به نام خدا

عنوان:

تکلیف یازدهم یادگیری ماشین

استاد:

دکتر پدرام

دانشجو:

محمدعلی مجتهدسلیمانی

4033904504

تاریخ:

9/10/1403

سوال 1

یادگیری تجمعی یک رویکردی است که با چندین مدل که اصطلاحا به آنها weak learner یا base model گفته میشود به نوعی با همدیگر ادغام شده اند برای حل یک مسئله خاص، استفاده از یادگیری تجمعی مزایای مختلفی دارد:

کاهش variance (کاهش احتمال بیش پردازش):

مدل های پیچیده به طور کلی مستعد overfitting هستند به این معنی که روی داده های آموزشی خیلی خوب هستند اما روی داده های از قبل دیده نشده ضعیف عمل میکنند.

روش های جمعی مثل bagging میانگین پیشبینی از مدل های مختلف را که روی بخش های متفاوتی از dataset آموزش دیده اند میگیرد. این فرآیند میانگین گیری سبب میشود که خطای هر مدل کمتر شود و باعث میشود به طور کلی واریانس کاهش پیدا کند و مدل تعمیم پذیر خواهد بود.

کاهش bias:

روش های جمعی مثل boosting سعی میکنند مدل ها را به صورت توالی قرار دهند تا هر مدل تمرکز کند تا خطای مدل قبلی را کاهش بدهد. این رویکرد باعث میشود که bias کاهش پیدا کند و accuracy بالا برود.

سوال 2

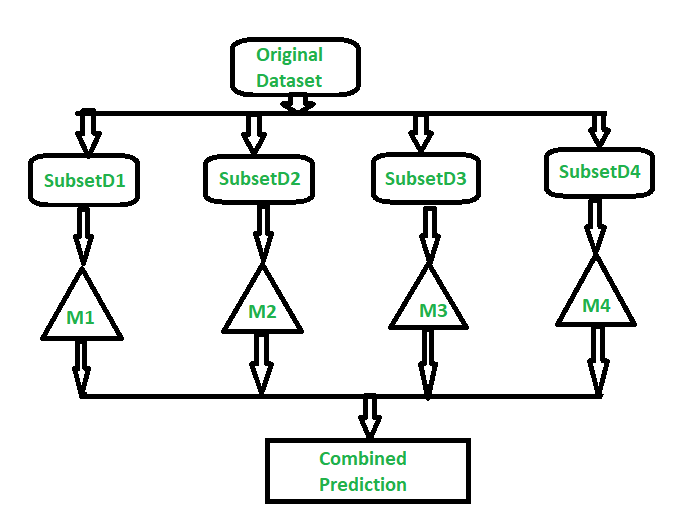
درخت های تصمیم به طور ذاتی دارای واریانس بالا و bias پایین هستند. یعنی میتوانند الگوهای پیچیده را به خوبی یاد بگیرند. اما به شدت به داده های آموزشی حساس هستند و ممکن است بیش از حد روی داده های آموزش بیش پردازش شوند و دچار overfit شویم و دارای واریانس بالا هستند. یعنی با یک تغییر کوچک در مجموعه آموزشی، ساختار درخت تغییر چشمگیری میکند.

در روش bagging ما با استفاده از روش bootstrap چندین مجموعه داده آموزشی میسازیم که با انتخاب و جایگزینی دوباره همراه است این روش. هر درخت روی یکی از این مجموعه ها آموزش میبیند.

به وسیله میانگین گیری از تخمین هر کدام از این درخت ها (یا استفاده از رای گیری برای کلاس بندی) این روش میتواند باعث کاهش واریانس شود. این رویکرد تجمعی باعث میشود که مدل کمتر به یک قسمت خاص از مجموعه داده آموزشی حساس شود.

نکته مهم دیگر این است که برای بیشینه کارآمدی درخت ها باید خطای هر درخت مستقل از بقیه درخت ها باشد تا بتوانیم تنوع خوبی روی داده های مختلف داشته باشیم برای پروسه میانگین گیری.

نکته مهم دیگری که باید اشاره بکنم این است که درخت های تصمیم به عنوان مدل های ناپایدار (unstable learners) شناخته میشوند. به این معنی که وقتی دارند روی مجموعه داده های مختلف آموزش میبینند به علت اینکه با یک تغییر کوچک ساختار درخت تغییر پیدا میکند، درخت ها به هم مشابه نیستند و ما تنوع داریم در مدل های خودمان و همه از یک جنس نخواهند بود.



سوال 3

وقتی روش bagging ما بهترین عملکرد ممکن را دارد که این شرایط را داشته باشیم: 1. مدل های base learner ما واریانس بالا داشته باشند به این معنی که حساس هستند به تنوع داده های آموزشی و مستعد بیش پردازش هستند. 2. مدل های ما دارای تنوع باشند به این معنی که خطای ایجاد شده توسط هر کدام کاملا مستقل از سایر درخت ها باشد. وقتی این وضعیت برقرار باشد میانگین گیری از پیشبینی به صورت موثر واریانس را کاهش میدهد و باعث افزایش دقت میشود.

پس تنوع بین مدل های پایه بسیار مهم است برای عملکرد خوب روش bagging حالا چگونه این تنوع را بدست بیاوریم؟ :

Bootstrap sampling: مکانیزم مرکزی که باعث تنوع در این روش میشود bootstrap است به این معنی که ما در این روش یک سری مجموعه های آموزشی میسازیم که هر کدام از دیتاست اصلی انتخاب میشوند قرار داده میشوند و بعد دوباره به دیتاست اصلی برمیگردند و در دور دوم دوباره امکان انتخاب آنها است، که با احتمال 63 درصد هر داده میتواند در زیر مجموعه های آموزشی حضور پیدا کنند. این مکانیزم باعث میشود ما اطمینان پیدا کنیم که برداشت های مختلفی از توزیع داده های مجموعه اصلی داشته باشیم. این باعث میشود که تنوع ایجاد شود مخصوصا برای مدل های ناپایدار (unstable) که به ازای کوچکترین تغییر در مجموعه آموزشی ساختار آنها تغییر میکند.