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

پاسخ سوال 1

واژگان اولیه: {win, deal} (تعداد کلمات واژگان یا همان V برابر است با 2)

مقدار هموارسازی (Alpha) برابر است با 1

کلاس اسپم (S): شامل 2 ایمیل

تعداد دفعات تکرار کلمه win >>> برابر 4 بار

تعداد دفعات تکرار کلمه deal >>> برابر 1 بار

مجموع کل کلمات در این کلاس >>> 5 کلمه

تعداد اسنادی که شامل win هستند >>> 1 سند

تعداد اسنادی که شامل deal هستند >>> 1 سند

کلاس هَم (H): شامل 2 ایمیل

تعداد دفعات تکرار کلمه win >>> برابر 2 بار

تعداد دفعات تکرار کلمه deal >>> برابر 1 بار

مجموع کل کلمات در این کلاس >>> 3 کلمه

تعداد اسنادی که شامل win هستند >>> 2 سند

تعداد اسنادی که شامل deal هستند >>> 1 سند

بخش الف) محاسبات مدل چندجمله‌ای (Multinomial NB)

در این مدل تعداد کل تکرارها مهم است

فرمول ساده: (تعداد تکرار کلمه + 1) تقسیم بر (مجموع کلمات کلاس + تعداد اعضای واژگان)

کلمه win در کلاس اسپم >>> (4 + 1) / (5 + 2) = 7/5 (تقریباً 0.71)

کلمه deal در کلاس اسپم >>> (1 + 1) / (5 + 2) = 7/2 (تقریباً 0.28)

کلمه win در کلاس هم >>> (2 + 1) / (3 + 2) = 5/3 (برابر 0.6)

کلمه deal در کلاس هم >>> (1 + 1) / (3 + 2) = 5/2 (برابر 0.4)

محاسبات مدل برنولی (Bernoulli NB) >>> در این مدل فقط تعداد اسنادی که کلمه در آن‌ها هست مهم است

فرمول ساده >>> (تعداد سندهای شامل کلمه + 1) تقسیم بر (کل سندهای کلاس + 2)

کلمه win در کلاس اسپم >>> (1 + 1) / (2 + 2) = 4/2 (برابر 0.5)

کلمه deal در کلاس اسپم >>> (1 + 1) / (2 + 2) = 4/2 (برابر 0.5)

کلمه win در کلاس هم >>> (2 + 1) / (2 + 2) = 4/3 (برابر 0.75)

کلمه deal در کلاس هم >>> (1 + 1) / (2 + 2) = 4/2 (برابر 0.5)

دقت کنید مدل چندجمله‌ای تکرار را می‌بیند، اما مدل برنولی فقط حضور یا غیبت را می‌سنجد

بخش ب)

ب-1) معمای ایمیل خالی ایمیلی که هیچ کلمه‌ای ندارد

در مدل چندجمله‌ای: چون کلمه‌ای نیست هیچ ضربی انجام نمی‌شود فقط احتمال اولیه کلاس‌ها (0.5) باقی می‌ماند هر دو کلاس امتیاز مساوی می‌گیرند مدل گیج می‌شود و نمی‌تواند تصمیم بگیرد

در مدل برنولی: این مدل غیبت کلمات را هم حساب می‌کند

امتیاز اسپم: (احتمال اولیه 0.5) ضربدر (احتمال نبودن win که 0.5 است) ضربدر (احتمال نبودن deal که 0.5 است) 0.125

امتیاز هَم: (احتمال اولیه 0.5) ضربدر (احتمال نبودن win که 0.25 است) ضربدر (احتمال نبودن deal که 0.5 است) 0.0625

نتیجه >>> مدل برنولی ایمیل خالی را اسپم تشخیص می‌دهد چون امتیاز اسپم بیشتر است

هموارسازی لاپلاس باعث می‌شود احتمال نبودن یک کلمه هرگز صفر نشود و مدل بتواند حتی برای هیچ یک احتمال حساب کند

ب-2)ایمیل طراحی شده>>> "win win win win win win"

چرا در مدل چندجمله‌ای شکست می‌خورد؟ در این مدل چون احتمال win در اسپم (0.71) بزرگتر از احتمال آن در هم (0.6) است هر چقدر این کلمه را بیشتر تکرار کنیم امتیاز اسپم به شدت افزایش یافته و مدل فریب می‌خورد که این یک اسپم است

چرا در مدل برنولی موفق است؟؟ برنولی فقط می‌بیند که کلمه win هست (یک بار) و کلمه deal نیست تکرار 6 باره win هیچ فرقی با 1 بار آمدن آن ندارد چون احتمال حضور win در کلاس هم (0.75) بیشتر از اسپم (0.5) است مدل به درستی آن را هم تشخیص می‌دهد

ب-3) گسترش واژگان و کلمه جدید (free)

واژگان جدید شد >>> {win, deal, free} پس V برابر 3 است کلمه free در آموزش نبوده است

محاسبات جدید در مدل چندجمله‌ای (تغییر مخرج):

احتمال free در اسپم: (0 + 1) / (5 + 3) = 8/1 (برابر 0.125)

احتمال free در هم: (0 + 1) / (3 + 3) = 6/1 (برابر 0.166)

محاسبات جدید در مدل برنولی >>> احتمال حضور free در هر دو کلاس برابر است با: (0 + 1) / (2 + 2) = 0.25

نتیجه آزمایش ذهنی (ایمیل شامل فقط کلمه free):

در چندجمله‌ای: امتیاز هم بیشتر می‌شود (0.083 در مقابل 0.0625) پس کلاس هم انتخاب می‌شود

در برنولی: چون کلمات win و deal غایب هستند جریمه غیبت برای کلاس هم که win در آن خیلی رایج بود بسیار سنگین‌تر است و مدل به سمت اسپم متمایل می‌شود

نکته نهایی >>> دانستن مقدار دقیق V (سایز واژگان) حیاتی است چون در مدل چندجمله‌ای تمام احتمال‌های قبلی را با تغییر مخرج دستخوش تغییر می‌کند اگر V غلط باشد مجموع احتمالات مدل از 1 بیشتر یا کمتر می‌شود و پیش‌بینی‌ها کاملاً غیرقابل اعتماد خواهند بود

پاسخ سوال 2

بخش الف) چرا الگوریتم Rocchio شکست می‌خورد؟؟

تحلیل مراکز ثقل (Centroids):

منطق ساده >>> الگوریتم Rocchio هر دسته را با میانگین نقاط آن می‌شناسد

در کلاس A >>> اسناد به صورت دایره‌ای دور نقطه (0 و 0) هستند چون توزیع متقارن است یعنی به همان اندازه که در سمت راست نقطه هست در سمت چپ هم هست میانگین تمام این نقاط دقیقاً همان مرکز یعنی (0 و 0) می‌شود

در کلاس B >>> این اسناد هم به صورت یک حلقه دور مرکز (0 و 0) هستند باز هم به دلیل تقارن، میانگین تمام نقاط این حلقه، همان نقطه (0 و 0) می‌شود

نتیجه>>> چون مرکز هر دو کلاس یکسان است الگوریتم فکر می‌کند هر دو دسته یکی هستند و نمی‌تواند مرزی بین آن‌ها بکشد

تحلیل بایاس-واریانس (Bias-Variance)

بایاس (Bias)>>> یعنی پیش‌فرض‌های مدل، مدل Rocchio پیش‌فرضش این است که هر دسته باید یک توده متمرکز شبیه توپ باشد به این می‌گوییم بایاس بالا این مدل نمی‌تواند بفهمد که یک دسته می‌تواند به شکل حلقه باشد

نتیجه>>> چون مدل بیش از حد ساده است (بایاس بالا) ساختار پیچیده حلقه را نمی‌بیند و شکست می‌خورد

چالش One-vs-Rest (یک در مقابل بقیه) >>> وقتی کلاس C (که دور است) را در مقابل بقیه (A و B) قرار می‌دهیم چون تعداد اسناد کلاس B خیلی زیاد است (3000 تا) میانگینِ بقیه به شدت تحت تاثیر کلاس B قرار می‌گیرد این باعث می‌شود مرز تصمیم‌گیری طوری کج شود که کلاس A (که تعدادش کمتر است) نادیده گرفته شود

بخش ب)

فرمول نگاشت پیشنهاد شده ما باید از مختصات معمولی (x1 و x2) به مختصات شعاعی برویم

ویژگی جدید اول (z1) = شعاع = جزر (x1 به توان دو + x2 به توان دو)

ویژگی جدید دوم (z2) = زاویه

در این فضای جدید چه اتفاقی می‌افتد؟ کلاس A همگی شعاع‌شان بین 0 تا 1 است کلاس B همگی شعاع‌شان بین 20 تا 21 است حالا به سادگی با یک خط راست (مثلاً در شعاع 10) می‌توان این دو را از هم جدا کرد

تحلیل اثر: این کار باعث می‌شود یک مرز دایره‌ای در دنیای واقعی برای کامپیوتر به شکل یک خط راست دیده شود

کلاس C: این کلاس که در نقطه (100 و 100) بود حالا شعاعش حدود 141 می‌شود و همچنان در جای دورتری نسبت به بقیه قرار می‌گیرد و جداپذیر باقی می‌ماند

پاسخ سوال شماره 2:

بخش الف) وقتی k برابر با کل داده‌ها (k=N) باشد

1. تحلیل اکثریت:در الگوریتم kNN برای برچسب زدن به یک نقطه رای‌گیری می‌شود وقتی k برابر با کل داده‌ها باشد یعنی همه اسناد (4020 سند) در رای‌گیری شرکت می‌کنند

چون کلاس B تعدادش 3000 تاست و از بقیه خیلی بیشتر است، در هر رای‌گیری کلاس B برنده می‌شود

نتیجه: هر کجای نقشه که دست بگذاریم سیستم می‌گوید این کلاس B است پس کلاس A که 1000 عضو دارد در تمام رای‌گیری‌ها می‌بازد و خطایش 100 درصد می‌شود

بخش ب) 1. چرا در فضای دو بعدی افزایش داده C بی‌اثر است؟

الگوریتم kNN یک الگوریتم "محلی" (Local) است. یعنی فقط به همسایه‌های نزدیک نگاه می‌کند.

وقتی ما در مرکز (کلاس A) هستیم، کلاس C در فاصله خیلی دوری (نقطه 100) قرار دارد حالا کلاس C می‌خواهد 20 تا باشد یا 20 هزار تا چون نزدیک ما نیست اصلاً در رای‌گیری برای نقاط نزدیک به A شرکت داده نمی‌شود

2. نفرین ابعاد (فضای 100,000 بعدی): در ابعاد بسیار بالا (مثل 100 هزار بعد)، یک اتفاق عجیب می‌افتد: فاصله همه چیز از هم تقریباً برابر می‌شود دیگر چیزی به نام نزدیک یا دور به آن معنای واقعی وجود ندارد در این حالت kNN که بر اساس فاصله کار می‌کرد گیج می‌شود

وقتی فاصله‌ها تقریباً یکی شد دیگر نزدیکی اهمیتی ندارد و فقط تعداد مهم می‌شود چون کلاس C تعدادش خیلی زیاد شده 20 هزار تا در هر کجای این فضای پربعد که باشیم وقتی بخواهیم 3 همسایه نزدیک را انتخاب کنیم به احتمال زیاد از کلاس C خواهند بود چون تعدادشان در کل فضا غالب است

نتیجه >>> در ابعاد بالا، هندسه و فاصله معنایش را از دست می‌دهد و کلاس پرجمعیت بقیه را می‌بلعد

پاسخ سوال 3

پاسخ سوال ۱-الف) در محیط‌های متنی با فضای ویژگی‌های بسیار بزرگ و داده‌های محدود روبرو هستیم یعنی خلوت بودن داده‌ها در این شرایط مدل به راحتی می‌تواند یک مرز جداکننده سخت پیدا کند که دقت ۱۰۰٪ در آموزش داشته باشد اما این دقت فریبه چون مدل به جای یادگیری «مفهوم» اخبار بر روی کلمات تصادفی و نویزهای موجود در داده‌های آموزشی Overfitting کرده است استفاده از پارامتر C کوچک در SVM حاشیه نرم یک راهبرد هوشمندانه برای مقابله با این پدیده است با کوچک کردن C ما به مدل اجازه می‌دهیم برخی داده‌ها را نادیده بگیرد تا عرض حاشیه افزایش یابد حاشیه پهن‌تر یعنی مدل به جای تمرکز بر جزئیاتِ فرعی بر جهت کلی بردارها تمرکز می‌کند که منجر به تعمیم‌پذیری بهتر در اخبار جدید می‌شود

جمع‌بندی >>> در متن جداپذیری آسان اما تعمیم‌دهی سخت است دقت ۱۰۰٪ در آموزش هشدارِ بیش‌برازش است C کوچک ابزارِ کنترلِ «ساده‌سازی مدل» برای رسیدن به پایداری است

پاسخ سوال ۱-ب) در استراتژی OvA کلاس بقیه در واقع مجموعه‌ای از کلاس‌های متنوع (ورزش، سلامت، فناوری و...) است از نظر هندسی این کلاس بقیه یک موجودیت واحد و خوش‌تعریف نیست بلکه مجموعه‌ای چندتکه و غیرمحدب در اطراف کلاس هدف است یک SVM خطی که از یک ابرصفحه (تیغ صاف) برای جداسازی استفاده می‌کند نمی‌تواند به درستی دور یک مجموعه غیرمحدب را خط بکشد این عدم توازن باعث می‌شود که امتیازدهیِ اطمینان (Confidence) مختل شود چرا که مرز تصمیم‌گیری به دلیل فشار غیرمتوازنِ کلاس‌های مختلفِ موجود در بقیه دچار سوگیری شده و اسناد نامرتبط ممکن است امتیازهای کاذب بالایی کسب کنند

جمع‌بندی کلاس بقیه یک آشوب هندسی است استراتژی OvA به دلیل عدم توازن ساختاری در تولید امتیازهای اطمینان قابل‌اعتماد برای رتبه‌بندی دچار خطا می‌شود

پاسخ سوال ۲-الف

اصل ترایایی در رتبه‌بندی یعنی اگر سند الف از ب بهتر باشد و ب از ج بهتر باشد حتماً الف باید از ج بهتر باشد در مدل RankSVM خطی تمام اسناد به یک عدد واحد (Score) تبدیل می‌شوند که روی یک خط مرتب شده‌اند بنابراین نظم منطقی همیشه حفظ می‌شود اما کرنل‌های غیرخطی (مانند RBF) با بررسی روابط پیچیده و محلی این نظم را به هم می‌ریزند یک کرنل غیرخطی ممکن است در نواحی مختلف فضا

قضاوت‌های متناقضی داشته باشد که منجر به ایجاد دور (Cycle) در رتبه‌بندی شود یعنی الف > ب > ج > الف

این وضعیت برای یک موتور جستجو فاجعه‌بار است زیرا سیستم نمی‌تواند یک لیست رتبه‌بندی واحد و پایدار ارائه دهد

جمع‌بندی در رتبه‌بندی نظم جهانی بر دقت محلی اولویت دارد مدل خطی تضمین‌کننده یک تابع امتیازدهی واحد است در حالی که مدل غیرخطی می‌تواند با نقض ترایایی سیستم رتبه‌بندی را دچار تناقض کند

پاسخ سوال ۲-ب

داده‌های حاصل از کلیک کاربران اغلب متناقض هستند مثلاً توافقِ همگانی روی برتری الف بر ب وجود ندارد، اگر پارامتر C بزرگ باشد مدل با حساسیت بیش از حد سعی می‌کند تمام این ترجیحات متناقض را در بردار وزن‌ها بگنجاند که باعث نویزپذیری شدید مدل می‌شود راهکار بهینه حاشیه پویا(Dynamic Margin) است در این رویکرد ما عرض حاشیه را ثابت نمی‌گیریم بلکه آن را بر اساس درجه توافق کاربران تنظیم می‌کنیم اگر توافق روی یک زوج سند بالا باشد حاشیه را بزرگ در نظر می‌گیریم تا مدل آن را یاد بگیرد و اگر توافق کم باشد حاشیه را کوچک می‌کنیم تا مدل با داده‌های مشکوک به راحتی کنار بیاید

جمع‌بندی مدل رتبه‌بندی باید بین اطمینان کاربر و سخت‌گیری مدل تعادل ایجاد کند تنظیم حاشیه بر اساس توافق راهکاری برای حفظ ساختار کلی رتبه‌بندی در برابر نویزهای انسانی است

پاسخ سوال 4

پاسخ بخش اول:

منطق مدل برنولی>>> مدل برنولی برای هر سند احتمال را بر اساس حضور کلمات موجود ضرب‌در عدم حضور کلمات غایب محاسبه می‌کند

اثر ابعاد یک میلیونی در یک کتیبه که فقط ۱۰ کلمه دارد مدل برنولی باید احتمال عدم حضور ۹۹۹,۹۹۰ کلمه دیگر را محاسبه کند یعنی حاصل‌ضرب نهایی تحت کنترل مطلق کلمات غایب قرار می‌گیرد

نقش مخرب هموارسازی لاپلاس: فرمول احتمال در برنولی با هموارسازی لاپلاس به این صورت است (تعداد اسناد شامل کلمه + ۱) تقسیم بر (تعداد کل اسناد کلاس + ۲)

در فضای یک میلیون بعدی وقتی alpha را برابر ۱ می‌گیریم این اثر در مخرج کسرها انباشته می‌شود چون تعداد کلمات غایب بسیار زیاد است مدل به سمتی متمایل می‌شود که "تعداد اسناد آموزشی بیشتری" دارد در واقع سیگنالِ ضعیف کلماتِ موجود در سند در میان نویزِ ناشی از هموارسازیِ یک میلیون کلمه غایب گم می‌شود

سوال ۲: مدل چندجمله‌ای فقط کلمات موجود در سند را می‌بیند اگر ۳ کلمه نادر مربوط به کلاس اول (C1) در سند باشد فقط احتمال همان ۳ کلمه در هم ضرب می‌شود

عدم جریمه برای غیبت >>> برخلاف برنولی مدل چندجمله‌ای برای ۹۹۹,۹۹۷ کلمه‌ای که در سند نیستند جریمه‌ای لحاظ نمی‌کند بنابراین حتی اگر سند بسیار کوتاه باشد هویت آن که توسط آن ۳ کلمه تعیین شده حفظ می‌شود

جمع‌بندی >>> استفاده از برنولی در فضای ویژگی بزرگ اشتباه است چون جریمه عدم حضور کلمات بر تایید حضور کلمات غلبه کرده و باعث بایاس شدید به سمت کلاس اکثریت می‌شود مدل چندجمله‌ای به دلیل نادیده گرفتن کلمات غایب در داده‌های متنی Sparse بسیار پایدارتر عمل می‌کند

پاسخ بخش دوم

پدیده تمرکز فواصل (Distance Concentration): در فضای یک میلیون بعدی به دلیل پراکندگی زیاد فاصله اقلیدسی بین هر دو سند تصادفی تقریباً با هم برابر می‌شود یعنی نزدیک‌ترین همسایه با دورترین همسایه تفاوت چندانی در فاصله ندارد

مشکل انتخاب k بزرگ (رادیکال N): وقتی فواصل با هم فرقی ندارند انتخاب تعداد زیادی همسایه (k بالا) باعث می‌شود مدل به جای پیدا کردن شباهت‌های محلی عملاً یک رای‌گیری عمومی از داده‌های

اطراف انجام دهد چون در کل پایگاه داده کلاس اکثریت تعداد نمونه بیشتری دارد احتمال اینکه در میان k همسایه اکثریت با کلاس بزرگتر باشد بسیار زیاد است این یعنی مدل k-NN دیگر هوشمند نیست و فقط کلاس اکثریت را برمی‌گرداند

سوال ۲:

عدم نیاز به هسته غیرخطی>>> در طبقه‌بندی متن وقتی تعداد ویژگی‌ها (یک میلیون) بسیار بیشتر از تعداد نمونه‌هاست داده‌ها معمولاً در همان فضای اصلی تفکیک‌پذیر خطی هستند استفاده از هسته RBF پیچیدگی محاسباتی و خطر اورفیت (Overfitting) را بی‌دلیل بالا می‌برد

فاجعه حاشیه سخت (Hard Margin): حاشیه سخت یعنی مدل حق ندارد هیچ اشتباهی در داده‌های آموزشی داشته باشد در داده‌های نویزی باستانی این کار باعث می‌شود مدل مرزهای تصمیم بسیار پیچیده و عجیبی دور تک‌تک داده‌های نویزی بسازد (واریانس بالا) این مرزها در مواجهه با داده‌های جدید کاملاً شکست می‌خورند

چرا Linear Soft-Margin SVM بهترین است؟؟

خطی بودن (Linear): سادگی را حفظ کرده و برای ابعاد بالا کفایت می‌کند

حاشیه نرم (Small C): به مدل اجازه می‌دهد از برخی داده‌های نویزی چشم‌پوشی کند تا مرز تصمیم صاف و کلی باقی بماند این کار باعث کاهش واریانس و افزایش قدرت تعمیم‌پذیری (Generalization) می‌شود

جمع‌بندی >>> در ابعاد بالا، فاصله اقلیدسی کارایی ندارد و k-NN به سمت پیش‌بینی کلاس اکثریت سقوط می‌کند همچنین استفاده از مدل‌های پیچیده مثل SVM با هسته RBF و حاشیه سخت روی داده‌های متنی منجر به یادگیری نویز (Overfitting) می‌شود بهترین انتخاب یک مدل خطی با حاشیه نرم است که تعادل بین بایاس و واریانس را رعایت کند

موفق و موید:(