دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

داده کاوی

پاسخنامه تمرین سری ۴

بهار ۱۴۰۳

فهرست مطالب

۱ پاسخ سوالات

۱.۱ پاسخ سوال ۱

۱. حجم مجموعه داده

در کل بیش‌برازش روی مجموعه داده‌های کوچک محتمل تر است چون به عنوان راهی برای کاهش خطا، امکان به خاطرسپاری داده‌های ورودی، به جای استخراج الگوهای معنادار و کلی وجود دارد.

۲. تعادل داده‌ها (داده‌های بالانس و غیربالانس)

با داده‌های آنبالانس احتمال بیش‌برازش بالا می‌رود. چون عمده‌ی داده‌ها از یک گروه هستند، مدل بدون یادگرفتن الگوهای صحیح عملکرد خوبی روی داده‌های آموزشی نشان میدهد اما آنچه آموخته است قابل توسعه به داده‌های دیده نشده نیست.

۳. ویژگی‌های نامربوط

ویژگی‌های نامربوط می توانند باعث افزایش احتمال بیش‌برازش بشوند. ممکن است مدل در هنگام آموزش به جای یادگیری الگوهای معنادار، بیشتر از مقادیر این ویژگی‌های نامربوط که اطلاعات مفید برای تسک در دست ندارند، استفاده بکند و الگوهایی را یاد بگیرد که قابلیت توسعه یافتن به داده‌های دیده نشده را نخواهند داشت. در این صورت روی داده‌های تست مطلوب عمل نخواهد کرد.

۴. تعداد ایپاک‌های آموزش

افزایش تعداد ایپاک ها می‌تواند به بیش‌برازش بیانجامد. این اتفاق در حالتی می‌افتد که ظرفیت مدل نسبت به پیچیدگی داده‌ها زیاد باشد یا داده‌ها نویز زیادی داشته باشند. در این صورت مدل شروع می‌کند به یادگیری جزئیات غیرقابل توسعه از داده‌های آموزشی(مانند نویز). برای پیشگیری از چنین اتفاقی می توان دقت روی داده‌های validation را زیر نظر داشت و از نقطه‌ای که روند بهبود آن تمام می‌شود و شروع به افت می‌کند آموزش را قطع کرد (به این کار early stopping می‌گویند)

۵. پیچیدگی مدل

هر چقدر مدل پیچیده‌تر می‌شود، امکان استخراج نشدن الگوهای معنادار و کلی توسط مدل و به خاطرسپاری داده‌های ورودی بالا می‌رود و این احتمال بیش برازش را افزایش می‌دهد.

۶. نشت داده‌ها (Data Leakage)

نشت داده‌ها باعث می‌شود اطلاعاتی در زمان آموزش در دسترس مدل قرار بگیرد که نباید در دسترس مدل باشد (چون پس از آموزش، برای داده‌های دیده نشده این اطلاعات وجود ندارند). این مسئله می تواند باعث شود بیش‌برازش نهفته باقی بماند چون اگر بیش‌برازش رخ داده باشد (مدل الگوهای صحیحی را نیاموخته باشد) بررسی دقت روی داده های آزمون این را آشکار نخواهد کرد (در این حالت دقت روی داده‌های آزمون هم بالا خواهد بود با این که الگوهای قابل توسعه آموخته نشده‌اند).

۲.۱ پاسخ سوال ۲

0.25

predict/actual	1	0
1	3	8
0	0	9

- err = $\frac{8}{20}$ = 0.4

0.5

predict/actual	1	0
1	3	2
0	0	15

- err = $\frac{2}{20}$ = 0.1

0.75

predict/actual	1	0
1	2	0
0	1	18

- err = $\frac{1}{20}$ = 0.05

۳.۱ پاسخ سوال ۴

i. Construct the classification matrix

Classification Matrix		
	Predicated Class	
Actual Class	Fraudulent = 1	Non Fraudulent = 0
Fraudulent = 1	30	32
Non Fraudulent = 0	58	920

ii. Calculate the error rate (overall and error rate of each class)

Error Rate:

Fraudulent Class = 32/62 = 0.52

Non Fraudulent = 58/978 = 0.06

Overall = 90 / 1040 = 0.0865

۴.۱ پاسخ سوال ۵

۱. برای داده های با نویز بالا، پس–هرس (post-pruning) بهتر از پیش–هرس (pre-pruning) می‌باشد. **نادرست**. پس–هرس روی داده‌های با نویز بالا بهتر از پیش–هرس عمل نمی‌کند. پیش–هرس از ابتدا و قبل از

این که درخت به یادگیری نویزها بپردازد جلوی رشد درخت را می‌گیرد و زودتر جلوی بیش‌برازش را در مجموعه داده‌های شدیداً نویزی میگیرد.

۲. برای داده های غیر بالانس، پیش-هرس ،(pre-pruning) بهتر از پس-هرس (post-pruning) می‌باشد

درست. درخت های تصمیم به سمت کلاس اکثریت بایاس می‌شوند و بخش عمده‌ی درخت صرف تلاش برای جدا کردن گروه های خالص کلاس اکثریت می‌شود (به جای یادگیری قوانین جامع برای جدا کردن دو کلاس از هم). اگر درخت خیلی عمیق شود، ممکن است بتواند جداسازی را انجام بدهد اما قوانینی غیرقابل توسعه به دست آورده است؛ بهتر است زود جلوی رشد درخت گرفته شود.

۳. وقتی مجموعه‌داده ها کوچک است احتمال بیش‌برازش بالا می‌رود و وقتی مجوعه داده بزرگ است احتمال کم‌برازش.

نادرست. کوچک بودن مجموعه داده‌ها ممکن است باعث بیش‌برازش بشود، ولی بزرگ بودن مجموعه داده باعث کم‌برازش نمی‌شود.

۴. درخت‌های تصمیم برای مجموعه داده‌هایی که فیچرها روابط غیرخطی با متغیر هدف دارند مناسب هستند.

نادرست. درخت‌های تصمیم می‌توانند برای مجموعه داده‌هایی که روابط غیر خطی پیچیده دارند چندان مناسب نباشند؛ با توجه به این که با برش‌های موازی با محور دسته بندی را انجام می‌دهند، برای حالاتی که مرزهایی پیچیده‌تری دارند چندان مطلوب عمل نمی‌کنند.

۵.۱ پاسخ سوال ۶

با بررسی اطلاعات کلی دیتاست (در لینک درج شده) متوجه می‌شویم مجموعه داده به شدت آنبالانس است (۰٫۱۷۲% داده‌ها متعلق به کلاس مثبت هستند). برای داده‌های نامتوازن، استفاده از stratified cross validation مناسب است که درصد یکسانی از هر کلاس در مجموعه ی train و validation قرار بگیرد.