تئوری تکلیف چهارم درس الگوریتم های علوم داده

نام و نام خانوادگی: فاطمه ترودی

شماره دانشجویی: 403422048

نام استاد: دکتر سعیدرضا خردپیشه

نیمسال دوم 04-1403

Q1

حسگر / ورودی : داده های خام را از محیط جمع آوری میکند .

استخراج ویژگی : داده های خام را به اطلاعات مفید و فشرده تبدیل می کند .

طبقه بند / الگوریتم یادگیری : با استفاده از ویژگی ها، الگوها را به دسته های مشخصی طبقه بندی می کند .

آموزش و منطق تصمیم گیری : در فاز آموزش، مدل را با داده های برچسب دار تنظیم کرده و در فاز طبقه بندی، برای داده های جدید پیش بینی انجام می دهد .

معلم مجموعه ای از ورودی ها همراه با لیبل کلاس صحیح فراهم میکند که به اصطلاح به آن مجموعه داده آموزشی گفته میشود.

الگوریتم یادگیری از این داده ها برای تنظیم پارامترهای مدل استفاده میکند تا خطا را مجموعه ی train کم کند .

معلم خود الگوریتم را تغییر نمی دهد، بلکه فقط نقش راهنما را با فراهم کردن پاسخ های صحیح ایفا می کند.

"جنبه"	"حالت آموزش"	"حالت طبقه بندی"
"هدف"	"یادگیری از داده های برچسب دار"	"پیش بینی برچسب برای داده های نادیده"
"ورودی"	"بردار های ویژگی + برچسب های کلاس شناخته شده"	"فقط بردار های ویژگی (بدون برچسب)"
"خروجی"	"مدل آموزش دیده (مانند مرزهای تصمیم گیری)"	"برچسب های کلاس پیش بینی شده"
"فرآیند"	"پارامترهای مدل بهینه سازی میشوند ","مدل ثابت است و فقط برای استنتاج استفاده میشود"	
"نظارت"	"نیاز به نظارت انسانی (داده های برچسب دار)"	"به صورت خودکار عمل می کند"

آموزش :مجموعه ای از تصاویر گربه/سگ با برچسب ها برای آموزش طبقه بند استفاده می شود .

طبقه بندی :از طبقه بند آموزش دیده برای تشخیص اینکه تصویر جدید گربه است یا سگ استفاده می شود.

Q2

تفاوت اصلی بین طبقه بندی و رگرسیون

طبقه بندی :

هدف: پیش بینی مقادیر گسسته یا دسته ای

برای مثال : تشخیص اینکه یک ایمیل اسپم است یا خیر (دسته ها: اسپم، غیراسپم).

رگرسیون :

هدف : پیش بینی مقادیر عددی پیوسته

برای مثال : پیش بینی قیمت یک خانه براساس متراژ و مکان (خروجی : یک عدد)

Q3

بردار ویژگی / تعریف و نقش: بردار ویژگی یک بردار-**n** بعدی از مقادیر عددی است که ویژگی ها یا خصوصیات اساسی یک شیء یا سیگنال ورودی را نشان می دهد.

نقش در تشخیص الگو: بردار ویژگی نتیجه فرآیند استخراج ویژگی است، جایی که داده های خام ورودی (مانند تصویر، صدا یا متن) به یک فرمت ساختاریافته تبدیل می شوند تا طبقه بند بتواند با آن کار کند. بردار ویژگی جنبه های اطلاعاتی و تمایزدهنده ورودی را ثبت می کند تا به تمایز بین کلاس های مختلف کمک کند.

حالت مخفی / تعریف و رابطه با طبقه بندی : حالت مخفی یک متغیر نهان (latent) در مدل است که به طور مستقیم از داده های ورودی قابل مشاهده نیست، اما بر ویژگی های مشاهده شده تأثیر می گذارد. در طبقه بندی، هدف اغلب استنباط محتمل ترین حالت مخفی با توجه به دنباله ای از بردارهای ویژگی مشاهده شده است. این حالت های مخفی معمولاً با برچسب های کلاس یا زیرکلاس ها مرتبط هستند.

Q4

Template Matching یک تکنیک تشخیص الگو است که در آن یک تصویر ورودی با مجموعه ای از الگوهای ذخیره شده مقایسه می شود. هر الگو نماینده یک کلاس (مانند حروف یا اعداد) است.

تصویر سمت چپ یک شبکه دوبعدی باینری از کاراکتر "A" را نشان می دهد .

این شبکه به یک بردار ویژگی یک بعدی تبدیل می شود .

در تطبیق الگو، این بردار با تمام بردارهای الگوی ذخیره شده با استفاده از یک معیار تشابه (مانند فاصله اقلیدسی یا همبستگی) مقایسه می شود .

الگویی که بیشترین شباهت را داشته باشد (یعنی نزدیک ترین تطابق)، خروجی طبقه بندی شده را تعیین می کند.

نکته این است که هیچ یادگیری ای در این روش انجام نمی شود؛ طبقه بندی صرفاً بر اساس مقایسه مستقیم با نمونه های موجود صورت می گیرد.

تطبیق الگو با تعداد محدود الگوها به دلایل زیر عملکرد ضعیفی دارد:

1. **تطبیق الگو صرفاً نمایش های دقیق را ذخیره می کند . و اگر الگوی ورودی حتی اندکی تغییر کند به دلیل نویز ممکن است با هیچ کدام از الگوهای ذخیره شده مطابقت نداشته باشد و منجر به طبقه بندی نادرست شود.**

2. **داده های واقعی اغلب شامل تغییرات هستند (مثل تفاوت در سبک نوشتاری یا زاویه تصویر). بدون داشتن الگوهای کافی برای پوشش این تغییرات، عملکرد سیستم کاهش می یابد.**

Q5

منطقه تصمیم گیری: منطقه تصمیم گیری بخشی از فضای ویژگی است که در آن همه الگوهای ورودی توسط طبقه بند به یک برچسب کلاس یکسان تخصیص داده می شوند.

مرز تصمیم گیری: مرز تصمیم گیری سطحی (یا خط در فضای دوبعدی) است که مناطق تصمیم گیری مختلف را از هم جدا می کند. این مرز نقطه ای است که طبقه بند از پیش بینی یک کلاس به کلاس دیگر تغییر می کند.

عدم قطعیت در طبقه بندی منجر به موارد زیر می شود:

1. **در مناطق با عدم قطعیت، مشخص نیست که یک ورودی به کدام کلاس تعلق دارد و مرزها به جای اینکه واضح و دقیق باشند، مبهم یا احتمالی می شوند.**

2. در این مناطق، احتمال خطای طبقه بندی افزایش می یابد و اعتماد **ما به** طبقه بند کاهش پیدا می کند.

3. **تعداد مناطق نامطمئن زیادی معمولا در نزدیکی مرزهای تصمیم گیری یافت میشوند.**

Q6

قضیه بیز در زمینه طبقه بندی : در طبقه بندی ، قضیه بیز برای محاسبه احتمال تعلق یک مشاهده به یک کلاس خاص استفاده میشود .

هدف: تخصیص مشاهده x به محتمل ترین کلاس

از آنجا که p(x) برای همه ی کلاس ها یکسان است ، میتوان آن را در فرایند ماکزیمم کردن نادیده گرفت.

کلاسی انتخاب میشود که حاصل ضرب این احتمالات را بیشینه کند .

: این احتمال نشان دهنده ی فراوانی یا احتمال پایه ای یک کلاس خاص در مجموعه ی داده هاست

: این احتمال نشان میدهد که چقدر احتمال دارد بردار ویژگی x را مشاهده کنیم ، با فرض اینکه متعلق به کلاس باشد.

: این احتمال به روز شده کلاس مذکور بعد از مشاهده ی بردار ویژگی x است . این مقدار نهایی برای تصمیم گیری در طبقه بندی استفاده میشود و نشان دهنده ی احتمال تعلق داده به یک کلاس خاص است .

: این اصطلاح احتمال کلی مشاهده بردار ویژگی X در تمام کلاس ها را نشان می دهد. به عنوان یک عامل نرمال سازی عمل می کند تا اطمینان حاصل شود که مجموع احتمالات پسین برای همه کلاس ها برابر با ۱ است

Q7

در طبقه بندی بیزی، از قضیه بیز برای محاسبه احتمال پسین استفاده می کنیم که فرمولش در سوال قبل ذکر شده .

در این فرمول تابع چگالی احتمال شرطی کلاس PDF است که نشان دهنده احتمال مشاهده بردار ویژگی x با فرض تعلق به کلاس است.

برای اعمال قانون تصمیم گیری بیز، باید این *PDF* برای هر کلاس از داده های آموزشی تخمین زده شود.

اهمیت:

دقت طبقه بندی به شدت به کیفیت تخمین وابسته است.

تخمین ضعیف PDF منجر به محاسبه نادرست احتمالات پسین می شود که نتیجه آن خطاهای طبقه بندی است.

تخمین پارامتری :

- **فرض :فرض می کند که PDF شکل ثابتی دارد (مثلاً توزیع گاوسی).**
- **پیچیدگی مدل :تعداد پارامترها ثابت است (مثلاً میانگین و ماتریس کوواریانس).**
- **نیاز به داده :با مجموعه داده های کوچک تا متوسط به خوبی کار می کند.**
- **انعطاف پذیری :انعطاف پذیری کمتری دارد و ممکن است شکل های پیچیده توزیع را از دست بدهد.**
- **مثال :فرض کنیم داده های هر کلاس از توزیع گاوسی پیروی می کنند. در این حالت، میانگین و ماتریس کوواریانس از داده های آموزشی تخمین زده می شود.**

تخمین غیرپارامتری :

- **فرض :هیچ فرض قوی ای درباره شکل PDF ندارد و توزیع را مستقیماً از داده ها مدل می کند.**
- **پیچیدگی مدل :با افزایش داده ها پیچیده تر می شود.**
- **نیاز به داده :به مقدار زیادی داده نیاز دارد تا تخمین دقیقی ارائه دهد.**
- **انعطاف پذیری :انعطاف پذیرتر است و می تواند توزیع های دلخواه و پیچیده را مدل کند.**
- **مثال :استفاده از (Kernel Density Estimation - KDE)برای ساخت PDF مستقیماً از داده ها، با استفاده از یک هسته (مثلاً هسته گاوسی) که روی هر نقطه داده متمرکز است.**

Q8

تصویر سمت چپ :

- **مرزهای تصمیم گیری بسیار پیچیده و نامنظم هستند،**
- **و طبقه بند تلاش کرده است هر نقطه داده (قرمز یا سیاه) را به طور کامل از کلاس دیگر جدا کند، و این کار به ایجاد مناطق تصمیم گیری بسیار تکه تکه و غیرقابل تعمیم منجر شده است.**
- **این نشان دهنده بیش برازش *(overfitting)*است،**

تصویر سمت راست :

- **مرزهای تصمیم گیری نرم تر و کمتر پیچیده هستند و مناطق تصمیم گیری منسجم تری را تشکیل می دهند.**
- **برخی نقاط به اشتباه طبقه بندی شده اند ، اما شکل کلی مناطق تصمیم گیری تعمیم پذیرتر است.**
- **این نشان دهنده استفاده از یک مدل ساده تر یا با تنظیمات (regularization) است که الگوهای کلی تر را در داده ها تشخیص داده و احتمالاً عملکرد بهتری روی داده های جدید خواهد داشت.**

وقتی ویژگی های کلاس ها هم پوشانی دارند، تعریف مرزهای تصمیم گیری واضح و دقیق دشوارتر می شود. در این حالت:

1. **در مناطقی که ویژگی های کلاس ها هم پوشانی دارند، نقاط داده از کلاس های مختلف در هم آمیخته می شوند. در نتیجه، دیگر نمی توان یک خط یا سطح کاملاً جداکننده برای تفکیک دقیق آنها ترسیم کرد.**
2. **وقتی کلاس ها هم پوشانی دارند، دستیابی به دقت بالا ذاتاً دشوارتر است.**
3. **طبقه بند باید یک مرز سازشی پیدا کند که خطاهای طبقه بندی را در منطقه هم پوشانی کمینه کند.**
   این ممکن است منجر به مرزهای نرم تر شود که روند کلی داده ها را دنبال می کنند، به جای اینکه هر نقطه را به طور دقیق جدا کنند.

Q9

قانون 1-Nearest Neighbor (1-NN)

یک نمونه جدید بدون برچسب را با پیدا کردن یک نمونه ی نزدیک ترین از مجموعه داده های آموزش، بر اساس معیار شباهت (یا معادل آن معیار فاصله) دسته بندی می کند. مراحل کار به این صورت است:

فرض کنید یک نمونه ورودی جدید *x* داریم، شباهت (یا فاصله) بین *x*و هر نمونه موجود در داده های آموزش را حساب میکنیم.

نمونه آموزشی ای که بیشترین شباهت (یا کمترین فاصله) با *x* دارد، به عنوان نزدیک ترین همسایه انتخاب می شود.

سپس برچسب کلاس این نمونه نزدیک ترین همسایه را به *x* نسبت می دهیم.

به جای شباهت میتوانیم از معیار های دیگر هم استفاده کنیم مثل نرم اقلیدسی یا غیره.

این طبقه بند مستقل از پیاده سازی است زیرا:

نیازی به مرحله آموزش یا ساخت مدل ندارد. فقط کافی است داده های آموزش ذخیره شده و در زمان دسته بندی، فاصله یا شباهت محاسبه شود.

هیچ فرض یا پارامتری درباره توزیع داده یا ساختار مدل از پیش تعیین نمی شود.

انتخاب معیار فاصله (یا شباهت) تاثیر قابل توجهی روی دقت ایین طبقه بند دارد زیرا:

معیار تعیین می کند که «نزدیک ترین» را چه معنا کنیم ؛ انتخاب نامناسب ممکن است همسایه هایی را انتخاب کند که واقعا مشابه نیستند.

اگر معیار به خوبی ساختار داده را نشان دهد، همسایه ها احتمالاً برچسب مشابهی خواهند داشت و دقت افزایش می یابد. و برعکس اگر معیار مناسب نباشد (مثلاً استفاده از فاصله اقلیدسی روی داده هایی که مقیاس های مختلف دارند یا روابط غیرخطی دارند)، همسایه های نزدیک می توانند گمراه کننده باشند و دقت کاهش می یابد.

معیارهای مختلف بر جنبه های متفاوت داده تأکید دارند: مثلا فاصله اقلیدسی حساس به مقیاس و شکل هندسی است.

و یا فاصله منهتن تفاوت های مطلق را مد نظر دارد.

Q10

روش پنجره پارزن یک روش غیرپارامتریک برای تخمین چگالی احتمال است که به این شکل عمل می کند:

فضای نمونه ها را به قسمت های کوچک (پنجره ها یا باکس ها) تقسیم می کند که اندازه هر پنجره با پارامتر h مشخص می شود.

برای یک نقطه جدید x ، تعداد نمونه های آموزش که داخل پنجره ای به اندازه h و x قرار دارند را شمارش می کند.

چگالی احتمال در x را به صورت نسبت تعداد نمونه های داخل پنجره به حجم پنجره تخمین می زند.

به عبارتی، هر پنجره یک تابع وزنی دارد و تعداد نقاط داخل آن تعیین کننده چگالی در آن نقطه است.

روش *K-NN* هم یک روش غیر پارامتریک است اما تفاوت اصلی آن با پارزن در این است که:

به جای ثابت نگه داشتن اندازه پنجره h، پنجره در روش *k-NN* متغیر است و طوری انتخاب می شود که دقیقاً k نمونه آموزش داخل آن باشد.به عبارت دیگر، برای نقطه x، فاصله بهk امین نزدیک ترین نمونه آموزش به x به عنوان اندازه پنجره تعیین می شود.

بنابراین، حجم پنجره متغیر است و بسته به چگالی داده ها در اطراف x تغییر می کند.

اگر h یا k خیلی کوچک باشد: پنجره ها خیلی کوچک یا تعداد همسایه ها کم است و تخمین چگالی نویزی و پراکنده می شود یا همان *overfitting* رخ میدهد .

اگر h یا k خیلی بزرگ باشد: پنجره ها یا تعداد همسایه ها زیاد می شود و تخمین چگالی خیلی صاف و مبالغه آمیز *(underfitting)* خواهد بود و جزئیات داده از بین می رود.

ا نتخاب بهینه این پارامترها بستگی به داده ها، توزیع آنها و هدف نهایی دارد. و معمولاً نیاز به آزمایش و ارزیابی عملکرد مدل با مقادیر مختلف دارد.

Q11

تابع Discriminant Functions

یک تابع ریاضی است که برای هر کلاس یک مقدار عددی به ورودی می دهد و هدف آن تصمیم گیری درباره تعلق نمونه به کدام کلاس است .در واقع، نمونه به کلاسی اختصاص داده می شود که مقدار این تابع در آن کلاس بیشترین باشد.

فرض کنیدکلاس داریم که هر کلاس با توزیع گاوسی چندمتغیره با میانگین و ماتریس کوواریانس مدل سازی شده است. تابع تفکیک برای کلاس به صورت زیر است:

: چگالی احتمال شرطی برای کلاس است

: احتمال پیشین کلاس است

برای توزیع گاوسی چندمتغیره:

با گرفتن لگاریتم و حذف کردن ثابت هایی که وابسته به کلاس نیستند، تابع تفکیک می شود:

مرز تصمیم بین دو کلاس و شامل نقاطی از فضای ویژگی ها است که تابع تفکیک برای هر دو کلاس برابر است:

بنابراین، مرز تصمیم همان مجموعه نقاطی است که در آن سیستم بین دو کلاس مردد است.

- **اگر ماتریس های کواریانس و متفاوت باشند، مرز تصمیم معمولاً یک منحنی درجه دو است. این حالت به عنوان تحلیل تفکیک مربعی - Quadratic Discriminant Analysis) (QDAشناخته می شود.**
- **اگر ماتریس کوواریانس همه کلاس ها یکسان باشد، مرز تصمیم به صورت خطی خواهد بود که به آن تحلیل تفکیک خطی (Linear Discriminant Analysis - LDA) می گویند.**

در نتیجه، توابع تفکیک شکل و محل مرزهای تصمیم در فضای ویژگی را تعیین می کنند.

Q12

الگوریتم (Expectation-Maximization - EM)یک روش تکراری است که برای تخمین پارامترها در مدل هایی با متغیرهای نهان مثل مدل مخلوط گاوسی Gaussian Mixture) (Models - GMMبه کار می رود. در مدل مخلوط گاوسی فرض می شود داده ها از ترکیب چند توزیع گاوسی تولید شده اند، اما عضویت هر داده به یکی از این مؤلفه ها (مخفی است).

مراحل:

1.پارامترها را مقداردهی اولیه می کنیم: میانگین ها ، ماتریس های کوواریانس و ضرایب ترکیب برای هر مؤلفه گاوسی

2. مسئولیت را محاسبه می کنیم؛ یعنی احتمال اینکه نمونه داده متعلق به مؤلفه باشد، با فرض پارامترهای فعلی:

3.پارامتر ها را به روز رسانی میکنیم:

4.مراحل 2 و3 را تکرار میکنیم تا پارامتر ها به پایداری برسند.

روش های دسته ای (Ensemble) ترکیبی از پیش بینی های چندین طبقه بند هستند که تصمیم نهایی را می سازند. این روش ها معمولاً بهتر از یک طبقه بند منفرد عمل می کنند زیرا:
1.ترکیب چند مدل باعث کاهش واریانس پیش بینی ها می شود و اشتباهات هر مدل جداگانه تا حدی خنثی می شوند.
2.روش های دسته ای نسبت به نویز و داده های پرت مقاوم ترند چون خطاها در بین مدل ها متوسط یا رای گیری می شوند.
3.ترکیب مدل های متنوع می تواند الگوهای متفاوتی در داده ها را یاد بگیرد و در نتیجه روی داده های جدید عملکرد بهتری داشته باشد.
4.مدل های مختلف ممکن است در بخش های متفاوتی از فضای ورودی قوی تر باشند. روش دسته ای از نقاط قوت همه آنها بهره می برد.