به نام خدا

عنوان:

تکلیف هشتم درس یادگیری ماشین- درخت تصمیم

استاد:

دکتر پدرام

نویسنده:

محمدعلی مجتهدسلیمانی

4033904504

تاریخ:

26/8/1403

* **سوال 1.**
* بخش 1.

Information gain یک واحدی است برای اینکه مشخص کنیم ویژگی (feature) انتخاب شده چه مقدار در جدا سازی و تفکیک داده ها خوب عمل کرده است. به این معنی است که ما از کدام یکی از ویژگی ها در هر node استفاده کنیم تا بیشترین تفکیک در داده ها داشته باشیم. برای بدست آوردن این مقدار ابتدا باید entropy را برای کل داده ها حساب بکنیم تا میزان نا خالصی یا بی نظمی در داده ها مشخص شود. entropy اگر 1 باشد یعنی داده به صورت نصف، نصف تقسیم بندی شده اند از 2 کلاس و بیشترین میزان هست و اگر 0 باشد که کمترین میزان هست یعنی تمام داده از جنس یک کلاس هستند. بعد از محاسبه entropy میتوانیم information gain را محاسبه کنیم. این امر با تفریق entropy کلی از میانگین وزن دار entropy فرزندان یک node محاسبه میشود.

* بخش 2.

یکی از معایب اصلی درختان تصمیم این است که وقتی با داده های عددی کار میکنند ممکن است بسیار پیچیده شوند بخاطر اینکه تقسیم بندی های ما معمولا باینری هست و درخت بزرگی تولید شود. هدف ما تولید درخت کوچک هست هر چه درخت بزرگتر باشد مستعد overfit شدن نسبت به داده های آموزشی خواهد بود.

یکی دیگر از معایب این است که در خروجی ها نمیتوان چندین ویژگی و ترکیبی از آنها را داشت و ویژگی های خروجی حتما باید رسته ای باشند.

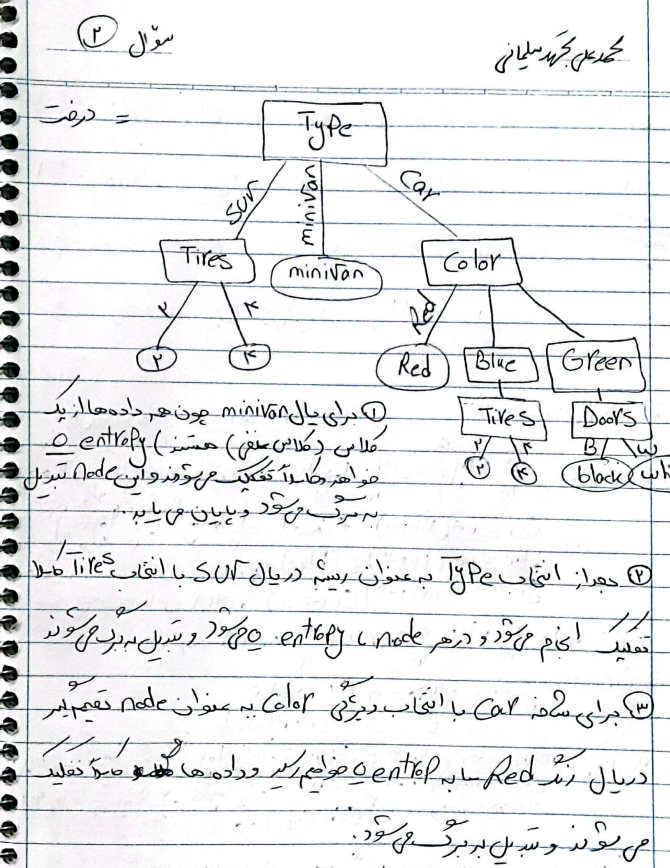
یکی دیگر از معایب این است که اگر تغییرات در داده های آموزشی به گونه ای باشد که موجب شوند انتخاب های مختلفی برای ویژگی در هر nodeیی در درخت وجود داشته باشد این الگوریتم ها ناپایدار میشوند. این مسئله وقتی شدید میشود که یک ویژگی در تمام زیر درخت ها تاثیر بگذارد.

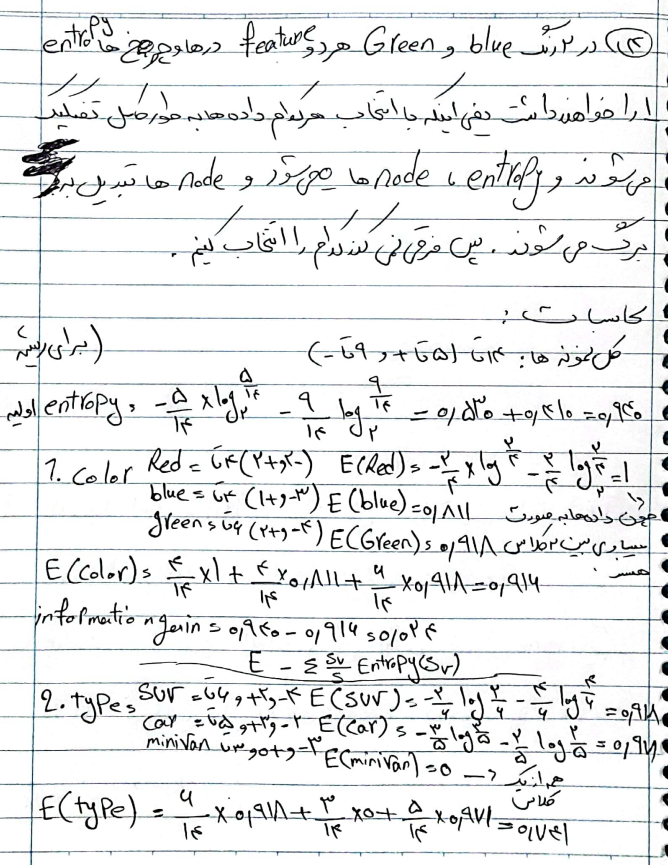
* بخش 3.

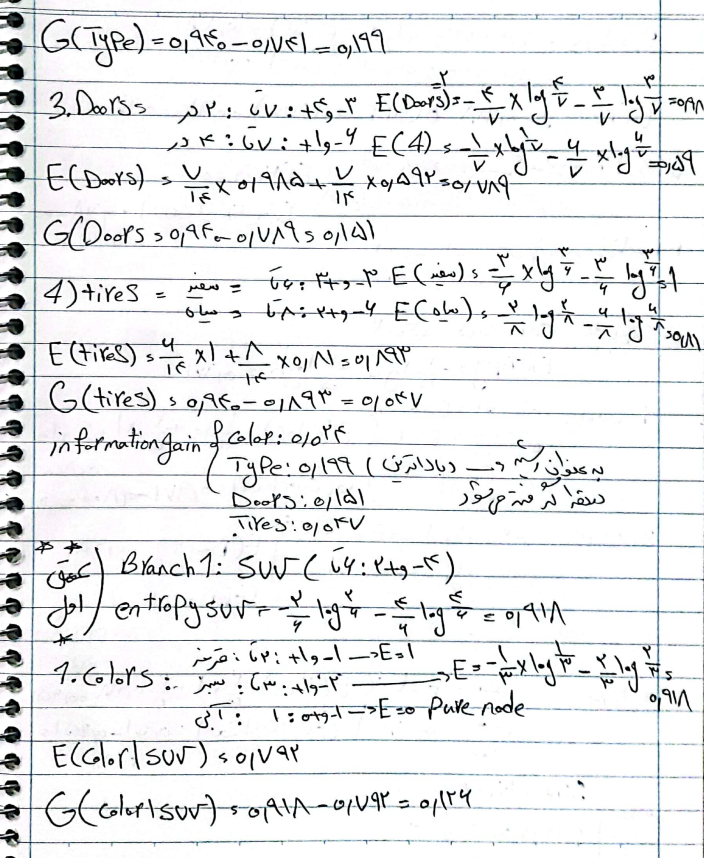
درختان تصمیم جز مجموعه nonparametric ها هستند یعنی اصلا فاز یادگیری آنها مانند شبکه های عصبی نیست که نیاز باشد به دنبال وزن ها باشند تا فرآیند یادگیری انجام شود در نتیجه بسیار سریعتر به جواب میرسند نسبت به شبکه های عصبی که فرآیند طولانی در یادگیری دارند زیرا که باید پارامتر های آزاد مثل وزن ها را مقدار دهی کنند و درختان تصمیم میتوانند با تفکیک داده ها بدون نیاز به فرآیند یادگیری این کار را انجام دهند. پیچیدگی درختان تصمیم بسیار ساده تر از شبکه های عصبی هست که از چندین لایه تشکیل شده اند. میزان محاسبات در درختان تصمیم بسیار کم و سریع هست نسبت به شبکه های عصبی که محاسبات پیچیده تری دارند.

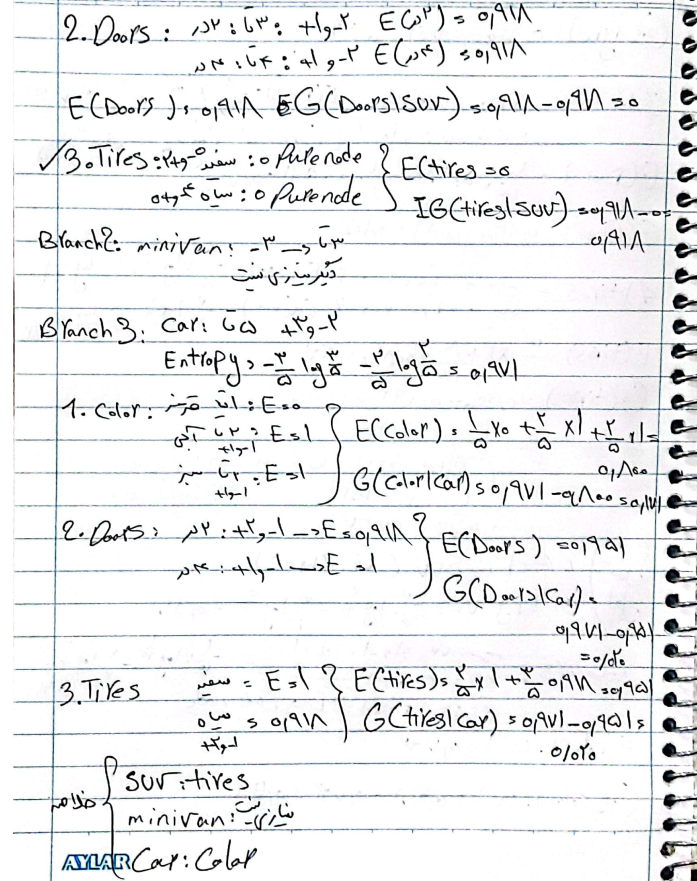
به دلیل همین سادگی به فضای حافظه کمتری نسبت به شبکه های عصبی نیاز دارند. درختان تصمیم بیشتر برای مسائل خطی مناسب هستند در صورتی که شبکه های عصبی برای مسائل غیر خطی بسیار مناسب هستند.

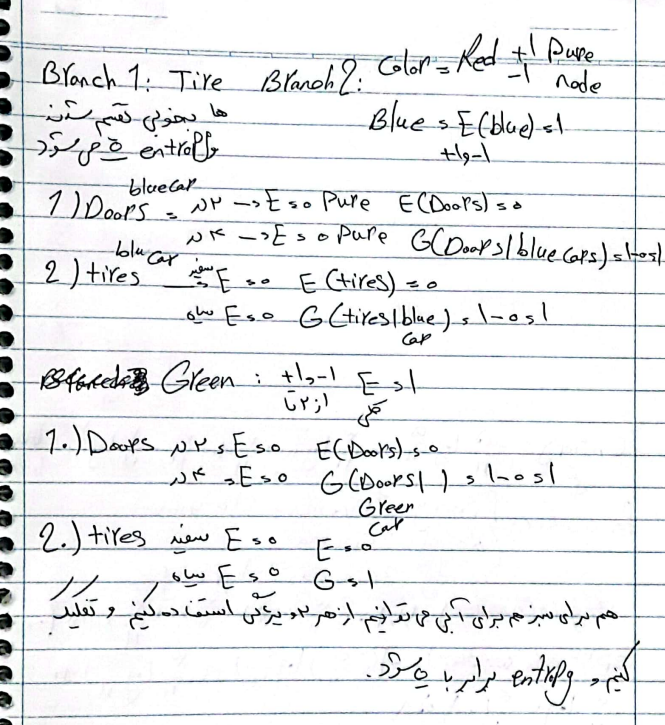
سوال 2.











سوال 3.