بسمه تعالی

زهرا آینه دست

تمرین اول ماشین لرنینگ

آکادمی یوتک

* سوال اول:

Polynomial feature بهترین رابطه تقریبی میان x و y را به ما میدهد

و همچنین مقدار عمیقی از انحنا ها را میتواند fit کند

اما

وجود تعداد کمی حتی یک و دو داده پرت خارج از محدوده میتواند به صورت جدی روی نتیجه آنالیز غیر خطی آن اثر بگذارد (که نتیجه آن کاهش دقت و افزایش بایاس و underfitting است)

در نتیجه برای توصیف داده ها خطی میتواند مناسب باشد چراکه داده های خطی کم واریانس بوده و عموما داده خیلی خارج از محدوده نخواهند داشت که این مشکل را ایجاد کند.

همچنین model validation tool های کمی برای تشخیص و detect داده های خارج از محدوده در مدل های nonlinear نسبت به linear دارد .

در نتیجه برای مدل های linear مناسب تر است .

یک مشکل دیگر استفاده درجه ها بالا در polynomial این است که آیا گسترش به یک چند جمله ای با درجه ۱۶ واقعا معنی دار است؟!آیا داده های معنی دار به ما میدهد؟!(داده از درجه ۱۰ به توان ۸!!)

و اگرچه این موضوع از نظر تئوری ممکن است اما از نظر عملی برای توان های محدود کار میکند.(ارورر نمبدهد اما تحلیل مناسبی هم ندارد.)

* سوال دوم:

برای یافتن مدل مناسب و یا پارامتر هایی چون توان مناسب

میتوان

۱)

از residual plot استفاده کرد

در این plot اگر شاهد residual های تصادفی و رندوم نباشیم و خطا دارای الگوی خاصی باشد

در این صورت باید مدل را عوض کنیم و مدل بهتری جایگزین کنیم.

۲)

از روش train- test validation استفاده کنیم

به اینگونه که داده های خود را تقسیم بندی کرده و مدل را روی train set آموزش داده و روی validation و در نهایت روی داده های تست ،بسنجیم.

دستیابی به بالاترین دقت( و کمترین خطا و هزینه) روی داده های تست به ما در رسیدن به بهترین مدل کمک میکند .

۳)استفاده از bias-variance trade off

ما به دنبال مدلی هستیم که low bias و low variance باشد و total eror آن حداقل باشد .

* سوال سوم:

کاهش خطا در زمان آموزش همواره مطلوب نیست

گاهی هنگامی که دقت بسیار بالا میرود و متعاقبا خطای مدل بسیار کاهش میابد ما با high variance model مواجه هستیم.

در این حالت ،مدل تلاش میکند تا تمام داده های آموزشی را حفظ کند و به مشکل over fitting میخورد که مطلوب ما نیست و ممکن است روی داده های تست دقت کمی داشته باشد .

برای بررسی های بیشتر از راه های بالا میتوان استفاده کرد.

در نتیجه دقت ،معیار مناسبی برای سنجش مدل نیست و گاهی اگر مدلی دقت معمولی روی داده های آموزشی داشته باشد بتواند روی داده های تست عملکرد بهتری داشته باشد

در زمانی که داده های آموزشی پیچیدگی بیشتری داشته باشند(سنگین و با نویز فراوان باشند) و داده های تست آسان تر باشد خطا رو داده آموزشی بیشتر از خطا روی داده های تست خواهد بود .

ممنون میشوم نکات و مشکلات مدنظرتان را به هر طریق که صلاح میدانید به بنده برسانید.

با تشکر از توجه شما.