Quiz — 5
MIR — Spring 2026
Instructor: Dr. Amin Eskandari
Head Teaching Assistants: Hamid Namjoo, Amir Hossein Hemati
Teaching Assistants: Ali Mojahed, Ali Nikvan, Amir Javvi

سوال 1

شما تحلیل‌گر امنیتی یک سامانهٔ پالایش اسپم هستید که از طبقه‌بند Naive Bayes روی یک مجموعه‌دادهٔ آموزشی بسیار کوچک با واژگان اولیهٔ {win , deal} استفاده می‌کند

داده‌های آموزشی ما شامل ایناست

اسپم (S):

a) ایمیل ۱: "win win win win" (۴ بار کلمهٔ win)

b) ایمیل ۲: "deal" (۱ بار کلمهٔ deal)

هَم (H):

a) ایمیل ۱: "win deal" (یک بار win و یک بار deal)

b) ایمیل ۲: "win" (یک بار win)

احتمالات پیشین برابرند: $P(S) = P(H) = 0.5$

برای برآورد پارامترها از نرم‌سازی لاپلاس با $\alpha = 1$ استفاده می‌کنیم

الف) محاسبهٔ پارامترها و درک تفاوت بنیادین

۱) پارامترهای احتمال شرطی هر کلمه به شرط کلاس را برای هر دو نسخهٔ Naive Bayes (یعنی مدل چندجمله‌ای و مدل برنولی) محاسبه کنید

۲) در یک جملهٔ شفاف و کوتاه، تمایز مفهومی اساسی این دو مدل را در نحوهٔ مدل‌سازی حضور/غیاب کلمات بیان کنید

ب) ایمیلی دریافت می‌شود که کاملاً خالی است (یعنی هیچ کلمه‌ای ندارد)

a) خروجی هر دو مدل (چندجمله‌ای و برنولی) را برای این ایمیل محاسبه کرده و کلاس پیش‌بینی‌شده را مشخص کنید

b) توضیح دهید که چرا مدل برنولی یک پیش‌بینی قطعی ارائه می‌دهد اما مدل چندجمله‌ای نمی‌تواند میان دو کلاس تمایز قائل شود

c) نقش هموارسازی لاپلاس در این پدیده چیست؟؟ اگر α=0بود، چه اتفاقی برای ایمیل خالی در مدل برنولی می‌افتاد؟؟؟

ب-۲) بر اساس یافته‌های بخش قبل، یک ایمیل غیرخالی طراحی کنید که این ویژگی هارا داشته باشد

a) برچسب واقعی آن «هَم» باشد

b) اما مدل چندجمله‌ای با اطمینان بالا آن را اسپم تشخیص دهد

c) در حالی که مدل برنولی به‌درستی آن را «هم» پیش‌بینی کند

ایمیل پیشنهادی خود را دقیقاً بنویسید

با محاسبهٔ احتمال پسین در هر دو مدل موفقیت حمله را نشان دهید سپس تحلیل کنید که چرا تکرار یک کلمهٔ «نسبتاً اسپم» برای مدل چندجمله‌ای فریب‌دهنده است اما مدل برنولی که فقط حضور/غیاب را می‌بیند فریب نمی‌خورد این ضعف چگونه از تفاوت ذاتی «مدل‌سازی فراوانی» در برابر «مدل‌سازی باینری» ناشی می‌شود؟؟

ب-۳) مدیر سامانه تصمیم می‌گیرد واژگان را با افزودن کلمهٔ 'free' گسترش دهد بنابراین واژگان جدید به‌صورت {win, deal, free} خواهد بود دقت کنید هیچ‌یک از ایمیل‌های آموزشی حاوی کلمهٔ 'free' نیستند

a) با همان داده‌های آموزشی و $\alpha = 1$، مقدار $P(\text{free} \mid S)$ و $P(\text{free} \mid H)$ را برای مدل چندجمله‌ای محاسبه کنید تمامی مراحل را با ذکر اعداد و نمادها نشان دهید

b) همان احتمال‌ها را برای مدل برنولی محاسبه کنید

c) سپس یک ایمیل شامل فقط یک‌بار کلمهٔ 'free' را با هر دو مدل طبقه‌بندی کرده و کلاس پیش‌بینی‌شده را برای هر مدل مشخص نمایید

d) چرا آگاهی از مقدار صحیح V (اندازهٔ واژگان) در هموارسازی مدل چندجمله‌ای حیاتی است؟؟ اگر به‌اشتباه V را برابر ۲ (واژگان اولیه) در نظر بگیریم چه تأثیری بر احتمال به‌دست‌آمده و طبقه‌بندی نهایی خواهد داشت؟

سوال 2

در یک موتور جستجوی اسناد هر سند در یک فضای دو بعدی با دو ویژگی (x1 و x2) نمایش داده می‌شود سه دسته سند (کلاس) با ویژگی‌های زیر در پایگاه داده وجود دارند

a) کلاس A (هسته مرکزی): شامل 1000 سند که با چگالی بالا در یک دایره به شعاع 1 حول مرکز (0 و 0) متمرکز شده‌اند

b) کلاس B (حلقه پیرامونی): شامل 3000 سند که در یک حلقه دایره‌ای بین شعاع 20 تا 21 به دور مرکز (0 و 0) پخش شده‌اند (نکته: مرکز هندسی این حلقه نیز همان نقطه 0 و 0 است)

c) کلاس C (شبح): شامل 20 سند که به دو گروه 10 تایی تقسیم شده‌اند گروه اول در نقطه (100+ و 100+) و گروه دوم در نقطه (100- و 100-) قرار دارند

بخش الف)الگوریتم Rocchio با محاسبه مرکز ثقل هر کلاس مرز تصمیم را تعیین می‌کند

a) توضیح دهید چرا مرکز ثقل (Centroid) کلاس A و کلاس B دقیقاً بر هم منطبق می‌شوند؟؟ با توجه به این موضوع چرا الگوریتم Rocchio در تفکیک این دو کلاس کاملاً ناتوان است؟

b) این شکست را با مفهوم بایاس مدل توضیح دهید چرا مدل Rocchio با وجود سادگی نمی‌تواند ساختار حلقوی را درک کند و کلاس A را در کلاس B هضم می‌کند؟؟

c) در استراتژی One-vs-Rest وقتی می‌خواهیم کلاس C را از "بقیه" (ترکیب A و B) جدا کنیم فراوانی بالای اسناد کلاس B چه تاثیری بر دقت تشخیص اسناد کلاس A دارد؟؟

بخش ب) فرض کنید فقط مجاز به استفاده از یک طبقه‌بند خطی هستید و نمی‌توانید ابعاد فضا را زیاد کنید (باید در فضای دو بعدی بمانید)

a) یک تابع انتقال برای ویژگی‌های (x1 و x2) پیشنهاد دهید که فضای جدید همچنان دو بعدی باشد اما در آن کلاس A و B توسط یک خط راست کاملاً از هم جدا شوند

b) توضیح دهید این انتقال چه تغییری در نمایش کلاس C ایجاد می‌کند؟؟ آیا کلاس C که در فضای قبلی به راحتی با خط جدا می‌شد در فضای جدید همچنان جداپذیر می‌ماند؟

سوال شماره 2 (2-2)

بخش الف)

فرض کنید الگوریتم k نزدیک‌ترین همسایه (kNN) را با مقدار k مساوی با کل اسناد موجود (k=N) اجرا می‌کنیم

a) ثابت کنید در این حالت نرخ خطای کلاس A برابر 100 درصد می‌شود این پدیده را با توجه به تعداد اسناد کلاس B (اکثریت) و مفهوم بایاس سیستماتیک توضیح دهید

b) اگر از روش رای‌گیری وزنی بر اساس عکس فاصله استفاده کنیم توضیح دهید چرا شانس زنده شدن کلاس A بالا می‌رود اما کلاس C همچنان در معرض حذف شدن توسط اکثریت قرار دارد؟؟

بخش ب)

a) فرض کنید در همین فضای دو بعدی، تعداد اسناد کلاس C را از 20 عدد به 20,000 عدد افزایش دهیم توضیح دهید چرا با وجود این افزایش عظیم مرزهای تصمیم kNN (با k=3) در اطراف کلاس A هیچ تغییری نمی‌کند؟؟

b) حالا فرض کنید فضا را از 2 بعد به 100,000 بعد می‌بریم با استفاده از پدیده تمرکز فاصله توضیح دهید چرا در ابعاد بالا افزایش تعداد اسناد کلاس C باعث می‌شود حتی اسناد مرکزی (کلاس A) به اشتباه در کلاس C دسته‌بندی شوند؟؟ چگونه در ابعاد بالا تعداد بر هندسه و فاصله غلبه می‌کند؟

سوال 3

یک موتور جست‌وجوی خبری پیشرفته اسناد را از منابع مختلف (خبرگزاری‌های رسمی، X ، وبلاگ‌ها و .....) جمع‌آوری می‌کند این سامانه دو وظیفه اصلی بر عهده دارد

1) طبقه‌بندی اسناد در ۵ دسته‌ی موضوعی (سیاسی، ورزشی، فناوری، سلامت، سرگرمی)

2) رتبه‌بندی نتایج جست‌وجو بر اساس اعتبار و ترجیحات کاربران

تمامی پیاده‌سازی‌ها بر اساس مدل‌های ماشین بردار پشتیبان (SVM) است

بخش اول: فرض کنید اسناد در فضای برداری TF-IDF با ۱۰۰،۰۰۰ ویژگی قرار دارند به دلیل خلوت بودن شدید داده‌ها (Sparsity) یک طبقه‌بند SVM خطی با حاشیه سخت (Hard Margin) می‌تواند داده‌های آموزشی را با دقت ۱۰۰٪ جدا کند

الف) حالا تحلیل کنید چرا در این فضای پربعد جداپذیری کامل لزوماً به معنای یافتن بزرگترین حاشیه واقعی نیست؟؟ توضیح دهید چگونه پدیده بردارهای پشتیبان تصادفی منجر به کاهش قدرت تعمیم‌دهی مدل می‌شود در نهایت مشخص کنید که پارامتر تنبیه خطا (C) در حاشیه نرم (Soft Margin) چگونه در این فضای خاص به جای پذیرش خطا نقش پایدارکننده مرز را ایفا می‌کند

ب) تیم فنی قصد دارد از استراتژی OvA برای تفکیک ۵ کلاس متوازن استفاده کند یکی از مهندسان ادعا می‌کند چون تعداد داده‌های هر کلاس برابر است هیچ مشکلی از بابت عدم توازن وجود ندارد

سوال: با یک استدلال هندسی ثابت کنید که حتی در صورت توازن کلاس‌های اصلی مجموعه بقیه (Rest) در هر مرحله از آموزش به احتمال زیاد یک مجموعه غیرمحدب (Non-convex) یا چندتکه خواهد بود این ویژگی ساختاری چه تاثیری بر حاشیه در یک SVM خطی دارد؟؟ همچنین توضیح دهید چرا در زمان امتیازدهی یک سند واضح از کلاس سیاسی ممکن است توسط طبقه‌بند سیاسی در برابر بقیه امتیازی کمتر از یک سند کاملاً بی‌ربط دریافت کند؟

بخش دوم

الف) برای استخراج الگوهای پیچیده از رفتار کاربران، پیشنهادی مبنی بر استفاده از RankSVM با کرنل غیرخطی (مانند RBF) مطرح شده است

نشان دهید چرا در مسائل رتبه‌بندی که بردار پرس‌وجو مدام تغییر می‌کند استفاده از کرنل غیرخطی در فضای تفاضلی می‌تواند منجر به نقض اصل ترایایی شود؟؟

استدلال کنید که چرا یک امتیازدهی یادگیری شده بر پایه SVM خطی (فرمول: Score = W.X) سازگاری منطقی رتبه‌بندی را بهتر از هر مدل غیرخطی پیچیده‌ای تضمین می‌کند

ب) در داده‌های واقعی کاربران اغلب نظرات متناقضی دارند

در الگوریتم RankSVM حاشیه (Margin) بیانگر حداقل فاصله مورد انتظار بین رتبه‌ها است توضیح دهید در صورت وجود این تناقضات افزایش پارامتر C چگونه باعث می‌شود مدل به جای یادگیری ترجیح عام روی نویز آماری قفل شود؟؟

به عنوان یک راهکار خلاقانه چگونه می‌توان مفهوم حاشیه را بازتعریف کرد تا به صورت پویا برای زوج‌سندهایی که کاربران روی آن‌ها توافق ندارند، انعطاف‌پذیرتر از زوج‌سندهای مورد اتفاق‌نظر عمل کند؟

سوال 4

شما به عنوان تحلیل‌گر ارشد در پروژه مایا وظیفه دارید سیگنال‌های متنی بازسازی شده از کتیبه‌های باستانی را به سه کلاس موضوعی (C1, C2, C3) طبقه‌بندی کنید داده‌های شما دارای ویژگی‌های فنی زیر هستند

a) ابعاد نجومی>>> تعداد ویژگی‌ها (الگوهای متنی) حدود یک میلیون است
b) توزیع به شدت پراکنده (Sparse) >>> به دلیل ماهیت باستانی 99 درصد از ویژگی‌ها در کل داده‌ها فقط یک یا دو بار ظاهر شده‌اند
c) طول متغیر >>> برخی کتیبه‌ها بسیار کوتاه و برخی بسیار طولانی هستند

علی استراتژیست تیم، سه پیشنهاد زیر را برای پیاده‌سازی سیستم مطرح کرده است

پیشنهاد اول: استفاده از مدل Naive Bayes برنولی به همراه هموارسازی لاپلاس (با پارامتر alpha=1) استدلال او این است که در داده‌های تنُک فقط حضور یا عدم حضور الگوها مهم است و شمارش تکرار (TF) باعث نویز می‌شود

پیشنهاد دوم : استفاده از طبقه‌بند k نزدیک‌ترین همسایه (k-NN) با مقدار (k = رادیکال N) که N تعداد کل نمونه‌هاست همچنین استفاده از فاصله اقلیدسی پس از نرم‌سازی طول بردارها

پیشنهاد سوم : استفاده از SVM غیرخطی با هسته RBF و حاشیه سخت به روش One-vs-One ، استدلال این است که چون تعداد ویژگی‌ها (d) بسیار بیشتر از تعداد نمونه‌ها (n) است داده‌ها قطعاً در فضای بالاتر تفکیک‌پذیر هستند و باید بدون خطا (حاشیه سخت) آموزش ببینند

بخش الف) با تحلیل دقیق اثر طول سند و تعداد ویژگی‌ها توضیح دهید چرا استفاده از Laplace Smoothing در فضایی با یک میلیون ویژگی باعث می‌شود مدل برنولی به جای تمرکز بر کلمات موجود در سند بر اساس کلمات غایب تصمیم‌گیری کند؟؟ چگونه این موضوع منجر به یک بایاس شدید به سمت کلاس اکثریت می‌شود؟؟؟

الف-2) فرض کنید یک سند بسیار کوتاه دارید که شامل 3 الگوی بسیار کلیدی و نادر است که فقط در کلاس C1 دیده شده‌اند چرا مدل پیشنهادی (برنولی با alpha=1) ممکن است به شکلی فاحش این سند را در کلاسی دیگر قرار دهد؟؟؟ در این شرایط مدل Multinomial NB چه رفتاری متفاوتی از خود نشان می‌دهد؟؟

ب) استدلال پیشنهاد دوم را نقد کنید: چرا در یک فضای یک میلیون بعدی حتی با نرم‌سازی بردارها پدیده تراکم فواصل(Distance Concentration) رخ می‌دهد؟؟

توضیح دهید چرا انتخاب (k = رادیکال N) در این فضای خاص عملاً مدل را به یک رای‌گیری ساده از کلاس اکثریت تبدیل کرده و مفهوم همسایگی محلی را نابود می‌کند؟

در پیشنهاد سوم چرا ترکیب هسته RBF و Hard Margin در شرایطی که d بسیار بزرگتر از n است مصداق بارز خطای استراتژیک در مبادله بایاس-واریانس محسوب می‌شود؟

توضیح دهید چگونه این مدل به جای یادگیری الگوی زبان جزیره‌های تصمیم دور داده‌های نویزی ایجاد می‌کند و چرا یک Linear SVM با حاشیه نرم انتخاب مهندسی‌تری است؟؟؟

موفق و موید :(