دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

تمرین دوم درس داده کاوی

نگارش
مهدیه سادات بنیس

استاد درس
دکتر احسان ناظرفرد

نیم سال دوم ۱۴۰۱

• بخش تئوری:

سوال اول

یکی از مباحثی که در درخت تصمیم مطرح میشود هرس درخت برای جلوگیری از بیش برازش است. توضیح دهید چرا نمیتوان از مجموعه داده جدا برای هرس درخت استفاده کرد؟ منظور این است که داده‌هایی که برای هرس استفاده میشوند با مجموعه داده ای که برای ساخت درخت استفاده میشود یکسان نباشند.

وقتی نوبت به هرس درخت می‌رسد، مهم است که از یک مجموعه داده برای ساخت درخت و هرس آن استفاده کنید. زیرا هرس فرآیندی است که هدف آن کاهش پیچیدگی درخت تصمیم با حذف شاخه ها یا گره های غیر ضروری است.

اگر بخواهیم از مجموعه داده‌های جداگانه برای رشد و هرس درخت استفاده کنیم، می‌تواند منجر به ناسازگاری‌ها و تصمیم‌های بالقوه هرس نابهینه شود. چند دلیل وجود دارد که چرا مجموعه داده های جداگانه برای هرس درختان مناسب نیستند:

- توزیع داده ناسازگار: مجموعه داده مورد استفاده برای هرس ممکن است توزیع یا ویژگی های متفاوتی در مقایسه با مجموعه داده مورد استفاده برای رشد درخت داشته باشد. این عدم تطابق می‌تواند منجر به تصمیم‌گیری‌هایی شود که معرف داده‌های واقعی نیستند، که منجر به تعمیم و عملکرد ضعیف در داده‌های دیده نشده می‌شود.

- از دست دادن اطلاعات: تصمیمات هرس بر اساس معیارهای آماری مختلفی مانند افزایش اطلاعات یا ناخالصی جینی است که به فراوانی و توزیع نقاط داده بستگی دارد. اگر از مجموعه داده های جداگانه استفاده شود، فرآیند هرس ممکن است به طور دقیق روابط و الگوهای مشاهده شده در طول رشد درخت را نشان ندهد و منجر به از دست رفتن اطلاعات مهم شود.

- تصمیمات هرس غیربهینه: هدف هرس حذف شاخه ها یا گره های غیر ضروری است که به قدرت پیش بینی درخت کمک قابل توجهی نمی کنند. با استفاده از یک مجموعه داده جداگانه، الگوریتم هرس ممکن است به طیف کاملی از داده‌های لازم برای تصمیم‌گیری آگاهانه در مورد اینکه کدام بخش از درخت باید هرس شود، دسترسی نداشته باشد. این می تواند منجر به هرس غیربهینه شود، جایی که شاخه ها یا گره های مهم حفظ می شوند یا برعکس.

به طور خلاصه، استفاده از مجموعه داده های جداگانه برای هرس درختان می تواند منجر به تناقضات، از دست دادن اطلاعات و تصمیمات هرس نابهینه شود. استفاده از مجموعه داده یکسان برای رشد و هرس درخت ضروری است تا از هرس منسجم و مؤثری که عملکرد پیش‌بینی درخت را به حداکثر می‌رساند، اطمینان حاصل شود.

سوال دوم

با توجه به مطالب تدریس شده در کلاس، برای داده‌های زیر یک درخت تصمیم درست کنید. (ذکر تمام مراحل و توضیح آنها لازم است)

آیا به مهمانی دعوت می شود؟	وزن	قد	رنگ لباس
خیر	لاغر	۱۷۰	قرمز
بله	چاق	۱۶۲	آبی
خیر	چاق	۱۶۵	سبز
بله	لاغر	۱۷۲	سبز
بله	لاغر	۱۶۰	آبی

$$H(D) = -(\frac{2}{5}lg(\frac{2}{5}) + \frac{3}{5}lg(\frac{3}{5})) = 0.97$$

$$Inforamtion\ Gain(color) = H(D) - H(D)_{color} = H(D) - (\frac{1}{5} \times 0 + \frac{2}{5} \times 0 + \frac{2}{5} \times -(\frac{1}{2}lg(\frac{1}{2}) + \frac{1}{2}lg(\frac{1}{2})))$$

$$Inforamtion\ Gain(color) = H(D) - H(D)_{color} = H(D) - (\frac{2}{5}) = 0.97 - 0.4 = 0.57$$

$$Inforamtion\ Gain(weight) = H(D) - H(D)_{weight} = H(D) - (\frac{3}{5} \times -(\frac{2}{3}lg(\frac{2}{3}) + \frac{1}{3}lg(\frac{1}{3})) + \frac{2}{5} \times 1))$$

$$Inforamtion\ Gain(weight) = H(D) - H(D)_{weight} = 0.02$$

گسسته سازی ویژگی قد:

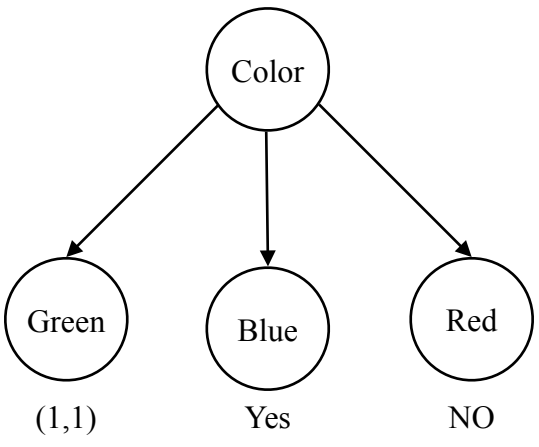
$$H(D)_{163.5} = (\frac{2}{5} \times 0 + \frac{3}{5} \times -(\frac{2}{3}lg(\frac{2}{3}) + \frac{1}{3}lg(\frac{1}{3}))) = 0.551$$

$$H(D)_{171} = (\frac{1}{5} \times 0 + \frac{4}{5} \times 1) = 0.8$$

چون $H(D)_{163.5}$ کمتر است پس Information Gain بیشتری میدهد.

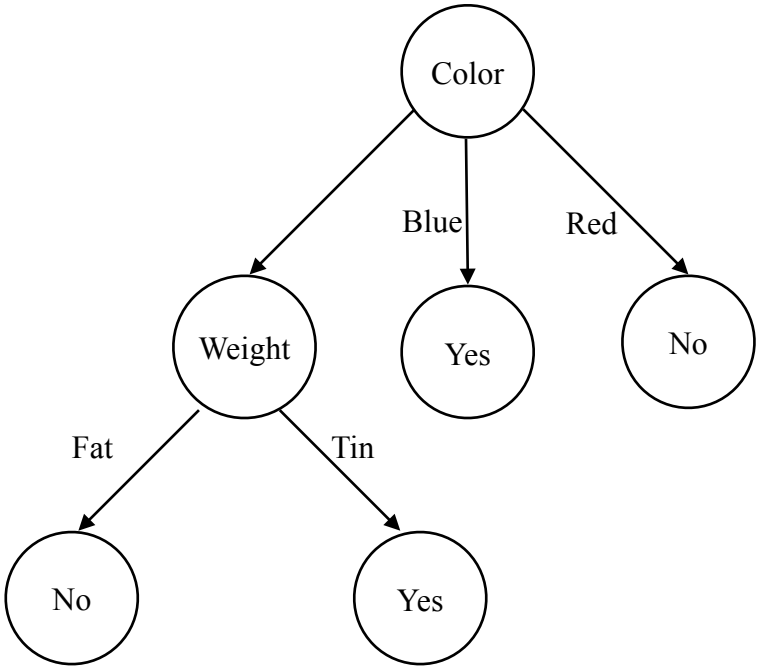
$$Inforamtion\ Gain(hight) = H(D) - H(D)_{hight} = H(D) - (\frac{2}{5}) = 0.97 - 0.551 = 0.419$$

از بین Information Gain سه ویژگی رنگ لباس بیشترین مقدار را داشت پس به عنوان ریشه درخت انتخاب میشود.



$$Inforamtion\ Gain(weight) = H(D)_{color} - H(D)_{weight} = H(D)_{color} - 0 = 1$$
$$Inforamtion\ Gain(hight) = H(D)_{color} - H(D)_{hight} = H(D)_{color} - 1 = 0$$

در مرحله بعد برای مقدار های رنگ لباس قرمز و آبی به برگ رسیدیم و برای حالت سبز ما با محاسبه Information Gain دو ویژگی مقدار وزن بیشتر است و در نهایت به برگ میرسیم و الگوریتم تمام میشود.



سوال سوم

در جدول داده شده زیر با استفاده از قانون بیز برچسب داده زیر را به دست آورید. در صورت صفر شدن احتمال از هموارسازی لاپلاس استفاده کنید.

(معدل = عالی ، مطالعه = بله ، حضور = خیر)

معدل	مطالعه برای امتحان	حضور در کلاس ها	پاس شدن
ضعیف	خیر	خیر	خیر
ضعیف	بله	بله	بله
متوسط	خیر	خیر	خیر
متوسط	بله	بله	بله
عالی	خیر	خیر	بله
عالی	بله	بله	بله

$$p(pass|Average\ grade = A, Study = yes, presence = no) = \frac{p(X|pass)p(pass)}{p(x)}$$

$$\rightarrow p(Average\ grade = A|pass) \times p(Study = yes|pass) \times p(presence = no|pass) \times p(pass)$$

$$= \frac{2}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{4}{6} = \frac{1}{16}$$

$$p(fail|Average\ grade = A, Study = yes, presence = no) = \frac{p(X|fail)p(fail)}{p(x)}$$

$$\rightarrow p(Average\ grade = A|fail) \times p(Study = yes|fail) \times p(presence = no|fail) \times p(fail)$$

$$= 0 \times 0 \times 1 \times \frac{2}{6} = 0$$

با توجه به نتایج به دست آمده برچسب داده پاس شدن است زیرا که احتمال شرطی بالاتری دارد.

سوال چهارم

همانطور که میدانیم یکی از معیارها برای ارزیابی مدلهای یادگیری نظارت شده صحت است. اما این معیار در برخی موارد ممکن است معیار مناسبی برای ارزیابی نباشد. موقعیت هایی که این معیار برای ارزیابی به خوبی عمل نمیکند را توضیح دهید.

دقت همیشه معیار خوبی برای ارزیابی عملکرد مدل‌های تحت نظارت در شرایط زیر نیست:

- داده های وابسته به زمان: هنگام برخورد با داده های وابسته به زمان، جنبه زمان دارای اهمیت فراوانی است. ممکن است دقت به تنهایی برای ثبت عملکرد مدل در طول زمان کافی نباشد. معیارهایی مانند precision, recall, F1 score، که مثبت‌های واقعی، مثبت کاذب و منفی‌های کاذب را در نظر می‌گیرند، می‌توانند بینش بیشتری در مورد اینکه مدل چگونه الگوهای مرتبط را ثبت می‌کند، به خصوص در مواردی که توزیع داده یا نسبت‌های کلاس در طول زمان تغییر می‌کند، ارائه دهد.

- مجموعه داده های نامتعادل: هنگامی که مجموعه داده دارای عدم تعادل کلاس قابل توجهی است، جایی که تعداد نمونه ها در کلاس های مختلف بسیار نامتناسب است، دقت می تواند گمراه کننده باشد. در چنین مواردی، مدلی که به سادگی کلاس اکثریت را برای هر نمونه پیش‌بینی می‌کند، می‌تواند به دقت بالایی دست یابد، حتی اگر نتواند الگوها را در کلاس(های) اقلیت ثبت کند. در سناریوهای نامتعادل، precision, recall, F1 score برای ارزیابی بهتر عمل میکنند.

- دسته بندی حساس به هزینه: در برخی از برنامه ها، هزینه طبقه بندی اشتباه کلاس های مختلف می تواند متفاوت باشد. به عنوان مثال، در تشخیص پزشکی، طبقه بندی اشتباه یک مورد مثبت به عنوان منفی ممکن است عواقب شدیدتری نسبت به طبقه بندی اشتباه یک مورد منفی به عنوان مثبت داشته باشد. دقت با تمام طبقه‌بندی‌های اشتباه به یک اندازه برخورد می‌کند، که ممکن است با هزینه‌های دنیای واقعی همخوانی نداشته باشد. در این موارد، مهم است که معیارهایی را در نظر بگیرید که فاکتور هزینه را در خود جای دهد، مانند دقت حساس به هزینه یا معیارهای وزنی.

- پیش‌بینی‌های مرتب یا پیوسته: دقت برای پیش‌بینی‌های دسته‌بندی طراحی شده است، جایی که خروجی گسسته است و نشان‌دهنده برچسب‌های کلاس است. اگر کار شامل پیش‌بینی مقادیر ترتیبی (مثلاً رتبه‌بندی از ۱ تا ۵) یا مقادیر پیوسته (مثلاً رگرسیون) باشد، دقت کمتر معنادار می‌شود. معیارهای ارزیابی مانند میانگین مربعات خطا (MSE)، میانگین خطای مطلق (MAE) یا ریشه میانگین مربعات خطا (RMSE) برای ارزیابی عملکرد مدل‌ها در این سناریوها مناسب‌تر هستند.

- شدت طبقه‌بندی اشتباه: دقت با همه طبقه‌بندی‌های اشتباه، بدون توجه به بزرگی خطا، به یک اندازه برخورد می‌کند. با این حال، در برخی موارد، طبقه‌بندی نادرست نمونه‌ها در نزدیکی مرز تصمیم ممکن است قابل قبول‌تر از طبقه‌بندی اشتباه نمونه‌هایی باشد که دور از مرز هستند. برای مثال، در تشخیص تقلب، طبقه‌بندی نادرست یک تراکنش مرزی به‌عنوان متقلبانه نسبت به طبقه‌بندی اشتباه یک تراکنش قانونی با ارزش بالا به‌عنوان متقلبانه، شدیدتر است. در چنین مواردی، مدل‌ها باید با استفاده از معیارهایی ارزیابی شوند که شدت طبقه‌بندی‌های اشتباه را نشان می‌دهند، مانند دقت وزنی یا معیارهای مبتنی بر هزینه.

مهم است که معیارهای ارزیابی را انتخاب کنید که با الزامات و ویژگی های خاص حوزه مشکل هماهنگ باشد تا درک جامعی از عملکرد یک مدل به دست آورید.

سوال پنجم

فرض کنید که برای انتخاب پارامتر α در مدل از روش 10 fold cross validation استفاده کرده‌ایم. بهترین روش برای انتخاب مدل نهایی و تخمین ارور کدام است؟

برای پیدا کردن یک پارامتر با استفاده از روش 10 fold cross validation و انتخاب مدل نهایی، می توانید این مراحل را دنبال کنید:

محدوده ای از مقادیر را برای پارامتری که می خواهید تنظیم کنید انتخاب کنید.
مجموعه داده خود را به ۱۰ زیرمجموعه با اندازه مساوی تقسیم کنید.
برای هر مقدار از پارامتری که می خواهید ارزیابی کنید، مراحل زیر را انجام دهید:
آ. مدل را با مقدار پارامتر انتخابی مقداردهی اولیه کنید.
ب برای هر فولد (از ۱ تا ۱۰) موارد زیر را انجام دهید:

i. از فولد فعلی به عنوان validation و از نه تای باقی مانده به عنوان مجموعه آموزشی استفاده کنید.

ii. مدل را روی مجموعه آموزشی آموزش دهید.

iii. عملکرد مدل را در مجموعه validation با استفاده از یک معیار ارزیابی انتخابی ارزیابی کنید.

ج. میانگین عملکرد در ۱۰ برابر برای هر مقدار پارامتر را محاسبه کنید.

مقدار پارامتر را با بهترین عملکرد متوسط انتخاب کنید. این مقدار پارامتر است که بالاترین عملکرد را در 10 fold ارائه می دهد.

مدل را با استفاده از کل مجموعه داده و مقدار پارامتر انتخابی بازسازی کنید. با استفاده از مقدار پارامتر بهینه، مدل را بر روی مجموعه داده کامل آموزش دهید.

به صورت اختیاری، عملکرد مدل را بر روی یک مجموعه آزمایشی جداگانه یا مجموعه داده نگهدارنده که در طول مراحل آموزش یا cross validation استفاده نشده است، ارزیابی کنید. این یک ارزیابی اضافی از عملکرد مدل بر روی داده های دیده نشده ارائه می دهد.

با دنبال کردن این مراحل، می توانید با استفاده از 10 fold cross validation، پارامتر بهینه را پیدا کنید و مدل نهایی را بر اساس عملکرد آن انتخاب کنید.

برای تخمین خطای مدل خود می توانید مراحل زیر را دنبال کنید:

مجموعه داده خود را به دو بخش تقسیم کنید: یک مجموعه آموزشی و یک مجموعه تست. مجموعه آزمایشی باید بخش جداگانه‌ای از داده‌های شما باشد که در طول فرآیند اعتبارسنجی متقابل استفاده نشده است.

از کل مجموعه آموزشی برای آموزش یک مدل جدید با مقدار پارامتر انتخابی استفاده کنید.

مدل آموزش دیده را در مجموعه آزمایشی ارزیابی کنید: مدل آموزش دیده را روی مجموعه آزمایشی اعمال کنید و معیارهای عملکرد مورد علاقه را محاسبه کنید (به عنوان مثال، accuracy, precision, recall، و غیره). این معیارها به شما تخمینی از عملکرد مدل شما بر روی داده های نادیده می دهد.

به صورت اختیاری، معیارهای خطای اضافی را محاسبه کنید: بسته به مشکل خاص خود، ممکن است بخواهید معیارهای خطای اضافی مانند میانگین مربعات خطا (MSE) برای وظایف رگرسیونی یا ناحیه زیر منحنی ROC (AUC-ROC) برای وظایف طبقه‌بندی باینری را محاسبه کنید.

تخمین های خطا را تفسیر کنید: تخمین های خطا به دست آمده از مجموعه آزمایشی نشان می دهد که مدل شما چقدر به داده های جدید و نادیده تعمیم می یابد. به خاطر داشته باشید که تخمین خطا در مجموعه آزمایشی ممکن است کاملاً معرف عملکرد مدل در داده‌های کاملاً غیرقابل مشاهده نباشد، اما یک تقریب معقول ارائه می‌کند.

توجه به این نکته ضروری است که هنگام ارزیابی عملکرد مدل در مجموعه آزمایشی، باید از انجام تنظیمات بیشتر در مدل خود بر اساس نتایج مجموعه آزمایش خودداری کنید. این تضمین می‌کند که از تطبیق بیش از حد مجموعه تست اجتناب کنید و ارزیابی بی‌طرفانه‌ای از قابلیت‌های تعمیم مدل خود را حفظ کنید.

سوال ششم

در الگوریتم boosting اگر هر کدام از موارد زیر رخ دهد ما یادگیری را متوقف میکنیم؟ برای پاسخهای خود دلیل بیاورید.

• میزان خطای طبقه بندی‌کننده ترکیبی در داده های آموزشی اصلی ۰ شود.

خیر، زیرا خطای تست ممکن است حتی پس از صفر شدن خطای آموزش کاهش یابد. در این روش ها هدف تقویت، بهبود مکرر عملکرد گروه طبقهبندی کننده با تمرکز بر نمونه های طبقهبندی شده اشتباه است. الگوریتمهای تقویتکننده، مانند AdaBoost، با اختصاص وزنهای بالاتر به نمونه های طبقه بندی شده اشتباه در هر تکرار کار میکنند، در نتیجه به طبقه بندی کننده های ضعیف بعدی اجازه میدهند تا روی آن نمونه ها تمرکز بیشتری داشته باشند. تکرارها تا زمانی ادامه مییابند که تعداد از پیش تعریفشدهای از طبقه بندی کننده های ضعیف آموزش داده شوند یا به حداکثر تعداد تکرار برسد.

ممکن است میزان خطای طبقه بندی‌کننده ترکیبی در داده های آموزشی اصلی ۰ شود، اما به این معنی نیست که مدل یک نمایش کامل از توزیع زیربنایی را یاد گرفته است یا اینکه به خوبی به دادههای دیده نشده تعمیم میدهد. هدف اصلی تقویت، بهبود عملکرد تعمیم مدل، کاهش سوگیری و واریانس است. بنابراین، آموزش معمولاً تا زمانی ادامه مییابد که یک معیار توقف برآورده شود، مانند رسیدن به حداکثر تعداد تکرار، یک آستانه خطای از پیش تعریفشده، یا زمانی که عملکرد مدل در یک مجموعه اعتبارسنجی جداگانه شروع به بدتر شدن کند.

اگر میزان خطای طبقه‌بندی‌کننده ترکیبی در داده‌های آموزشی اصلی به صفر برسد، به این معنی است که طبقه‌بندی‌کننده ترکیبی به طبقه‌بندی کامل در مجموعه آموزشی دست یافته است. در چنین حالتی، boosting ممکن است متوقف شود زیرا بعید است تکرارهای بیشتر باعث بهبود عملکرد داده های آموزشی شود.

با این حال، توجه به این نکته مهم است که دستیابی به خطای صفر در داده‌های آموزشی تضمین نمی‌کند که مدل بر روی داده‌های دیده نشده یا مجموعه آزمایشی عملکرد کاملی داشته باشد. تطابق بیش از حد با داده های آموزشی ممکن است رخ دهد، جایی که مدل بیش از حد به مجموعه آموزشی تخصصی می شود و نمی تواند به خوبی به نمونه های جدید تعمیم یابد. بنابراین، ارزیابی عملکرد مدل تقویت شده در یک مجموعه آزمون مستقل برای ارزیابی توانایی تعمیم آن ضروری است.

• میزان خطای طبقه‌بندی‌کننده ضعیف فعلی روی داده‌های تمرین وزندار ۰ است.

بله، اگر میزان خطای طبقه‌بندی‌کننده ضعیف فعلی روی داده‌های تمرین وزن‌دار صفر باشد، به این معنی است که طبقه‌بندی‌کننده ضعیف قادر است نمونه‌ها را با وزن‌های مربوطه به‌طور کامل طبقه‌بندی کند. در این مورد، اگر الگوریتم از معیار توقف بر اساس عملکرد طبقه‌بندی ضعیف پیروی کند، boosting ممکن است متوقف شود.

یکی از معیارهای توقف متداول در الگوریتم‌های boosting این است که اگر میزان خطای طبقه‌بندی‌کننده ضعیف کمتر یا مساوی ۰ باشد، متوقف شود. تکرارهای بیشتر boosting ممکن است عملکرد کلی را به طور قابل توجهی بهبود نبخشد.

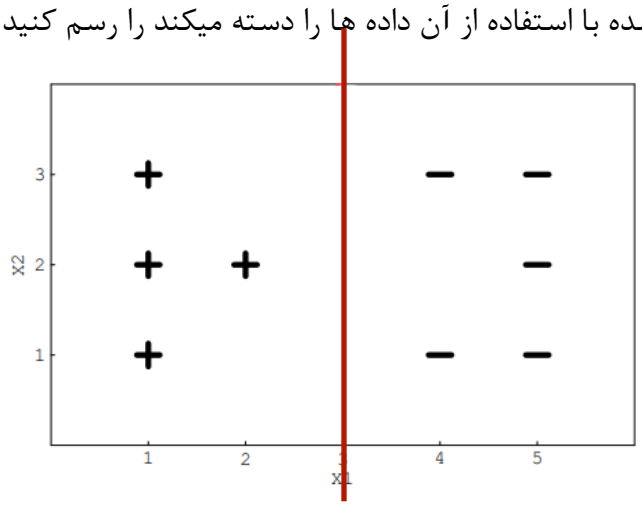
با این حال، توجه به این نکته مهم است که توقف در این مرحله ممکن است لزوما به بهترین عملکرد ممکن منجر نشود. هدف الگوریتم‌های تقویت، ترکیب چند طبقه‌بندی ضعیف برای ایجاد یک طبقه‌بندی قوی با تعمیم بهبود یافته است. توقف خیلی زود، حتی اگر طبقه‌بندی ضعیف خطای صفر در داده‌های وزنی داشته باشد، ممکن است پتانسیل تقویت را برای اصلاح بیشتر مدل محدود کند.

استفاده از معیارهای توقف اضافی یا پارامترهای کنترلی، مانند حداکثر تعداد تکرار یا آستانه بهبود عملکرد، معمول است تا اطمینان حاصل شود که تقویت تا رسیدن به سطح رضایت بخشی از تعمیم ادامه می یابد.

سوال هفتم

فرض کنید برای داده‌های زیر از طبقه‌بندی کننده SVM خطی بدون کرنل استفاده میکنیم و پارامتر C در این طبقه بندی کننده بسیار بزرگ در نظر گرفته شده است. (اگر در مورد این پارامتر اطلاعی ندارید این لینک را مطالعه کنید).

الف) خطی که SVM گفته شده با استفاده از آن داده ها را دسته میکند را رسم کنید و علت انتخاب این خط را توضیح دهید.



این خط به گونه ای انتخاب می شود که جمع حاشیه مرز یعنی فاصله تا نزدیک ترین داده ها ماکسیمم شود. همان طور که میبینیم دو داده منفی که $x = 4$ است و داده مثبت $x = 2$ با نزدیکترین داده به خط هستند و باتوجه به فرض گفته شده باید خط وسط این دو قرار گیرد.

ب) در شکل بالا نقاطی را انتخاب کنید که حذف آنها باعث میشود خطی که SVM داده ها را جدا میکند متفاوت از حالت (الف) شود. دلیل انتخاب این نقاط را توضیح دهید.

هر کدام از سه نقطه ذکر شده در قسمت قبل(داده های منفی در $x = 4$ و مثبت در $x = 2$) چون حاشیه مرز را تعیین میکنند اگر حذف شوند در تعیین خط موثر خواهد بود.

سوال هشتم

صحیح یا غلط بودن موارد زیر را با دلیل مشخص کنید :

الف) الگوریتم بیز ساده نمیتواند وابستگی بین متغیرها را مشخص کند. غلط

الگوریتم ساده بیز، بنا به ماهیت خود، استقلال بین متغیرهای ورودی (ویژگی ها) را فرض می کند. این فرض به عنوان "ساده" شناخته می شود زیرا فرآیند مدل سازی را با فرض اینکه همه ویژگی ها با توجه به متغیر کلاس مستقل از یکدیگر هستند، ساده می کند. این فرض محاسبه احتمالات شرطی مورد استفاده در الگوریتم را بسیار ساده می کند.

با توجه به این فرض، الگوریتم ساده بیز ممکن است نتواند وابستگی ها یا تعاملات پیچیده بین متغیرها را به دقت ثبت کند. فرض بر این است که وجود یا عدم وجود یک ویژگی خاص با وجود یا عدم وجود هر ویژگی دیگر، با توجه به متغیر کلاس، ارتباطی ندارد.

در حالی که بیز ساده ممکن است نتواند به طور مستقیم وابستگی بین متغیرها را تعیین کند، هنوز هم می تواند در موقعیت های خاص موثر باشد، به ویژه زمانی که فرض استقلال به خوبی برقرار است یا زمانی که متغیر کلاس بسیار آموزنده است.

اگر دانش قبلی یا شواهد قوی از وابستگی بین متغیرها در مجموعه داده خود دارید، سایر الگوریتم‌های یادگیری ماشینی که استقلال را فرض نمی‌کنند، مانند درخت‌های تصمیم، جنگل‌های تصادفی یا شبکه‌های عصبی، ممکن است برای گرفتن این وابستگی‌ها مناسب‌تر باشند.

ب) هنگامی که یک درخت تصمیم به سمت یک درخت پر پیش میرود احتمال اینکه نویز را هم پوشش دهد بیشتر میشود. درست

هنگامی که یک درخت تصمیم به سمت یک درخت کامل حرکت می‌کند، به این معنی است که با ایجاد شاخه‌ها و گره‌های بیشتر برای پوشش داده‌های آموزشی در حال گسترش و رشد عمیق‌تر است. این می تواند احتمال بیش از حد برازش را افزایش دهد، به این معنی که مدل بیش از حد پیچیده می شود و شروع به به خاطر سپردن نویز یا الگوهای نامربوط در داده های آموزشی می کند.

همانطور که درخت تصمیم پیچیده تر می شود، تمایل بیشتری برای تطبیق بیشتر داده های آموزشی از جمله نویز یا نقاط پرت موجود در داده ها دارد. این می‌تواند منجر به کاهش عملکرد تعمیم در داده‌های دیده نشده شود، زیرا مدل بسیار خاص به مجموعه آموزشی می‌شود و نمی‌تواند الگوهای واقعی زیربنایی را ثبت کند.

در مقابل، یک درخت تصمیم با عمق محدود یا ساختار ساده‌تر ممکن است یک مدل کلی‌تر و قوی‌تر ارائه دهد. بر روی ثبت ویژگی‌ها و الگوهای مهم در داده‌ها بدون تناسب بیش از حد با نویز یا جزئیات نامربوط تمرکز می‌کند.

برای کاهش خطر بیش از حد برازش و درج نویز، می توان از تکنیک هایی مانند هرس، تعیین حداکثر عمق برای درخت یا استفاده از روش های منظم سازی استفاده کرد. این تکنیک ها به جلوگیری از رشد بیش از حد درخت تصمیم کمک می کند و تعادل بهتری بین ثبت الگوهای مرتبط و اجتناب از نویز ایجاد می کند.

ج) در روش k نزدیک ترین همسایه اگر $k=1$ الگوریتم نسبت به داده های نویز مقاوم تر از حالتی است که $k=5$ در نظر گرفته شود. غلط

زیرا اگر $k=1$ باشد با پیدا کردن نزدیک ترین داده به هر داده جدید برچسب زنی را انجام میدهیم و اگر نویز در نزدیکی داده باشد برچسب زنی اشتباه میشود اما اگر k بیشتر باشد قدرت نویز در برچسب زنی کمتر می شود چون بقیه داده ها هم برچسب زنی را تحت تاثیر قرار می دهند.

سوال نهم

فرض کنید در حال طراحی یک سیستم برای تشخیص خستگی راننده در اتومبیل هستید. بسیار مهم است که مدل شما خستگی را تشخیص دهد تا از هر گونه حادثه ای جلوگیری شود. کدام یک از معیارهای زیر بهترین معیار برای ارزیابی هست : Accuracy Precision, Recall, Loss Value دلیل انتخاب خود را شرح دهید.

در زمینه طراحی سیستمی برای تشخیص خستگی راننده در خودرو، مهم ترین معیار برای ارزیابی، Recall خواهد بود.

یادآوری که به عنوان حساسیت یا نرخ مثبت واقعی نیز شناخته می‌شود، توانایی مدل را در شناسایی صحیح موارد خستگی راننده در بین تمام موارد واقعی خستگی اندازه‌گیری می‌کند. به عبارت دیگر، درصد درایورهای واقعی خسته که به درستی توسط سیستم شناسایی شده اند را کمیت می کند.

دلیل اینکه Recall مهمترین معیار در این سناریو است، اولویت اجتناب از تصادف است. هدف این سیستم با بهینه‌سازی برای یادآوری بالا، به حداقل رساندن منفی‌های کاذب است، که در مواردی است که راننده خسته به عنوان خسته تشخیص داده نمی‌شود. وجود برخی از نکات مثبت کاذب (رانندگان غیرخسته به اشتباه به عنوان خسته علامت گذاری شده اند) قابل قبول تر است تا عدم شناسایی راننده خسته، زیرا این امر به طور بالقوه می تواند منجر به تصادف یا سایر موقعیت های خطرناک شود.

در حالی که دقت، دقت، و ارزش از دست دادن نیز معیارهای ارزیابی مهم هستند، ممکن است نگرانی اصلی در این مورد خاص نباشند. اگر داده‌ها نامتعادل باشند، ممکن است دقت به تنهایی ارزیابی دقیقی ارائه نکند، و Loss Value ممکن است اهداف و خطرات خاص مرتبط با تشخیص خستگی راننده را نشان ندهد.

دقت، که نسبت رانندگان خسته به درستی شناسایی شده را در بین تمام رانندگان خسته شناسایی شده اندازه گیری می کند، نیز مهم است اما به اندازه یادآوری در این سناریو حیاتی نیست. مثبت کاذب ممکن است با علامت گذاری اشتباه رانندگان غیرخسته به عنوان خسته، ناراحتی ایجاد کند، اما در مقایسه با عواقب احتمالی از دست دادن یک راننده واقعاً خسته، اولویت کمتری دارد.

به طور خلاصه، معیار ارزیابی اولیه باید Recall باشد، زیرا بر به حداقل رساندن منفی کاذب و به حداکثر رساندن تشخیص موارد واقعی خستگی راننده تمرکز دارد که با هدف اجتناب از تصادفات و تضمین ایمنی راننده هماهنگ است.

سوال دهم

علاوه بر شاخص آنتروپی برای ساخت درخت تصمیم، شاخص دیگری نیز وجود دارد که میتوان به جای آنتروپی از آن برای ساخت درخت استفاده کرد. این شاخص را معرفی کنید و بگویید تفاوت آن با آنتروپی چیست؟ بالا یا پایین بودن این شاخص چه معنایی دارد و چگونه محاسبه میشود.

یکی دیگر از معیارهایی که معمولا برای ساخت درخت های تصمیم استفاده می شود، Gini impurity است که به عنوان شاخص جینی یا ضریب جینی نیز شناخته می شود. Gini impurity میزان ناخالصی یا بی نظمی یک گره از درخت تصمیم را اندازه گیری می کند.

برخلاف آنتروپی، که مقدار متوسط اطلاعات یا عدم قطعیت را در یک گره تعیین می‌کند، Gini بر احتمال طبقه‌بندی اشتباه یک عنصر به‌طور تصادفی انتخاب شده در یک گره تمرکز می‌کند. این احتمال را محاسبه می کند که یک عنصر به طور تصادفی انتخاب شده در یک گره، اگر به طور تصادفی بر اساس توزیع کلاس ها در آن گره برچسب گذاری شود، به اشتباه برچسب گذاری شود.

Gini با جمع کردن مجذور احتمالات هر برچسب کلاسی که انتخاب می شود ضرب در (۱ - احتمال آن برچسب کلاس) محاسبه می شود. فرمول ناخالصی جینی به شرح زیر است:

$$Gini = 1 - (p1^2 + p2^2 + ... + pk^2)$$

که در آن p1، p2، ...، pk احتمال هر برچسب کلاس را در گره نشان می دهد.

اگر شاخص Gini بالا باشد، نشان دهنده سطح بالاتری از ناخالصی یا اختلال در گره است. این بدان معنی است که کلاس های گره بیشتر مخلوط شده و کمتر به خوبی از هم جدا شده اند. Gini بالا نشان می دهد که گره نیاز به تقسیم بیشتر برای بهبود جداسازی کلاس ها دارد.

برعکس، اگر شاخص Gini پایین باشد، نشان دهنده سطح کمتر ناخالصی یا بی نظمی در گره است. این بدان معنی است که کلاس های گره همگن تر و به خوبی از هم جدا شده اند. ناخالصی جینی کم نشان می دهد که گره در حال حاضر نسبتاً خالص است و ممکن است نیازی به تقسیم بیشتر نداشته باشد.

به طور خلاصه، Gini یک معیار جایگزین برای آنتروپی برای ساخت درختان تصمیم است. ناخالصی یا بی نظمی در یک گره را بر اساس احتمال طبقه بندی اشتباه کمی می کند. Gini بالا نشان دهنده گره مخلوط تر است، در حالی که ناخالصی جینی کم نشان دهنده گره همگن تر است. الگوریتم‌های درخت تصمیم، مانند CART (درخت طبقه‌بندی و رگرسیون)، بسته به پیاده‌سازی و الزامات خاص، می‌توانند از Gini یا آنتروپی به عنوان معیار تقسیم استفاده کنند.

سوال یازدهم

درمورد مسائل رگرسیون به سوالات زیر پاسخ دهید :

الف) simple linear regression و multiple linear regression با یکدیگر مقایسه کرده و تفاوت و شباهت های آنها را بیان کنید.

رگرسیون خطی ساده و رگرسیون خطی چندگانه هر دو تکنیک‌های آماری هستند که برای مدل‌سازی رابطه بین یک متغیر وابسته و یک یا چند متغیر مستقل استفاده می‌شوند. با این حال، آنها از نظر تعداد متغیرهای مستقل درگیر و پیچیدگی رابطه ای که می توانند دریافت کنند، متفاوت هستند.

رگرسیون خطی ساده:

رگرسیون خطی ساده فقط شامل یک متغیر مستقل و یک متغیر وابسته است.

این یک رابطه خطی بین متغیر مستقل و متغیر وابسته را فرض می کند که می تواند با یک خط مستقیم در نمودار پراکندگی نشان داده شود.

هدف از رگرسیون خطی ساده، یافتن بهترین خطی است که مجموع اختلاف مجذور بین نقاط داده مشاهده شده و مقادیر پیش بینی شده روی خط را به حداقل برساند.

رگرسیون خطی ساده تخمین‌هایی را برای شیب و قطع خط ارائه می‌کند که به ترتیب نشان‌دهنده رابطه و نقطه شروع خط است.

معمولاً زمانی استفاده می شود که یک متغیر پیش بینی کننده وجود داشته باشد که تصور می شود تأثیر خطی بر متغیر پاسخ دارد.

رگرسیون خطی چندگانه:

رگرسیون خطی چندگانه شامل دو یا چند متغیر مستقل و یک متغیر وابسته است.

این اجازه می دهد تا رابطه پیچیده تری بین متغیرهای مستقل و متغیر وابسته وجود داشته باشد، زیرا اثرات ترکیبی پیش بینی کننده های متعدد را در نظر می گیرد.

هدف از رگرسیون خطی چندگانه یافتن ابرصفحه با بهترین تناسب در یک فضای چند بعدی است که مجموع اختلاف مجذور بین نقاط داده مشاهده‌شده و مقادیر پیش‌بینی‌شده روی ابر صفحه را به حداقل برساند.

رگرسیون خطی چندگانه تخمین‌هایی را برای ضرایب متغیرهای مستقل ارائه می‌کند که نشان‌دهنده بزرگی و جهت تأثیر آنها بر متغیر وابسته است.

معمولاً زمانی استفاده می‌شود که متغیرهای پیش‌بینی‌کننده متعددی وجود داشته باشند که اعتقاد بر این است که تأثیر خطی روی متغیر پاسخ دارند، و زمانی که می‌خواهیم اثرات فردی و ترکیبی این پیش‌بینی‌کننده‌ها را درک کنیم.

شباهت ها:

هر دو رگرسیون خطی ساده و چندگانه بر اساس فرض یک رابطه خطی بین متغیرهای مستقل و متغیر وابسته است.

هدف هر دوی آنها یافتن بهترین خط یا ابر صفحه است که تفاوت بین داده های مشاهده شده و مقادیر پیش بینی شده را به حداقل می رساند.

هر دو تکنیک شامل تخمین ضرایب برای تعیین کمیت رابطه بین متغیرهای مستقل و متغیر وابسته است.

تفاوت:

تفاوت اصلی در تعداد متغیرهای مستقل درگیر نهفته است. رگرسیون خطی ساده با یک متغیر مستقل سروکار دارد، در حالی که رگرسیون خطی چندگانه با دو یا چند متغیر مستقل سروکار دارد.

رگرسیون خطی چندگانه امکان مدل‌سازی انعطاف‌پذیرتر و پیچیده‌تر رابطه را با در نظر گرفتن اثرات ترکیبی پیش‌بینی‌کننده‌های متعدد فراهم می‌کند.

تفسیر ضرایب برآورد شده بین دو تکنیک متفاوت است. در رگرسیون خطی ساده، ضریب نشان دهنده تغییر در متغیر وابسته مرتبط با تغییر یک واحدی در متغیر مستقل است. در رگرسیون خطی چندگانه، ضرایب نشان دهنده تغییر در متغیر وابسته مرتبط با تغییر یک واحدی در متغیر مستقل مربوطه است، در حالی که سایر پیش بینی ها را ثابت نگه می دارند.

به طور کلی، رگرسیون خطی چندگانه قابلیت‌های رگرسیون خطی ساده را با اجازه دادن به گنجاندن پیش‌بینی‌کننده‌های متعدد گسترش می‌دهد و امکان تحلیل جامع‌تری از روابط بین متغیرها را فراهم می‌کند.

ب) یکی از راه های جلوگیری از بیش برازش استفاده از منظمسازی است که به دو نوع L1 و L2 تقسیم میشود. به نوع اول Lasso Regression و به نوع دوم Ridge regression گفته میشود. تفاوت این دو روش را از نوع بهینه سازی بیان کرده و نحوه کار آنها را توضیح دهید.

تفاوت:

اصطلاح پنالتی: Lasso از پنالتی هنجار L۱ استفاده می کند، در حالی که Ridge از پنالتی نرمال L۲ استفاده می کند.

پراکندگی: Lasso می تواند برخی از ضرایب را دقیقاً به صفر برساند و در نتیجه مدل های پراکنده ایجاد شود. Ridge فقط ضرایب را به سمت صفر کوچک می کند بدون اینکه تناقض را اعمال کند.

انتخاب ویژگی: Lasso با حذف ویژگی های نامربوط، انتخاب خودکار ویژگی را انجام می دهد. Ridge تمام ویژگی ها را حفظ می کند اما تاثیر آنها را کاهش می دهد.

تفسیرپذیری: Lasso می‌تواند با شناسایی و حذف ویژگی‌های نامربوط، مدل قابل تفسیرتری ارائه دهد. Ridge تمام ویژگی‌ها را حفظ می‌کند، که می‌تواند تفسیر را چالش‌برانگیزتر کند.

به طور خلاصه، رگرسیون Lasso (قانونی سازی L1) و رگرسیون Ridge (قانونی سازی L2) تکنیک های منظم سازی هستند که هدف آنها جلوگیری از برازش بیش از حد در مدل های رگرسیون خطی است. Lasso با هدایت برخی ضرایب دقیقاً به صفر، پراکندگی و انتخاب ویژگی را ارتقا می دهد، در حالی که Ridge همه ویژگی ها را حفظ می کند اما تأثیر آنها را کاهش می دهد. انتخاب بین Lasso و Ridge به مشکل خاص، اهمیت انتخاب ویژگی و قابلیت تفسیر مطلوب مدل بستگی دارد.

ج) در جدول زیر سن و فشار خون چند بیمار قلبی داده شده است. معادله رگرسیون به فرم $y = \beta_0 + \beta_1 x$ به دست آورید. همچنین با استفاده از معادله به دست آمده فشار خون یک بیمار ۴۰ ساله را پیش بینی کنید.) متغیر x نشان دهنده سن و متغیر y نشان دهنده فشار خون است)

Patient	A	B	C	D	E	F	G
x	42	74	48	35	56	26	60
y	98	130	120	88	182	80	135

$$X = \begin{bmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{bmatrix}, Y = \begin{bmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T Y = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 74 & 48 & 35 & 56 & 26 & 60 \end{bmatrix} \begin{bmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 74 & 48 & 35 & 56 & 26 & 60 \end{bmatrix} \begin{bmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T Y = \left(\begin{bmatrix} 7 & 341 \\ 341 & 18181 \end{bmatrix} \right)^{-1} \begin{bmatrix} 833 \\ 42948 \end{bmatrix} = \begin{bmatrix} 45.46741 \\ 1.50947 \end{bmatrix}$$

$$y = 45.46741 + 1.50947x$$

$$x = 40 \rightarrow y = 45.46741 + 1.50947x = 105.84621$$