به نام خداوند بخشنده مهربان

# طراحی سیستم‌های یادگیری ماشین

دکتر سیدصالحی
پاییز ۱۴۰۴

تمرین سری اول

**پرسش‌های تئوری** (۱۱۰ نمره) مهلت ارسال: ۱۴۰۴/۰۹/۰۷ ساعت ۲۳:۵۹:۵۹، با تاخیر: ۱۴۰۴/۰۹/۱۰ ساعت ۲۳:۵۹:۵۹

**پرسش ۱** (۱۵ نمره)

## استخدام نیروهای خود را به ما بسپارید!

فرض کنید شما یکی از اعضای تیم یادگیری ماشین در یک شرکت بزرگ استخدامی هستید که از یک سیستم هوشمند برای غربال رزومه‌ها و رتبه‌بندی متقاضیان شغلی استفاده می‌کند. شرکت می‌خواهد مدل فعلی خود را ارتقا دهد تا استخدام سریع‌تر و دقیق‌تری انجام دهد.

در حال حاضر سه نسخه از مدل در دست آزمایش است:

جدول ۱: مقایسه مدل‌های پیشنهادی

| مدل | توضیح | تأخیر پاسخ | دقت | انصاف | تفسیرپذیری |
|---|---|---|---|---|---|
| مدل A | مدل عمیق با بالاترین دقت ممکن، آموزش‌دیده‌شده روی داده‌های تاریخی شرکت | ۲ ثانیه | ٪۹۶ | پایین | بسیار پایین |
| مدل B | مدل تقطیرشده از مدل A ، سریع‌تر با کمی کاهش دقت | ۱/۱ ثانیه | ٪۹۳ | بهبود یافته (سوگیری قابل کنترل) | متوسط |
| مدل C | مدل خطی قابل تفسیر (مثل Logistic Regression) | ۱ ثانیه | ٪۶۵ | متوسط | بسیار بالا |

هیئت‌مدیره از شما خواسته است یکی از این مدل‌ها را برای استقرار نهایی انتخاب کنید. هر سه مدل آزمایش‌های اولیه را گذرانده‌اند، اما شرایط زیر وجود دارد:

- **تیم منابع انسانی:** سیستم باید عادلانه و قابل توضیح باشد، تا در برابر شکایات قانونی پاسخ‌گو باشند.
- **تیم محصول:** بر سرعت پاسخ در مصاحبه‌های آنلاین تأکید دارد (کاربران نباید منتظر بمانند).
- **تیم فنی:** تمایل دارد دقت مدل بالا باشد تا از رد شدن افراد مناسب جلوگیری شود.
- **مدیریت ارشد:** می‌خواهد احتمال شکایت و آسیب به برند شرکت به حداقل برسد.

با توجه به داده‌های جدول و محدودیت‌های عملیاتی و اخلاقی شرکت، تصمیم بگیرید کدام مدل (A, B, C) باید در محیط تولید استفاده شود و چرا. در پاسخ خود موارد زیر را تحلیل کنید:

- تأثیر تأخیر بر تجربهٔ کاربر و بهره‌وری سیستم
- نقش انصاف (Fairness) در اعتماد کاربران و ریسک حقوقی شرکت
- اهمیت تفسیرپذیری در تصمیم‌های انسانی حساس
- چگونگی ایجاد موازنه میان این سه عامل

**پرسش ۲** (۱۵ نمره)

## دقت را می‌خواهی، تاخیرش را هم بخر!

یک مدل ensemble شامل $n$ زیرمدل است که دقت (Top-1 Accuracy) را به‌صورت خطی به میزان $\alpha n$ افزایش می‌دهد، اما تأخیر یا زمان پاسخ را به‌صورت نمایی طبق رابطه‌ی زیر افزایش می‌دهد:

$$L = L_{\cdot}(۱/۳)^n$$

۱. مقدار $n^*$ را بیابید که تابع مطلوبیت زیر را بیشینه کند:

$$U = \alpha n - \lambda L$$

که در آن $\lambda$ جریمه‌ی مربوط به تأخیر (latency penalty) است.

۲. نتیجه را از دید طراحی سامانه تفسیر کنید — توضیح دهید اگر آستانه‌ی قابل‌قبول بین افزایش دقت و تأخیر مشخص باشد، چند زیرمدل باید در محیط تولید (production) مستقر شوند.

**پرسش ۳** (۱۰ نمره)

## بچ یا استریم؟

فرض کنید یک feature store تعداد $M$ ویژگی را برای $K$ مدل هم‌زمان فراهم می‌کند. هر ویژگی دارای هزینه‌ی بازمحاسبه‌ی $C_i$ و نرخ کهنگی یا افت تازگی $\lambda_i$ است.

۱. تابع زیان کلی تازگی را به‌صورت زیر مدل کنید:

$$L = \sum_i ۱ - e^{-\lambda_i \Delta t_i}$$

که در آن $\Delta t_i$ بازه‌ی زمانی بین دو بازمحاسبه‌ی متوالی ویژگی است.

۲. سپس، تابع هزینه‌ی کل را به‌شکل زیر بنویسید:

$$J = \sum_i \left( \frac{C_i}{\Delta t_i} + \beta(۱ - e^{-\lambda_i \Delta t_i}) \right)$$

و بازه‌ی زمانی بهینه‌ی $\Delta t_i^*$ را به‌گونه‌ای بیابید که $J$ را کمینه کند.

۳. نتیجه را تفسیر کنید: در چه شرایطی بهتر است یک ویژگی از حالت batch به حالت streaming منتقل شود؟

**پرسش ۴** (۱۰ نمره)

## طراحی سامانه‌ی پیشنهاددهنده‌ی بلادرنگ

فرض کنید در یک شرکت تجارت الکترونیکی بزرگ (مانند آمازون یا دیجیکالا) کار می‌کنید که از یک سیستم پیشنهاددهندهٔ بلادرنگ Real-time recommendation system برای نمایش محصولات به کاربران استفاده می‌کند. داده‌های کاربران از طریق Kafka به‌صورت لحظه‌ای ارسال می‌شوند و مدل‌های یادگیری ماشین از این داده‌ها برای پیش‌بینی پیشنهادها استفاده می‌کنند. شرکت در نظر دارد قالب پیام‌های خود را از JSON به Protobuf تغییر دهد تا کارایی و سرعت سیستم افزایش یابد.

۱. توضیح دهید چرا در سامانه‌های بلادرنگ، قالب‌های باینری مانند Protobuf نسبت به قالب‌های متنی مانند JSON عملکرد بهتری دارند.

۲. فرض کنید سامانهٔ شما سه بخش دارد:

بخش تراکنشی (پرداخت و سفارش) در MySQL،

جریان رویدادهای کاربری در Kafka،

سرویس استنتاج بلادرنگ از طریق gRPC.

توضیح دهید هرکدام از این سه بخش از چه نوع جریان داده‌ای (پایگاه‌داده‌ای، پیام‌محور یا RPC) استفاده می‌کنند و چرا این انتخاب برای کارایی و پایداری مناسب است.

۳. در عمل، مدل‌های پیشنهاددهنده باید هر شب بازآموزی شوند تا با رفتار جدید کاربران هماهنگ بمانند. توضیح دهید چگونه داده‌های Protobuf در Kafka می‌توانند برای آموزش مدل در زمان‌های مشخص جمع‌آوری شوند، و چرا هماهنگی میان Pipeline بلادرنگ (Online) و Pipeline آموزشی (Offline) حیاتی است.

**پرسش ۵** (۱۰ نمره)

## گمشده‌ای در جدول

فرض کنید داده‌ای در اختیار دارید که حدود ۱۵٪ مقادیر مفقود (missing values) دارد و این مقادیر به‌صورت غیر یکنواخت میان ویژگی‌ها (features) توزیع شده‌اند.
چهار روش مختلف برای تخمین یا جایگزینی مقادیر مفقود فهرست کنید و هر کدام را از نظر سوگیری (bias)، قابلیت تفسیر (interpretability)، و امکان‌پذیری در محیط تولیدی (production feasibility) ارزیابی کنید.

**پرسش ۶** (۱۰ نمره)

## بوی نَشت داده

تیم شما پس از افزودن صدها ویژگی جدید به مدل، کاهش محسوسی در عملکرد مشاهده می‌کند.
سه تحلیل تشخیصی (diagnostic analysis) پیشنهاد دهید تا مشخص شود آیا علت افت عملکرد تکرار و هم‌بستگی ویژگی‌ها (feature redundancy)، نشت داده (data leakage)، یا رانش ویژگی‌ها (feature drift) است.
برای هر تحلیل، متریکی یا ابزاری را که استفاده می‌کنید مشخص کنید: مانند اطلاعات متقابل (Mutual Information)، شاخص پایداری جمعیت (Population Stability Index – PSI)، یا کاهش اهمیت ویژگی‌ها در طول زمان (Feature Importance Decay).

**پرسش ۷** (۱۵ نمره)

## فهم درست مسأله، نیمی از حل مسأله است!

فرض کنید در یک پلتفرم ویدیو (مثل فیلیمو یا نتفلیکس) می‌خواهید سیستمی طراحی کنید که پیش‌بینی کند کاربر بعدی کدام ویدیو را تماشا خواهد کرد. دو روش زیر برای مدل‌سازی پیشنهاد شده است:

۱. طبقه‌بندی چندکلاسه (Multiclass Classification): هر ویدیو یک کلاس است و مدل احتمال تماشای هر ویدیو را پیش‌بینی می‌کند.

۲. مدل رگرسیونی امتیازدهی (Regression Scoring): برای هر کاربر–ویدیو، یک احتمال علاقه‌مندی پیش‌بینی می‌شود.

الف) توضیح دهید در این سناریو، تفاوت این دو روش از نظر مقیاس‌پذیری و هزینهٔ بازآموزی چیست و کدام‌یک برای سیستمی با هزاران ویدیو مناسب‌تر است؟

ب) فرض کنید مدل A باعث افزایش نرخ کلیک (CTR) شده اما میانگین زمان تماشای کاربران کاهش یافته است. این اتفاق چه چیزی دربارهٔ عدم هم‌راستایی متریک‌های فنی با اهداف تجاری نشان می‌دهد و چگونه می‌توان متریک مناسب‌تری تعریف کرد؟

**پرسش ۸** (۱۵ نمره)

## هم خدا رو می‌خوایم هم خرما رو!

در طراحی یک سیستم رتبه‌بندی خبر، برقراری تعادل میان اهداف متضاد مانند بیشینه‌سازی تعامل کاربر و کاهش اطلاعات نادرست اهمیت دارد. دو رویکرد زیر را مقایسه و تحلیل کنید:

۱. استفاده از یک مدل واحد با تابع زیان ترکیبی (Combined Loss Function)

۲. استفاده از چند مدل مستقل، هرکدام بهینه‌شده برای یک هدف خاص

تفاوت‌های آن‌ها را از نظر انعطاف‌پذیری، تفسیرپذیری، هزینهٔ بازآموزی و نگهداشت توضیح دهید، و بیان کنید این موازنه‌ها چگونه چالش‌های واقعی بهینه‌سازی چندهدفه را بازتاب می‌دهند.

لطفا برای این موارد، روابط loss های مدنظر که در اسلایدهای درس هم آمده‌بودند را بیان کنید و باتوجه به روابط این موارد را توضیح دهید. به توضیح موارد سوال بدون اشاره به روابط ریاضی loss نمره‌ای تعلق نخواهد گرفت.

**پرسش ۹** (۱۰ نمره)

## نامتوازن بودن داده

شرکتی در حال ساخت مدل کشف تقلب است. تنها ۱٪ تراکنش‌ها تقلبی‌اند و داده‌ها (۱۰۰ میلیون رکورد) عمدتاً از یک منطقه جغرافیایی جمع‌آوری شده‌اند.

۱. شرح دهید به‌صورت خلاصه هر یک از این روش‌ها چگونه می‌تواند نمایندگی یا سوگیری را تغییر دهد:

نمونه‌گیری تصادفی ساده

نمونه‌گیری طبقه‌ای (stratified)

نمونه‌گیری وزنی یا اهمیت‌محور (weighted / importance)

۲. چرا معیار accuracy برای این مسئله گمراه کننده است و برای حل آن چه روشی پیشنهاد می‌دهید؟

۳. در مورد روش‌های زیر در این مسئله توضیح دهید و برای هر کدام مزیت و ریسکش را بیان کنید.

Undersampling، SMOTE (oversampling)، Focal Loss

## پرسش‌های عملی (۹۰ نمره)

مهلت ارسال: ۱۴۰۴/۰۹/۰۷ ساعت ۲۳:۵۹:۵۹، با تاخیر: ۱۴۰۴/۰۹/۱۰ ساعت ۲۳:۵۹:۵۹

**پرسش ۱** (۴۰ نمره)

### بلای آوای خاموش الدوریا

**روایت ماجرا**

آگاه باش، ای کاتب سلطنتی! پادشاهی به ذهن تیزبین تو نیازمند است.

سفر پیمان "سنگ خورشید" با فاجعه‌ای هولناک روبرو شده است. یک کاروان بزرگ از رهروان، جستجوگران حقیقت و مریدان معنوی، که از اقصی نقاط پادشاهی الدوریا گرد آمده بودند، سفر خود را به سوی برج افلاکی "آستراس" آغاز کرد. آن‌ها آرمان‌های یک نسل را با خود حمل می‌کردند.

اما هنگامی که کاروان از میان جنگل‌های باستانی و مه‌آلود "شادوفن" عبور می‌کرد، نیرویی ناشناخته و هولناک در دل جنگل بیدار شد. بدون هیچ هشدار، بدون هیچ صدا یا اثری، تقریباً نیمی از رهروان به سادگی... ناپدید شدند.

پادشاه تو را، که نگهبان دفاتر سلطنتی هستی، مأمور مأموریتی حیاتی کرده است. وظیفه تو این است که با استفاده از اطلاعاتی که از دفتر کل سفر و دیگر اسناد سلطنتی در اختیار داری، الگوی پنهان در پس این بلا را کشف کنی. این بلا چه کسانی را با خود برد؟ آیا تصادفی بود یا منطقی در پس این جادوی تاریک نهفته است؟

بینش تو تنها امید پادشاهی برای ساختن یک طلسم محافظ است تا از تکرار این واقعه برای همیشه جلوگیری کند.

**توضیحات مجموعه داده**

وظیفه شما پیش‌بینی این است که آیا یک رهرو در طی بلای آوای خاموش ناپدید شده است یا خیر. در این مسیر، مجموعه‌ای از سوابق مسافران و ویژگی‌هایشان که از دفتر کل بازیابی شده است، در اختیارتان قرار می‌گیرد.

**توضیحات فایل‌ها و ستون‌ها**

**train.csv** - سوابق مربوط به حدود دو سوم (تقریباً ۸۷۰۰ نفر) از رهروان که باید به عنوان داده‌های آموزشی استفاده شود.

**PilgrimId** - یک شناسه منحصر به فرد برای هر رهرو.

**HomeRealm** - قلمرویی که رهرو سفر خود را از آنجا آغاز کرده است.

**MeditativeTrance** - نشان می‌دهد که آیا رهرو برای تمرکز افکار و اتصال به نیروی درونی خود، سفر را در حالت خلسه عمیق جادویی انجام داده است یا خیر.

**Wagon** - شماره ارابه‌ای که رهرو در آن اقامت دارد.

**DestinationSanctuary** - حریم والایی که رهرو قصد بازدید از آن را در برج افلاکی داشته است.

**Age** - سن رهرو.

**NobleBlood** - مشخص می‌کند که آیا رهرو از یک خاندان اشرافی شناخته‌شده است یا خیر.

**TavernBill**، **MarketSpend**، **ArtisanGuilds**، **HealersHut**، **DivinationDen** - مقادیر مربوط به مقدار طلا یا سکه‌ای که رهرو در میخانه‌ها، بازارها، غرفه‌های هنرمندان، کلبه شفاگر و کاشانه پیشگویی خرج کرده است.

**Name** - نام و نام خانوادگی رهرو.

**Vanished** - (ستون هدف) آیا رهرو به دلیل بلای آوای خاموش ناپدید شده است یا خیر.

**test.csv** - سوابق مربوط به یک سوم باقی‌مانده (تقریباً ۴۳۰۰ نفر) از رهروان که برای آزمون مدل استفاده می‌شود. این فایل شامل همان ستون‌های ورودی train.csv است، به جز ستون Vanished. وظیفه شما پیش‌بینی مقدار Vanished برای این افراد است.

**sample_submission.csv** - یک فایل نمونه برای ارسال پاسخ در قالب صحیح.

**PilgrimId** - شناسه هر رهرو در مجموعه داده آزمون.

**Vanished** - ستون هدف. برای هر رهرو، یکی از مقادیر True یا False را پیش‌بینی کنید.

**پرسش ۲** (۵۰ نمره)

## تشخیص احساسات از روی توییت

در این بخش، شما وظیفه تشخیص احساسات را با استفاده از مجموعه‌ای از توییت‌های کوتاه بر عهده دارید. این توییت‌ها حاوی متونی هستند که شیوه بیان حالات احساسی مختلف کاربران را نشان می‌دهند. برای هر توییت، برچسبی مشخص شده است که نشان‌دهنده نوع احساس غالب در آن متن است. برچسب‌ها ممکن است شامل دسته‌بندی‌هایی مانند: شادی، غم، خشم، ترس و سایر موارد.

از شما انتظار می‌رود که پاک‌سازی داده‌ها و مهندسی ویژگی را بر روی این مجموعه داده انجام دهید و سپس مدل‌سازی خود را در قالب یک نوت‌بوک (Jupyter) یا هر محیط تحلیلی مکمل پیاده‌سازی کنید. در طول فرآیند، باید مراحل کاری، پیش‌فرض‌ها، و تصمیمات مدل‌سازی خود را در نوت‌بوک جوپیتر یا یک سند جداگانه تشریح کنید.

برای ارزیابی نهایی، توصیه می‌شود از هر مدل کلاسیک که برای دستیابی به دقت بالا مناسب می‌دانید (مانند Logistic Regression، SVM، Random Forest و غیره) استفاده نمایید (استفاده از مدل‌های یادگیری عمیق زمینی یا شبکه‌های عصبی کم‌عمق نیز قابل قبول است، اما شبکه‌های عصبی عمیق مجاز نیست).

شما باید مدل خود را با استفاده از معیارهای مناسب بر روی یک مجموعه اعتبارسنجی ارزیابی کرده و از روش‌هایی مثل تفکیک داده‌ها به آموزش/اعتبارسنجی یا اعتبارسنجی متقابل K ـ لایه طبقه‌بندی‌شده استفاده کنید.

شما باید یک پایپ‌لاین جامع بسازید که شامل مراحل پیش‌پردازش داده، تبدیل متن به ویژگی‌های عددی (مثلاً با استفاده از Bag-of-Words، TF-IDF یا تکنیک‌های تعبیه‌سازی سبک)، و اعمال مدل طبقه‌بندی باشد. این پایپ‌لاین باید به وضوح در نوت‌بوک شما تشریح شود.

یک مجموعه داده آزمون بدون برچسب‌های واقعی در اختیار شما قرار می‌گیرد. پس از آموزش مدل بر روی داده‌های آموزشی، باید روی مجموعه داده آزمون، همان مراحل پیش‌پردازش و استخراج ویژگی‌ها را اعمال کرده و با استفاده از مدلی که بر روی داده‌های آموزشی خام آموزش داده‌اید، پیش‌بینی‌ها را انجام دهید.

لطفاً توجه داشته باشید که کسب دقت بالاتر از ۶۵٪ بر روی مجموعه داده اعتبارسنجی، نشان‌دهنده عملکرد قابل قبول مدل خواهد بود. با این حال، هدف اصلی این تمرین، درک و به‌کارگیری مراحل کامل یک پروژه یادگیری ماشین متنی است و تمرکز تنها بر روی بیشینه کردن دقت نهایی نخواهد بود. بدین معنا که مستندسازی فرآیند، تحلیل خطاها، و ارائه دلایل منطقی برای انتخاب روش‌ها و مدل‌ها، بخش مهمی از نمره نهایی را به همراه خواهد داشت و نمره‌دهی به صورت رقابتی نیست.