گزارش کار تمرین دوم

درس یادگیری ماشین

استاد درس: دکتر کمندی

دانشجویان گروه:

محیا معتمدی ۸۱۰۸۹۷۰۵۳

مینو احمدی ۸۱۰۸۹۷۰۳۲

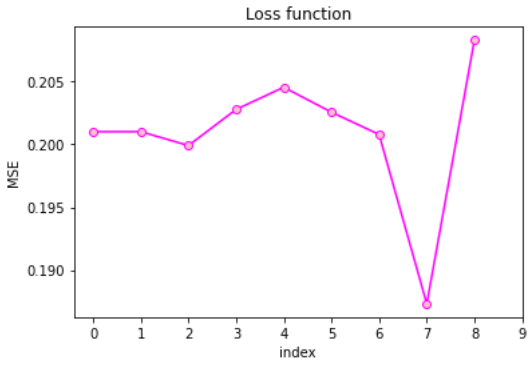
دانشکده علوم مهندسی، دانشگاه تهران

بهار ۱۴۰۱

سوال ۱)

همان طور که با دیدن دیتا ها مشهود است این دیتا ویژگی های categorical دارد که باید تبدیل به ویژگی های عددی شود به همین دلیل با تخصیص اعداد (label encoding)، سپس ان ها را به صورت رندوم به ده قسمت تقسیم کردیم و درخت تصمیم را هر بار با استفاده از یکی از ۹ دسته ساخته ایم و بعد با دسته ی دهم تست کرده ایم.
ویژگی های categorical را به numerical تبدیل کرده ایم

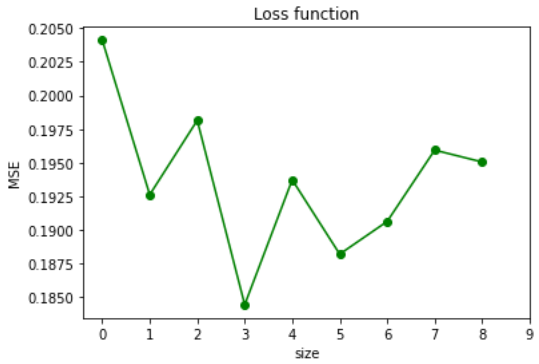
نمودار خطا های تشخیص مدل در زیر امده است :



همان طور که در جدول خطا ها میبینیم کمترین خطا را مدلی دارد که در ایندکس هفت ساخته شده (با استفاده از دسته ی هشتم) است .
در کل چون در این بخش دیتا ها شافل شده اند و به صورت رندوم پخش شده اند تقریبا شبیه به هم دسته بندی شده اند هر درخت تصمیم چون با توجه به دیتا های مخصوص به دسته خود ساخته شده است یک مقداری با درخت های تصمیم دیگر تفاوت دارد ولی تعداد دیتا های موجود در هر دسته چون به نسبت زیادند، تقریبا تخمین قابل قبول و با خطای کمی به ما داده است.

سوال ۲)

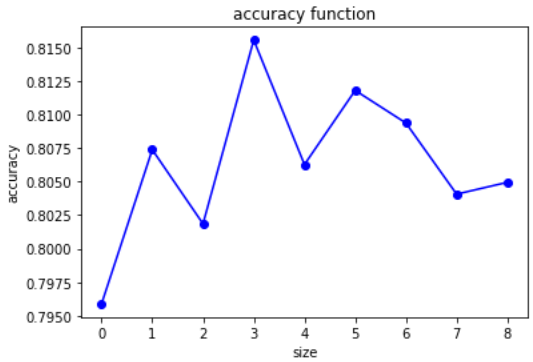
در این سوال ما درخت را ابتدا با یک دسته اموزش داده سپس دو دسته و سپس سه و.. به همین ترتیب بر تعداد دسته های اموزش افزودیم جدول میزان خطای تشخیص درخت های ایجاد شده بر اساس تعداد دسته های اموزش در زیر اورده شده است :



برای مثال این تابع خطا در سایز سه (استفاده از چهار دسته برای اموزش) کمترین خطا را داشته است و بعد از ان در سایز ۵ (استفاده از ۶ دسته برای اموزش) خطای کمی داریم.

تحلیل :

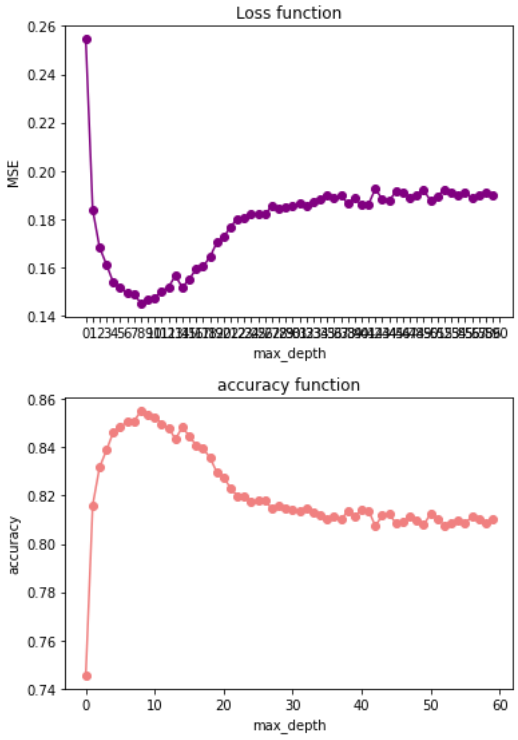
در این نمودار میبینیم که ابتدا با افزایش تعداد دسته ها برای اموزش، خطا کاهش یافته است و از یک جایی به بعد، به جای کاهش، اتفاقا افزایش خطا را داریم که نشان میدهد افزایش میزان دیتای اموزش تا حدی مناسب است ولی بیش از ان مقدار، باعث ایجاد over fitting میشود و مدل در داده های اموزش نتیجه ی خوبی دارد ولی روی داده های تست نتیجه ی مطلوبی نمیگیریم که این موضوع باعث افزایش خطا میشود.



در اینجا نمودار accuracy یا صحت تشخیص مدل را میبینیم که کاملا با نمودار خطا هم خوانی دارد . بدیهی است که هرچه خطا تشخیص کمتر باشد صحت و درستی تشخیص مدل بالاتر است که میبینیم در سایز سه که کمترین خطا بوده بالا ترین صحت تشخیص را داریم.

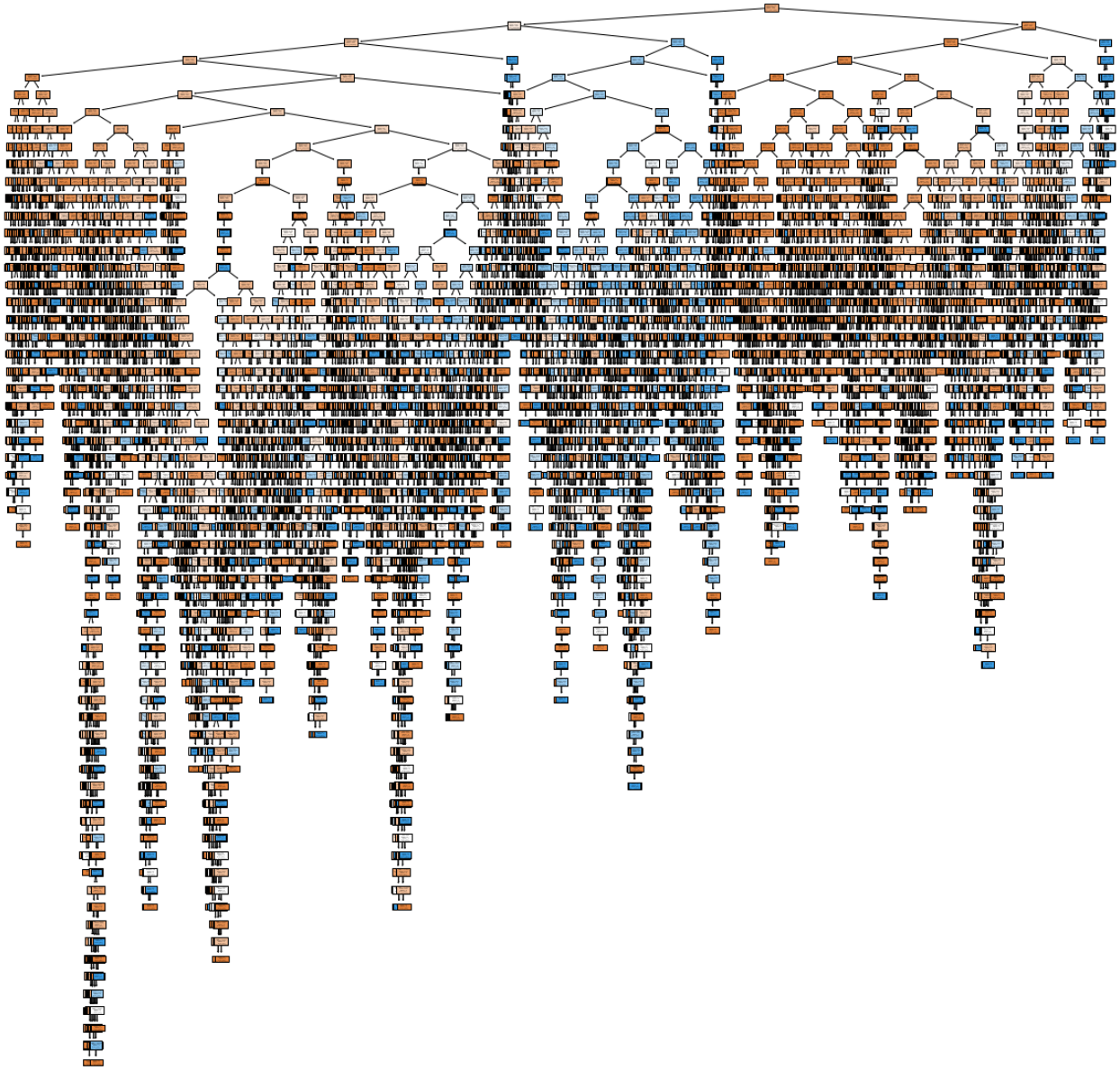
سوال ۳)

درخت تصمیم با عمق بالا دچار over fitting میشود در نتیجه از روش هرس کردن یا به عبارتی کاهش عمق استفاده میکنیم تا از این مشکل جلوگیری شود.
بیشترین عمق درخت ساخته شده ۶۱ است در نتیجه تمامی عمق ها از ۱ تا ۶۱ بررسی شده و همان طور که در نمودار مشخص است تا عمق ۸ خطا کاهش پیدا کرده ولی از این عمق به بعد از انجایی که over fitting رخ میدهد خطا افزایش پیدا میکند و accuracy کاهش میابد.

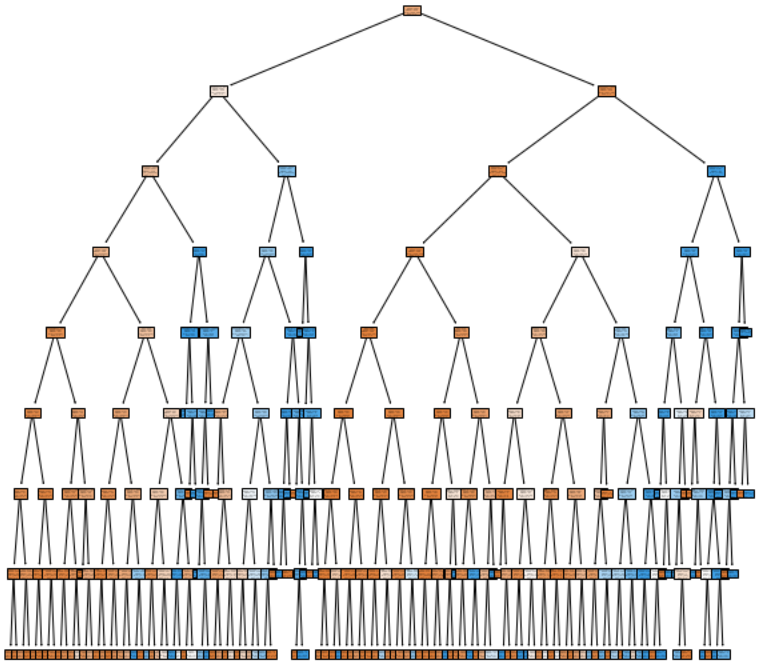


در ادامه تصویر درخت هرش نشده و هرس شده را برای مقایسه بهتر تاثیر هرس شوندگی اورده ایم.

در زیر درخت با عمق ۶۱ را میبینید:

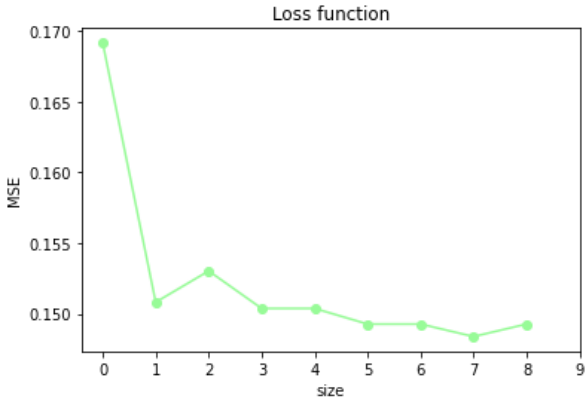


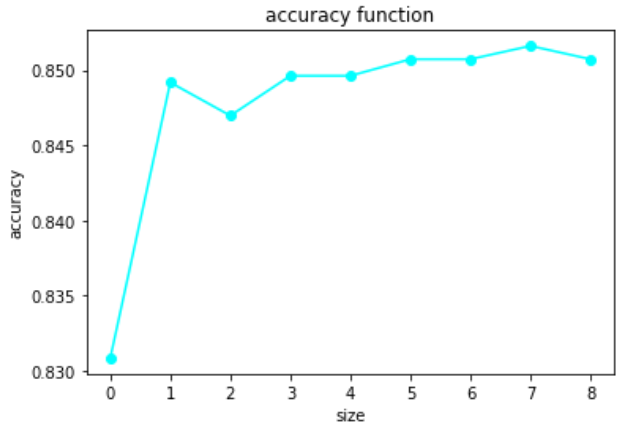
و در این بخش درخت هرس شده با عمق ۸ را میبینیم :



به وضوح با مقایسه دو درخت تاثیر هرس کردن را در تسریع رسیدن به درخت مطلوب را میبینیم.

در ادامه این سوال ما منحنی اموزش را برای درخت هرس شده رسم کردیم که در زیر امده است :





حال که درخت تصمیم هرس شده است میبینیم که این منحنی اکنون روند بسیار منطقی تری را طی میکند.

سوال ۴)

در این قسمت ما برای انتخاب ویژگی مناسب برای هر گره از روش رندوم استفاده کردیم که میزان خطا و صحت ان در زیر امده است :

```
acuuracy: 0.8246351172047767
loss: 0.17536488279522336
```

این موضوع نشان میدهد که روش رندوم دقت کمتری نسبت به روش information gain دارد (در روش رندوم accuracy حدود هشتاد و دو درصد بود که از هشتاد و شش درصد کمتر است).

روش دیگر gain ratio میباشد که میزان خطا و صحت ان در زیر امده است :

```
acuuracy: 0.8538257408226448
loss: 0.14617425917735516
```

که نسبت به رندوم بهتر بود ولی نسبت به information gain دقت کمتری داشت.