تمرین Cross-Validation

محمد صالح علی اکبری

تاریخ تحویل: ۲ خرداد

سؤال ۱

جمع‌بندی و نکات مهم در انتخاب روش Cross-Validation را تشریح کنید.

پاسخ

انتخاب روش Cross-Validation بستگی به ویژگی‌های داده و ماهیت مسئله دارد. نکات مهم در زیر فهرست شده‌اند:

- CV K-Fold: روش عمومی برای مسایل رگرسیون و طبقه‌بندی است که داده را به K بخش مساوی تقسیم می‌کند و در هر دور یک بخش برای آزمون و باقی برای آموزش استفاده می‌شود. این روش واریانس برآورد خطا را کاهش می‌دهد اما با افزایش K هزینه محاسباتی بیشتر می‌شود. -contentRefer: ence[oaicite:۰]index=۰
- CV Leave-One-Out (LOOCV): حالت خاص K-Fold که در آن K برابر تعداد نمونه‌هاست. در هر مرحله یک نمونه برای آزمون و مابقی برای آموزش استفاده می‌شود. مناسب برای داده‌های بسیار کوچک اما هزینه محاسباتی بالایی دارد. :contentReference[oaicite:۱]index=۱
- CV K-Fold Stratified: نسخه‌ای از K-Fold برای مسائل طبقه‌بندی با توزیع نامتوازن کلاس‌ها که نسبت نمونه‌های هر کلاس در هر فولد حفظ می‌شود. موجب پایداری بیشتر در ارزیابی دقت می‌شود. :contentReference[oaicite:۲]index=۲
- Split Series Time: ویژه سری‌های زمانی است که در آن ترتیب داده‌ها باید حفظ شود و از داده‌های آینده در آموزش استفاده نمی‌شود. در هر مرحله از آخرین نقاط گذشته برای آموزش و دوره‌ای از داده‌های بعدی برای آزمون بهره می‌برد. :contentReference[oaicite:۳]index=۳

- K-Fold Group: برای داده‌هایی با ساختار گروهی (مانند مطالعات پزشکی) که تضمین می‌کند نمونه‌های یک گروه همگی در آموزش یا آزمون قرار گیرند تا از نشت اطلاعات جلوگیری شود. -con: ۴]index=۴tentReference[oaicite:

برای مقایسه روش‌ها معمولاً از معیارهایی مانند RMSE، Accuracy Score و نظایر آن استفاده می‌شود و انتخاب عدد مناسب K بر اساس توازن بین دقت تخمین و هزینه‌ی محاسباتی صورت می‌گیرد.

سؤال ۲

فقط روش‌های عمومی (برای رگرسیون و طبقه‌بندی) را نام برده و دو روش با انتخاب خودتان را تشریح کنید.

پاسخ

فهرست روش‌های عمومی:

- Holdout (یا Hold-Out Validation)
- Cross-Validation K-Fold
- CV Leave-One-Out (LOOCV)
- CV K-Fold Stratified
- CV K-Fold Repeated
- CV Shuffle-Split
- Split Series Time
- CV K-Fold Group

۱. Cross-Validation: K-Fold در این روش، مجموعه‌داده به K بخش (فولد) مساوی تقسیم می‌شود. در هر تکرار، یک فولد به عنوان داده آزمون و $K - ۱$ فولد باقی‌مانده به عنوان داده آموزش استفاده می‌شوند. این فرایند K بار تکرار شده و نتایج ارزیابی میانگین‌گیری می‌شود.

- مزایا: استفاده بهینه از داده (تمام نمونه‌ها هم در آموزش و هم در آزمون مشارکت دارند)، کاهش واریانس در برآورد خطا.
- معایب: هزینه محاسباتی افزایش می‌یابد، مخصوصاً برای K بزرگ.
- موارد کاربرد: مسائل رگرسیونی و طبقه‌بندی عمومی که وابستگی زمانی یا گروه‌بندی داده وجود ندارد.

۲. CV: K-Fold Stratified نسخه بهبودیافته‌ای از K-Fold برای مسائل طبقه‌بندی با داده‌های نامتوازن است. در این روش، نسبت نمونه‌های هر کلاس در هر فولد مشابه نسبت کل داده حفظ می‌شود.

- مزایا: جلوگیری از ایجاد فولدهای با پراکندگی نامناسب کلاس‌ها، بهبود ثبات و دقت ارزیابی در مسائل طبقه‌بندی.
- معایب: پیچیدگی اندکی بیشتر در پیاده‌سازی، فقط برای مسائل طبقه‌بندی کاربرد دارد.
- موارد کاربرد: طبقه‌بندی دودویی یا چندکلاسه با توزیع نامتوازن کلاس‌ها.