

دانشگاه تهران – دانشکده مهندسی برق و کامپیوتر راهنمای حل تشریحی کامل (نسخه دستیار آموزشی)

درس علم داده – بهار ۱۴۰۴

ویرایش ۰۲

هدف این سند: یکدست‌سازی نمره‌دهی، تعریف معیارهای علمی، و ارائه پاسخ مرجع برای تمام سوالات تمرین نهایی توسعه‌یافته.

فلسفه نمره‌دهی

- صحت علمی: فرمول، استدلال و تفسیر باید دقیق باشند.
- کیفیت مهندسی: کد ماژولار، قابل اجرا، و بازتوانیدن باشد.
- شفافیت: مفروضات، محدودیت‌ها و ریسک‌ها صریح بیان شوند.
- مسئولیت‌پذیری: تحلیل عدالت و سیاست مداخله انسانی الزامی است.

Q1) چرخه عمر علم داده و صورت‌بندی مسئله

حد انتظار پاسخ عالی:

۱. تعریف روش مسئله: «پیش‌بینی احتمال مهاجرت برای پشتیبانی تصمیم‌گیر».
۲. تعریف متريک‌ها: ROC-AUC برای رتبه‌بندی، Recall@K برای یافتن موارد حساس، و کالیبراسيون برای قابلیت اتكا.
۳. چرخه کامل: framing -> data -> validation -> modeling -> deployment -> monitoring
۴. ثبت ریسک‌ها: نشت داده، concept drift، تعییر سیاست، و کیفیت برچسب.

خطاهای متداول:

- ادعای علیّ از داده مشاهده‌ای بدون قیود.
- نبود برنامه پایش پس از استقرار.

Q2) عملیات داده و EDA

پاسخ مرجع:

- حسابرسی .dtype/null/duplicate/range
- حداقل ۶ تا ۸ نمودار معنادار با تفسیر تصمیم‌محور.
- پیاده‌سازیتابع پیش‌پردازش تکرار‌پذیر (ترجیحاً با آزمون واحد).

حداقل نمودارهای پیشنهادی:

۱. توزیع متغیر هدف.

۲. توزیع ویژگی‌های کلیدی (GitHub_Activity, Research_Citations)

۳. همبستگی ویژگی‌های عددی.

۴. نرخ مهاجرت به تقییک کشور.

۵. اثر Education_Level بر هدف.

۶. نمودار پرتهای برای ۲ ویژگی حساس.

استنباط آماری Q³

الگوی پاسخ معتبر:

• تعریف فرض صفر/مقابل.

• انتخاب آزمون مناسب با نوع داده (مثلًا آزمون دو نمونه‌ای یا chi-square).

• تفسیر صحیح p-value: احتمال مشاهده داده (یا شدیدتر) تحت درست بودن H_0 .

• تفسیر بازه اطمینان: بازه‌ای از مقادیر سازگار با داده در سطح اطمینان مشخص.

نکته نمره‌دهی: اگر داشجو p-value را احتمال درست بودن H_0 تفسیر کند، کسر نمره قابل توجه اعمال شود.

طراحی بصری و روایت داده Q⁴

شاخص‌های پاسخ قوی:

• KPI‌ها به تصمیم واقعی متصل باشند (مثلًا نرخ موفقیت دعوت به برنامه مهاجرت).

• استفاده درست از ادراک بصری (area/color position/length) بهتر از

• نمایش حداقل یک خطای طراحی (مثل قطع محور (y و نسخه اصلاح شده.

پیشرفته SQL Q⁵

الگوی مرجع میانگین متحرک و رتبه‌بندی:

```
( AS citation_velocity WITH
,Research_Citations ,Year ,Country_Origin ,UserID SELECT
( OVER AVG(Research_Citations)
Country_Origin BY PARTITION
Year BY ORDER
ROW CURRENT AND PRECEDING 2 BETWEEN ROWS
moving_avg_citations AS )
Professionals_Data FROM
)
```

```
( OVER DENSE_RANK() ,* SELECT
DESC moving_avg_citations BY ORDER Country_Origin BY PARTITION
country_rank AS )
citation_velocity; FROM
```

معیارهای نمره‌دهی:

- استفاده صحیح از PARTITION BY/ORDER BY/window frame
- رعایت منطق زمانی و جلوگیری از نشت آینده.
- کیفیت کوئری cohort با CTE

Q6) نشت داده و معماری کلانداده

تشخیص نشت:

- نشت مستقیم (پسار و بدادی). Visa_Approval_Date
- نشت زمانی بالقوه. Last_Login_Region
- پراکسی زمانی بالقوه. Passport_Renewal_Status
- در صورت محاسبه نقطه-نقطه زمانی قابل قبول. Years_Since_Degree

معماری مورد قبول:

- لایه‌های Bronze/Silver/Gold
- point-in-time join با feature store
- همخوانی offline/online features
- پایش drift و صحت داده ورودی.

Q7) Elastic Net و تفسیر آماری

تابع هزینه:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda_1 \sum_{j=1}^n |\theta_j| + \frac{\lambda_2}{2} \sum_{j=1}^n \theta_j^2$$

مشتق مختصه‌ای:

$$\nabla_{\theta_j} J = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} + \lambda_1 \partial |\theta_j| + \lambda_2 \theta_j$$

زیرگرادیان $|\theta_j|$:

$$\partial |\theta_j| = \begin{cases} 1 & \theta_j > 0 \\ -1 & \theta_j < 0 \\ [-1, 1] & \theta_j = 0 \end{cases}$$

تفسیر آماری نمونه: اگر $p-value=0.003$ و بازه اطمینان $[0.18, 0.86]$ باشد:

- فرض صفر $H_0 : \beta = 0$ رد می‌شود.
- چون صفر داخل بازه نیست، اثر مثبت معنی‌دار است.

Q⁸) بهینه‌سازی: SGD در برابر Momentum و Adam

پدیده ravine: خمیدگی زیاد در یک محور و کم در محور دیگر باعث نوسان SGD می‌شود.
Momentum:

$$v_t = \beta v_{t-1} + \eta g_t, \quad \theta_{t+1} = \theta_t - v_t$$

میانگین‌گیری زمانی از گرادیان، نوسان‌های عالمت‌عرضکن را می‌کاهد و حرکت در جهت پایدار را تقویت می‌کند.

Adam:

- ممان اول و دوم گرادیان را نگه می‌دارد.
- نرخ یادگیری موثر را به صورت پارامتری تنظیم می‌کند.
- در ناهمگنی مقیاس ویژگی‌ها معمولاً پایدارتر است.

Q⁹) مدل‌های غیرخطی و تنظیم پیچیدگی

الف) SVM-RBF

اگر مدل بیشبرازش دارد، γ باید کاهش یابد تا حوزه اثر هر نقطه آموزشی بزرگ‌تر شده و مرز تصمیم هموارتر شود.

ب) هرس درخت

$$R_\alpha(T) = R(T) + \alpha|T|$$

با افزایش α ، جریمه پیچیدگی بزرگ‌تر شده و درخت کوچک‌تر می‌شود (بایاس بیشتر، واریانس کمتر).

Q¹⁰) کاهش بعد و PCA

برای مقادیر ویژه $\lambda_1, \lambda_2, \lambda_3$:

$$\text{EVR}_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \lambda_3}$$

نسبت واریانس دو مؤلفه اول:

$$\text{EVR}_{1,2} = \frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i}$$

تفسیر: مقدار ویژه، واریانس توضیح‌داده شده توسط آن مؤلفه اصلی است.

Q¹¹) خوشبندی و Elbow

دلیل هندسی/تحلیلی:

- WCSS با افزایش K همواره کاهش می‌یابد.
- نرخ کاهش WCSS بهتریج کم می‌شود (بازده نزولی).
- نقطه آرنج، جایی است که افزایش K بهبود معنادار نسبت به هزینه پیچیدگی ندارد.

Q¹²) شبکه‌های عصبی و مدل توالی

پاسخ کامل باید شامل:

- معماری،تابع زیان،optimizer و برنامه آموزش.
- راهکار کنترل بیشبرازش (dropout/early stopping/weight decay).
- مقایسه با baseline کلاسیک و تحلیل خطای.

Q۱۳) مدل زبانی و LLM Agent

طرح پیشنهادی:

- Plan: شکستن مسئله به گام‌های قابل کنترل.

- Retrieve: بازیابی شواهد معتبر.

- Reason: استدلال کنترل شده بر اساس شواهد.

- Verify: راستی‌آزمایی خروجی و ثبت عدم‌قطعیت.

شاخص‌های ارزیابی:

- Faithfulness

- نرخ Hallucination

- اینمی محتوایی و انطباق سیاستی

Q۱۴) عدالت، سوگیری و حاکمیت

انتظارات اصلی:

- گزارش متريک زيرگروهي (کشور/تحصيلات/...).

- تحليل خطر proxy و بازتوليد سوگيري تاریخي.

- طراحی مسیر human override و بازبینی پرونده.

Q۱۵) کالibrاسيون و سیاست آستانه

موارد لازم:

- منحنی کالibrاسیون و تفسیر شکاف با خط ایده‌آل.

- يك يا دو معیار کالibrاسیون (مثلًا ECE و Brier).

- دو آستانه:

۱. آستانه بهینه برای بیشینه کردن F1.

۲. آستانه بهینه بر اساس هزینه نامتقارن خطا (هزینه FN < FP).

- توصیه نهایی باید بر مبنای هزینه/کاربرد باشد، نه پیشفرض ۰,۵.

Q۱۶) تشخيص درفت و پایش

موارد لازم:

- تقسیم داده به پنجره مرجع/جاری (ترجیحاً زمانی؛ در نبود زمان، تقسیم تصادفی).

- محاسبه و رتبه‌بندی PSI برای ویژگی‌های عددی؛ سیاست تفسیر:

– $100 < \text{PSI}$: پایین

– $100 - 250$: متوسط

– ≥ 250 : بالا (بررسی و اقدام)

- يك شاخص درفت برای دسته‌ای‌ها (مثلًا JS divergence توزیع کشور).

- پیشنهاد SOP پایش: آستانه هشدار/حرانی و حرک بازآموزی.

Q17) ریکورس مقابله‌ای

موارد لازم:

- تعریف ویژگی‌های قابل مداخله (مثلًا Industry_Experience، Research_Citations، GitHub_Activity)
- جستجوی کمینه تغییر برای عبور از آستانه تصمیم، با سقف‌های واقع‌گرایانه.
- گزارش نرخ موفقیت ریکورس، میانه تغییر برای هر ویژگی، و بحث امکان/اخلاق.

SHAP تبیین‌پذیری Capstone)

مفاهیم کلیدی:

- مقدار پایه مدل (میانگین خروجی در داده مرجع). base_value
- خروجی نهایی برای نمونه خاص. output_value
- مجموع مقادیر SHAP برابر اختلاف این دو مقدار است.

تفسیر سناریوی نمونه: اگر داولطلب با Research_Citations بالا پیش‌بینی «عدم مهاجرت» بگیرد، باید نشان داده شود کدام ویژگی‌های منفی (مثلًا الگوهای تجربه/کشور/سایر عوامل) اثر مثبت استنادات را خنثی کردند.

جدول راهنمای سریع کسر/اعطای نمره

وضعیت پاسخ	سیاست نمره‌دهی
فرمول درست + تفسیر دقیق + نمره کامل پیاده‌سازی بازنگردانی	پیاده‌سازی بازنگردانی
فرمول درست ولی تفسیر آماری ناقص کسر ۲۰ تا ۴۰ درصد همان سوال کسر سنگین (تا ۵۰ درصد بلوك مرتبه)	عدم کنترل نشت داده
کسر ۳۰ درصد کسر کامل بلوك اخلاق	نتیجه خوب بدون شواهد اجرای کد تحلیل عدالت ناقص/حذف شده

توصیه به تصحیح‌کنندگان: اگر دانشجو فرضیات را شفاف کرده، محدودیت‌هارا صادقانه گزارش داده و مسیر مهندسی قابل بازنگردانی کرده است، حتی با متريک متوسط نيز پاسخ باکیفیت محسوب می‌شود.