستی دانلود نکنید.
- مهلت ارسال پاسخ‌ها: پنج‌شنبه ۱۷ آبان ساعت ۲۳:۵۹
- طراحان این تمرین: آرمان غفارنیا — مرتضی مهدوی

سؤالات تئوری

۲۵+۵ نمره

سؤال ۱ (۱۵ نمره)

فرض کنید که شما وظیفه دارید مشخص کنید که آیا یک شخص محصولی را بر اساس دو ویژگی سن و دانشجو بودن خریداری می‌کند یا خیر. به جدول داده‌های زیر که از ۸ شخص مختلف گرفته شده است، توجه کنید.

سن	دانشجو	خرید محصول
جوان	خیر	✗
جوان	خیر	✗

جوان	بله	✅
میان‌سال	خیر	✅
پیر	خیر	✅
پیر	بله	✅
پیر	بله	❌
میان‌سال	بله	✅
جوان	خیر	❌
پیر	خیر	✅

الف) آنتروپی مجموعه داده‌ها را با توجه به متغیر هدف «خرید محصول» محاسبه کنید. (۲ نمره)

ب) آنتروپی «خرید محصول» را به شرط دانشجو بودن یا نبودن شخص محاسبه کنید. (۵ نمره)

پ) میزان کسب اطلاعات[1] را برای تقسیم مجموعه داده‌ها بر اساس سن (بر اساس هر یک از دسته‌بندی‌های سنی) محاسبه کنید. (۶ نمره)

ت) با توجه به محاسبات بخش قبل، اگر بخواهیم درخت تصمیم را یک مرحله توسعه دهیم، بهتر است تقسیم‌بندی را بر اساس سن انجام دهیم یا دانشجو بودن؟ دلایل خود را توضیح دهید. (۲ نمره)

سؤال ۲ (۵ نمره)

در هنگام استفاده از الگوریتم کِی‌ان‌ان[2] ممکن است با مسائلی روبه‌رو شویم؛ برای مثال، می‌توان به مسئله تعداد زیاد ابعاد[3] و مسئله مقیاس‌بندی ویژگی‌ها[4] اشاره کرد. هر کدام از این مشکلات را توضیح دهید و راهکارهای رفع آن‌ها را نام ببرید.

سؤال ۳ (۵ نمره)

جنگل تصادفی[5] یک روش یادگیری گروهی است که برای طبقه‌بندی[6]، رگرسیون و وظایف دیگر استفاده می‌شود. درباره نحوه عملکرد آن و مزایای آن نسبت به درخت تصمیم توضیح دهید.

---

[1] Information Gain
[2] K-Nearest Neighbors (KNN)
[3] Curse of Dimensionality
[4] Feature Scaling
[5] Random Forest
[6] Classification

سؤال ٤ (٥ نمره امتیازی)

دو الگوریتم بهبودیافته مبتنی بر کی‌ان‌ان، کی‌ان‌ان وزن‌دار[1] و ان‌سی‌ای[2] نام دارند. هر دو الگوریتم و نحوه کارکرد هر کدام را توضیح دهید.

سؤالات عملی — ۱۵+۷۵ نمره

سؤال ۱ (۱۵+۲۵ نمره)

دادگان data1.csv در اختیارتان قرار گرفته و هدف این است که با توجه به اطلاعات بیماران، پیش‌بینی شود که آیا مبتلا به دیابت هستند یا خیر. این دادگان شامل چند متغیر پیش‌بینی پزشکی و یک متغیر نتیجه (ستون آخر) است. متغیرهای پیش‌بینی شامل موارد مختلفی است از جمله تعداد دفعاتی شخص باردار بوده، شاخص توده بدنی[3]، سطح انسولین، سن و ... .

الف) یک درخت تصمیم را برای پیش‌بینی اینکه آیا بیمار دیابت دارد یا خیر، روی دادگان آموزش دهید و پس از مرحله تست برای مدل خود صحت[4]، دقت[5] و فراخوانی[6] را گزارش کنید. (۱۰ نمره)

ب) دو ابرپارامتر[7] از ابرپارامترهای مدل خود را که با تغییر آن‌ها تغییر محسوسی در صحت نتایج ایجاد می‌شود، انتخاب کنید و حداقل ۱۰ مقدار متفاوت برای هر کدام امتحان کنید و بهترین مقدار را گزارش کنید. میتوانید این تغییرات را با نمودار نشان دهید. (۱۵ نمره)

پ) یک مدل جنگل تصادفی را بر روی دادگان آموزش دهید. موارد بالا را مجدد انجام دهید، نتیجه تغییرات صحت نتایج را با تغییر ابرپارامترها نشان دهید و با تغییرات درخت تصمیم در بخش قبل مقایسه کنید. (۱۵ نمره امتیازی)

سؤال ۲ (۵۰ نمره)

سکته مغزی که به آن حادثه عروق مغزی[8] نیز گفته می‌شود، زمانی رخ می‌دهد که بخشی از مغز از دریافت خون محروم شود؛ در نتیجه، ناحیه‌ای از بدن که سلول‌های مغزی آن را کنترل می‌کنند، از کار می‌افتد. این محرومیت از خون می‌تواند

---

[1] Weighted KNN
[2] Neighborhood Component Analysis (NCA)
[3] Body Mass Index (BMI)
[4] Accuracy
[5] Precision
[6] Recall
[7] Hyperparameter
[8] CVA

به دلیل کاهش جریان خون یا خون‌ریزی در بافت مغز باشد. سکته مغزی یک وضعیت اضطراری پزشکی است زیرا ممکن است منجر به مرگ یا ناتوانی دائمی شود. درمان‌هایی برای این وضعیت وجود دارد، اما باید در چند ساعت اولیه پس از بروز علائم سکته آغاز شود. دادگان `data2.csv` در اختیار شما قرار داده شده که باید برای حل مسئله مورد استفاده قرار گیرد.

الف) در بیشتر مسائل واقعی، تعداد داده‌های کلاس‌های مختلف یکسان و متوازن نیست. سه مورد از راهکارهای موجود برای رفع این مشکل را نام ببرید و نحوه عملکرد هر یک را شرح دهید. (۶ نمره)

ب) پیش‌پردازش و مهندسی ویژگی را با استفاده از تکنیک‌های موجود و مواردی که در کلاس آموخته‌اید انجام دهید. برای این منظور می‌توانید از روش‌هایی مانند مقیاس‌بندی، نرمال‌سازی، استانداردسازی، کدگذاری تک‌روشن[1] و ... و همچنین تبدیل ویژگی‌های غیرعددی به عددی استفاده کنید. در نظر داشته باشید ممکن است هر روشی الزاماً منجر به بهبود نشود. (۵ نمره)

پ) دادگان را با استفاده از یکی از روش‌های گفته‌شده در قسمت الف متوازن کنید و به دو بخش آموزش و تست تقسیم کنید. تقسیم‌بندی به‌صورتی باشد تعداد مناسبی از داده‌های هر کلاس در بخش‌های آموزش و تست قرار گیرد. ترجیحاً از یک `random state` ثابت استفاده کنید. (۵ نمره)

ت) الگوریتم کی‌ان‌ان را بدون استفاده از توابع آماده و از صفر پیاده‌سازی کنید. این الگوریتم را روی دادگان اعمال کنید و عملکرد آن را به‌کمک `classification report` از بخش `metrics` کتابخانه `sklearn` گزارش دهید. (۱۲ نمره)

ث) بدون اینکه تقسیم‌بندی دادگان خود را تغییر دهید (با همان دادگان آموزش و تست ایجادشده)، روش‌های مهندسی ویژگی را تغییر دهید به‌گونه‌ای که در نتایج بهبود ایجاد شود. سپس تغییرات در روش‌های مهندسی ویژگی را گزارش دهید. (۱۰ نمره)

[1] One-hot Encoding

ج) مقادیر کِی را تغییر دهید تا به بهترین کِی ممکن دست یابید. همچنین دو روش دیگر محاسبه فاصله را تست کنید و توضیح دهید که نتایج چه تغییری می‌کنند. در این بخش می‌توانید برای راحتی کار از مدل آماده کی‌ان‌ان استفاده کنید. (۱۲ نمره)

سالم و موفق باشید.