

دانشگاه تهران – دانشکده مهندسی برق و کامپیوتر

راهنمای حل تشریحی کامل (نسخه دستیار آموزشی)

درس علم داده – بهار ۱۴۰۴

ویرایش ۰.۲

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 بازتولیدپذیر باشد.
- **شفافیت:** مفروضات، محدودیت‌ها و ریسک‌ها صریح بیان شوند.
- **مسئولیت‌پذیری:** تحلیل عدالت و سیاست مداخله انسانی الزامی است.

Q۱) چرخه عمر علم داده و صورت‌بندی مسئله

حد انتظار پاسخ عالی:

۱. تعریف روشن مسئله: «پیش‌بینی احتمال مهاجرت برای پشتیبانی تصمیم‌یار».
۲. تعریف متریک‌ها: ROC-AUC برای رتبه‌بندی، Recall@K برای یافتن موارد حساس، و کالبراسیون برای قابلیت اتکا.
۳. چرخه کامل: framing -> data -> validation -> modeling -> deployment -> monitoring.
۴. ثبت ریسک‌ها: نشت داده، concept drift، تغییر سیاست، و کیفیت برچسب.

خطاهای متداول:

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

Q۲) عملیات داده و EDA

پاسخ مرجع:

- حسابرسی dtype/null/duplicate/range.
- حداقل ۶ تا ۸ نمودار معنادار با تفسیر تصمیم‌محور.
- پیاده‌سازی تابع 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 تفکیک کشور.
۵. اثر Education_Level بر هدف.
۶. نمودار پرت‌ها برای ۲ ویژگی حساس.

Q۳) استنباط آماری

الگوی پاسخ معتبر:

- تعریف فرض صفر/مقابل.
 - انتخاب آزمون متناسب با نوع داده (مثلاً آزمون دو نمونه‌ای یا chi-square).
 - تفسیر صحیح p-value: احتمال مشاهده داده (یا شدیدتر) تحت درست‌بودن H0.
 - تفسیر بازه اطمینان: بازه‌ای از مقادیر سازگار با داده در سطح اطمینان مشخص.
- نکته نمره‌دهی: اگر دانشجو p-value را احتمال درست‌بودن H0 تفسیر کند، کسر نمره قابل توجه اعمال شود.

Q۴) طراحی بصری و روایت داده

شاخص‌های پاسخ قوی:

- KPIها به تصمیم واقعی متصل باشند (مثلاً نرخ موفقیت دعوت به برنامه مهاجرت).
- استفاده درست از ادراک بصری (position/length بهتر از area/color).
- نمایش حداقل یک خطای طراحی (مثل قطع محور y) و نسخه اصلاح‌شده.

Q۵) SQL پیشرفته

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.

Q۶) نشت داده و معماری کلان داده

تشخیص نشت:

- Visa_Approval_Date: نشت مستقیم (پسارویدادی).
- Last_Login_Region: نشت زمانی بالقوه.
- Passport_Renewal_Status: پراکسی زمانی بالقوه.
- Years_Since_Degree: در صورت محاسبه نقطه به نقطه زمانی قابل قبول.

معماری مورد قبول:

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

Q۷) 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 $p\text{-value}=0.003$ 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.

Q۸) بهینه‌سازی: SGD در برابر Adam و Momentum

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 علامت‌عوض‌کن را می‌کاهد و حرکت در جهت پایدار را تقویت می‌کند.

Adam:

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

Q۹) 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).

Q۱۰) کاهش بُعد 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 مولفه اول:

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

تفسیر: مقدار ویژه، واریانس توضیح‌داده‌شده توسط آن مولفه اصلی است.

Q۱۱) خوشه‌بندی و Elbow

دلیل هندسی/تحلیلی:

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

Q۱۲) 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 خطا.

Q۱۳) مدل زبانی و LLM Agent

طرح پیشنهادی:

- Plan: شکستن مسئله به گام‌های قابل کنترل.
- Retrieve: بازیابی شواهد معتبر.
- Reason: استدلال کنترل‌شده بر اساس شواهد.
- Verify: راستی‌آزمایی خروجی و ثبت عدم قطعیت.

شاخص‌های ارزیابی:

- Faithfulness
- نرخ Hallucination
- ایمنی محتوایی و انطباق سیاستی

Q۱۴) عدالت، سوگیری و حاکمیت

انتظارات اصلی:

- گزارش متریک زیرگروهی (کشور/تحصیلات/...).
- تحلیل خطر proxy و بازتولید سوگیری تاریخی.
- طراحی مسیر human override و بازبینی پرونده.

Q۱۵) کالیبراسیون و سیاست آستانه

موارد لازم:

- منحنی کالیبراسیون و تفسیر شکاف با خط ایده‌آل.
- یک یا دو معیار کالیبراسیون (مثلاً Brier و ECE).
- دو آستانه:
- ۱. آستانه بهینه برای بیشینه کردن F1.
- ۲. آستانه بهینه بر اساس هزینه نامتقارن خطا (هزینه $FP < FN$).
- توصیه نهایی باید بر مبنای هزینه/کاربرد باشد، نه پیش‌فرض ۰,۵.

Q۱۶) تشخیص 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:
- $PSI < ۱۰.۰$: پایین
- $۱۰.۰ - ۲۵.۰$: متوسط
- $۲۵.۰ \geq$: بالا (بررسی و اقدام)
- یک شاخص درفت برای دسته‌ای‌ها (مثلاً JS divergence توزیع کشور).
- پیشنهاد SOP پایش: آستانه هشدار/بحرانی و محرک بازآموزی.

Q۱۷) ریکورس مقابله‌ای

موارد لازم:

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

Q۱۸) اعتبارسنجی زمانی و افت عملکرد

موارد لازم:

- طراحی rolling backtest زمانی؛ در نبود ستون زمانی، fallback باید شفاف و مستند باشد.
 - گزارش متریک‌های هر fold (حداقل AUC و F1) و مقدار افت نسبت به fold اول.
 - تحلیل رابطه افت عملکرد با شاخص درفت (مانند میانگین PSI).
 - خروجی‌های الزامی: q18_temporal_backtest.csv و q18_temporal_degradation.png.
- نکته تصحیح: اگر fallback زمانی استفاده شده باشد ولی در گزارش صریح توضیح داده نشود، کسر نمره اعمال شود.

Q۱۹) کمی‌سازی عدم قطعیت

موارد لازم:

- پیاده‌سازی روش conformal (یا معادل معتبر) برای بازه/نمره اطمینان.
 - گزارش پوشش تجربی در چند سطح اطمینان (مثلاً ۸۰/۹۰/۹۵ درصد).
 - تحلیل پهنای بازه و ریسک کمپوششی.
 - خروجی‌های الزامی: q19_uncertainty_coverage.csv و q19_coverage_vs_alpha.png.
- نکته تصحیح: گزارش سطح اطمینان بدون سنجش پوشش تجربی کافی نیست.

Q۲۰) آزمایش مداخله عدالت الگوریتمی

موارد لازم:

- خط پایه عدالت زیرگروهی (حداقل demographic parity gap یا equal opportunity gap).
 - اجرای یک مداخله مشخص (مثلاً reweighing) و مقایسه قبل/بعد.
 - تحلیل مصالحه عدالت-عملکرد و بررسی قید سیاستی (مثلاً سقف افت AUC/F1).
 - خروجی‌های الزامی: q20_fairness_mitigation_comparison.csv و q20_fairness_tradeoff.png.
- نکته تصحیح: اگر دانشجو بهبود عدالت را گزارش کند اما قید سیاستی/کسب‌وکاری نداشته باشد، نمره کامل داده نشود.

Capstone تبیین‌پذیری SHAP

مفاهیم کلیدی:

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

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

جدول راهنمای سریع کسر/اعطای نمره

| وضعیت پاسخ | سیاست نمره‌دهی |
|-------------------------------------|-----------------------------------|
| فرمول درست + تفسیر دقیق + نمره کامل | پیاپی‌سازی بازتولیدپذیر |
| فرمول درست ولی تفسیر آماری ناقص | کسر ۲۰ تا ۴۰ درصد همان سوال |
| عدم کنترل نشئت داده | کسر سنگین (تا ۵۰ درصد بلوک مرتبط) |
| نتیجه خوب بدون شواهد اجرای کد | کسر ۳۰ درصد |
| تحلیل عدالت ناقص/حذف‌شده | کسر کامل بلوک اخلاق |

توصیه به تصحیح‌کنندگان: اگر دانشجو فرضیات را شفاف کرده، محدودیت‌ها را صادقانه گزارش داده و مسیر مهندسی قابل بازتولید ارائه کرده است، حتی با متریک متوسط نیز پاسخ باکیفیت محسوب می‌شود.