بسمه تعالی

مهدی وحیدمقدم

۴۰۲۱۳۰۷۴

تمرین اضافی درس یادگیری ماشین

لینک گوگل کولب:

سوال ۱)

https://colab.research.google.com/drive/1_Wysdhw8UUvAupdgqB_3Rc90OGw-Pcbb?usp=sharing

سوال ۳)

https://colab.research.google.com/drive/1i5FH2eKBKWaDdBoz1wzhGy9IDLXYx8F-?usp=sharing

لینک گیت‌هاب:

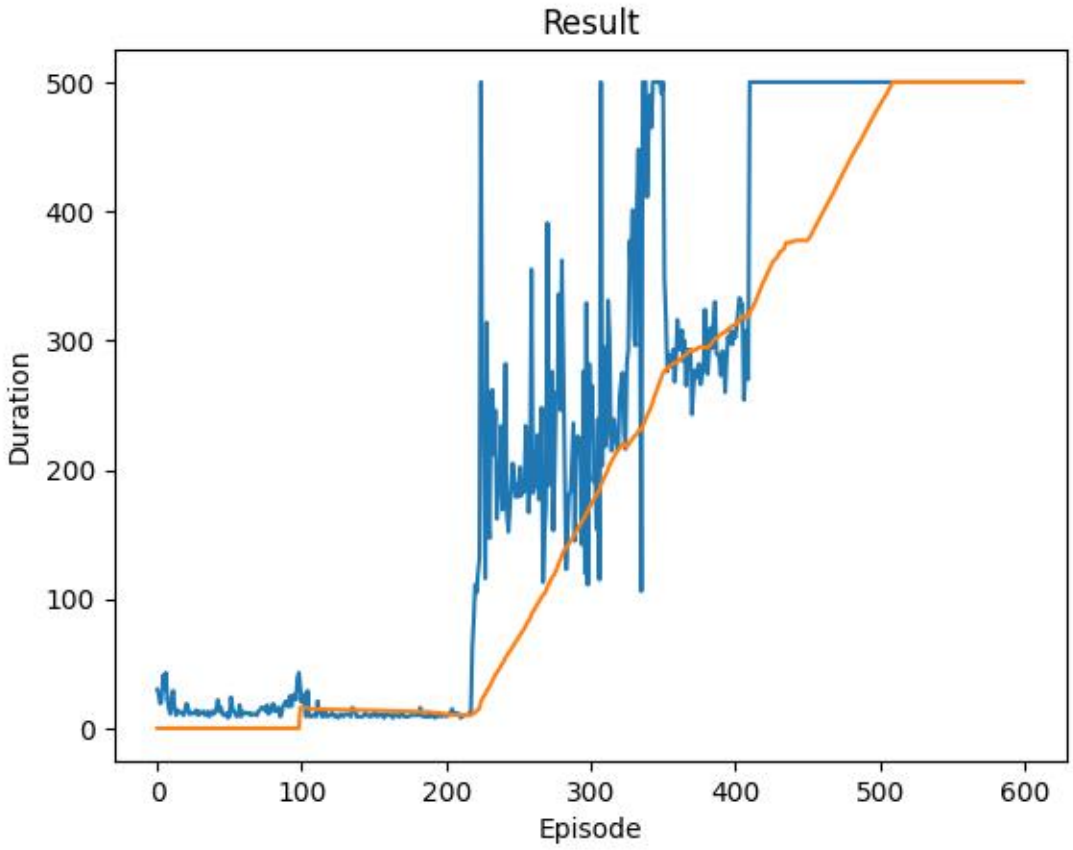
https://github.com/mvmoghadam1999/ML403_40213074/tree/main/Extra_ML_40213074

فهرست مطالب

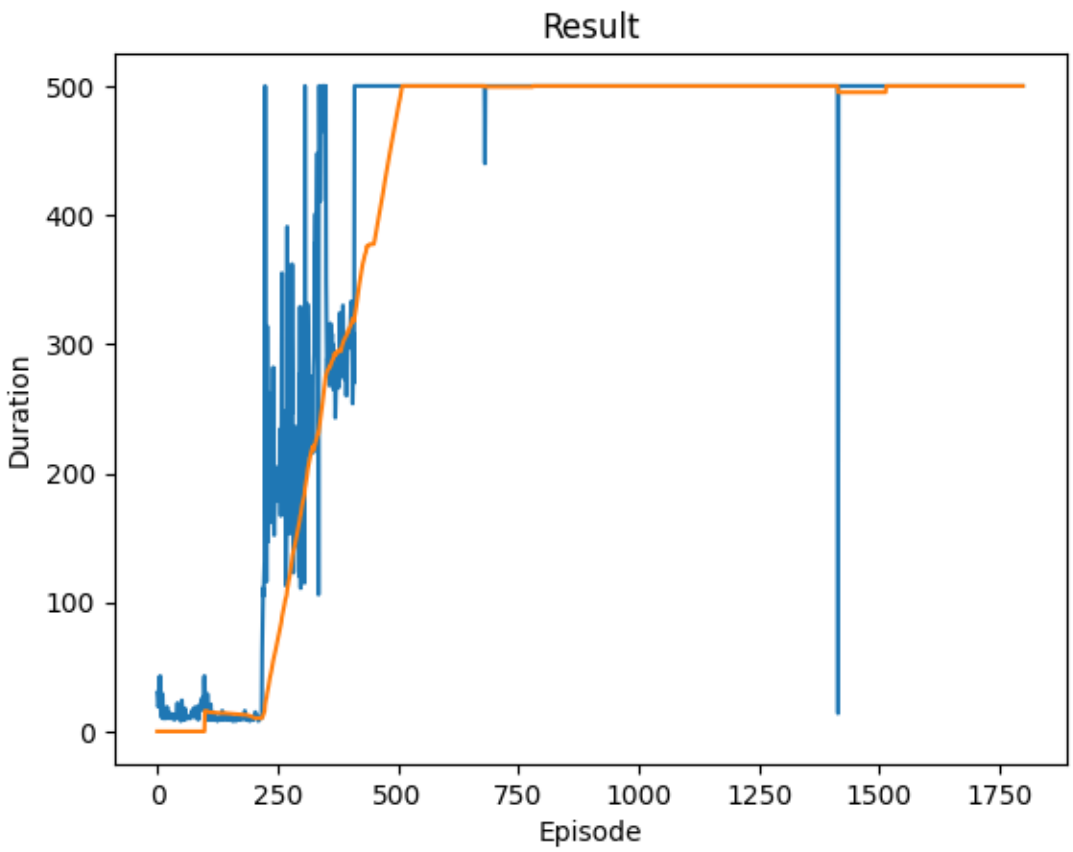
سوال ۱)

- با به‌کارگیری از GPU و قرار دادن device به cuda، کد مربوط به پاندول معکوس برای ۶۰۰ اپیزود آموزش داده شود و روند همگرایی مشاهده شود.

به ازای ۶۰۰ اپیزود خروجی به صورت زیر است:



برای این که بهتر مشخص شود همگرایی رخ داده یا نه، ۱۲۰۰ اپیزود اضافه می‌شود تا روند بهتر مشاهده شود.به ازای ۱۸۰۰ اپیزود داریم:



همان‌طور که مشاهده می‌شود همگرایی رخ داده است و سیستم کنترلی به خوبی عمل کرده است.

سوال ۲)

- در معیار آماری Kullback-Leibler-divergence که نحوه‌ی توزیع داده‌ها مورد بررسی قرار می‌گیرد،به صورت کلی نحوه‌ی توزیع $|X - Y|$ و نحوه‌ی توزیع $|Y - X|$ یکی هستند؟در واقع آیا رابطه‌ی زیر برقرار است؟

$$D_{KL}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) = D_{KL}(Q||P)$$
$$= \sum_x Q(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

در واقع Kullback-Leibler divergence یک معیار آماری است که برای اندازه‌گیری تفاوت بین دو توزیع احتمالی P و Q استفاده می‌شود. این معیار به ما کمک می‌کند تا میزان اطلاعات از دست رفته زمانی که از یک توزیع به جای دیگری استفاده می‌کنیم را بسنجیم. به طور خاص، K-L divergence بین دو توزیع P و Q به صورت زیر تعریف می‌شود:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

این رابطه در حالت پیوسته، به صورت زیر است:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

یکی از ویژگی‌های مهم KL-divergence این است که این معیار نامتقارن است، به این معنی که تغییر ترتیب توزیع‌ها معمولاً به مقادیر متفاوتی از KL-divergence منجر خواهد شد.به صورت کلی داریم:

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$$

این عدم تقارن به این معنی است که تفاوت اطلاعاتی بین دو توزیع P و Q بسته به اینکه کدام یک را به عنوان توزیع اصلی و کدام را به عنوان توزیع جایگزین در نظر بگیریم، متفاوت خواهد بود.

برای درک بهتر این موضوع، فرض کنید دو توزیع احتمالی P و Q داریم. اگر P توزیع "حقیقی" داده‌ها باشد و Q توزیعی باشد که مدل ما تولید می‌کند، $D_{KL}(P \parallel Q)$ KL divergence نشان می‌دهد که چقدر اطلاعات از دست رفته است وقتی که Q به جای P استفاده می‌شود. این

مقدار به ما می‌گوید که چقدر تفاوت بین توزیع واقعی و توزیعی که مدل ما تولید می‌کند وجود دارد.

اما اگر $D_{KL}(Q \parallel P)$ را محاسبه کنیم، این به ما می‌گوید که چقدر اطلاعات از دست رفته است وقتی که P به جای Q استفاده می‌شود. این دو مقدار معمولاً برابر نیستند مگر در موارد خاص که P و Q دقیقاً یکسان باشند.

لازم به ذکر است که KL divergence همیشه غیرمنفی است و تنها زمانی که P و Q یکسان باشند مقدار آن صفر می‌شود. به عبارت دیگر داریم:

$$D_{KL}(P \parallel Q) \geq 0$$

و همچنین:

$$D_{KL}(P \parallel Q) = 0 \; if \; and \; only \; if \; P = Q$$

بنابراین، KL divergence نشان‌دهنده‌ی میزان اطلاعاتی است که از دست می‌رود زمانی که از یک توزیع به جای دیگری استفاده می‌شود و این میزان از دست رفتن اطلاعات بستگی به ترتیب توزیع‌ها دارد.

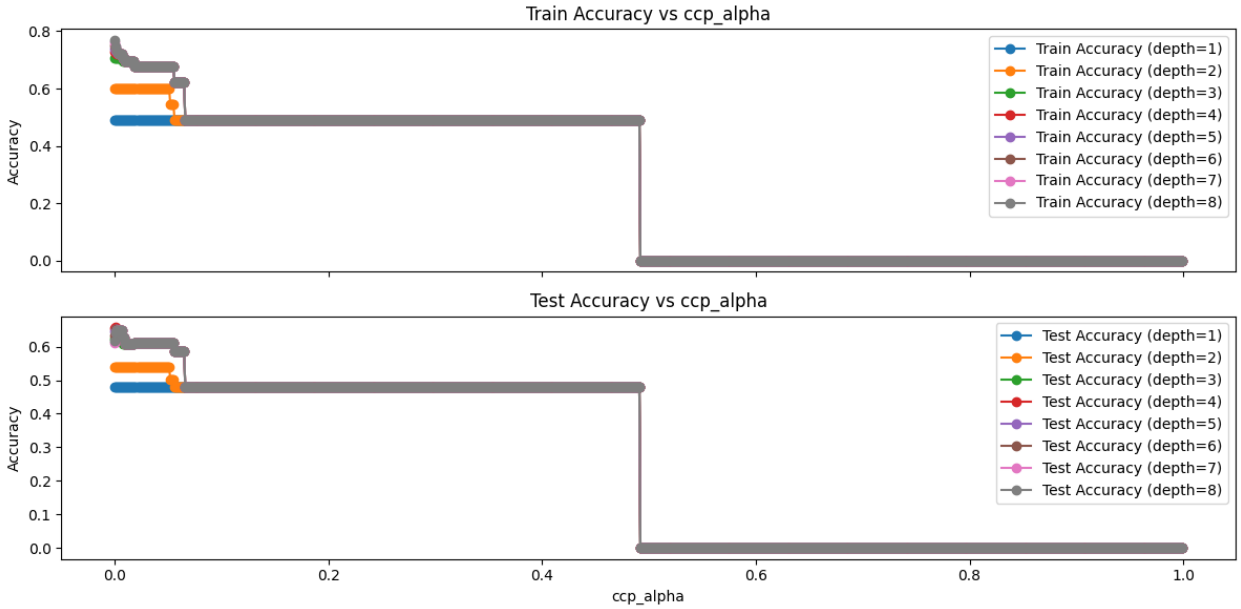
همچنین درباره‌ی توزیع $|X - Y|$ و $|Y - X|$، باید بگوییم که این دو توزیع باید یکسان باشند. به‌عبارتی، اگر X و Y دو متغیر تصادفی باشند، $|X - Y|$ و $|Y - X|$، همیشه برابرند، چون تفاوت مطلق بین دو مقدار همیشه مثبت و یکسان است. اما KL divergence $D_{KL}(Q \parallel P)$ و $D_{KL}(P \parallel Q)$ برابر نیست و به ترتیب دو توزیع بستگی دارد. به این ترتیب، KL divergence یک معیار نامتقارن است که به ترتیب توزیع‌ها حساس است و تغییر ترتیب می‌تواند به مقادیر متفاوتی منجر شود.

سوال ۳)

- پارامترهای یک درخت تصمیم برای رسیدن به بهترین دقت عملکرد تغییر کنند تا به بهترین دقت برسیم.

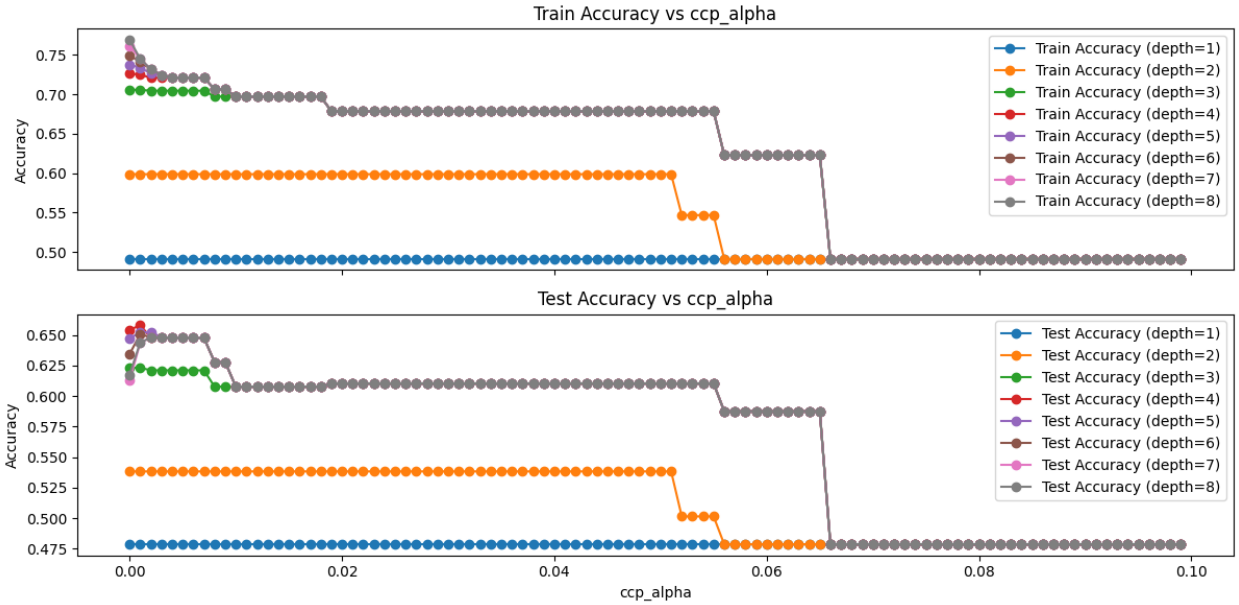
در بخش اول آموزش به ازای مقادیر مختلف max_depth از ۱ تا ۹ و به ازای تغییرات مقدار ccp_alpha از ۰.۰۰۰۰۱ تا ۱ دقت‌های به دست آمده نشان داده می‌شوند.پارامتر max_depth حداکثر عمقی است که برای درخت در نظر می‌گیریم.(طبیعتا این پارامتر هر چه بیشتر باشد،دقت بیشتر خواهد بود.همچنین پارامتر ccp_alpha مربوط به میزان هرس درخت است.این پارامتر هم هر چه بیشتر باشد،هرس بیشتری خواهیم داشت و طبیعتا دقت کمتر خواهد شد.

در شکل زیر مقادیر دقت به ازای مقادیر مختلف پارامترها را مشاهده می‌کنیم:



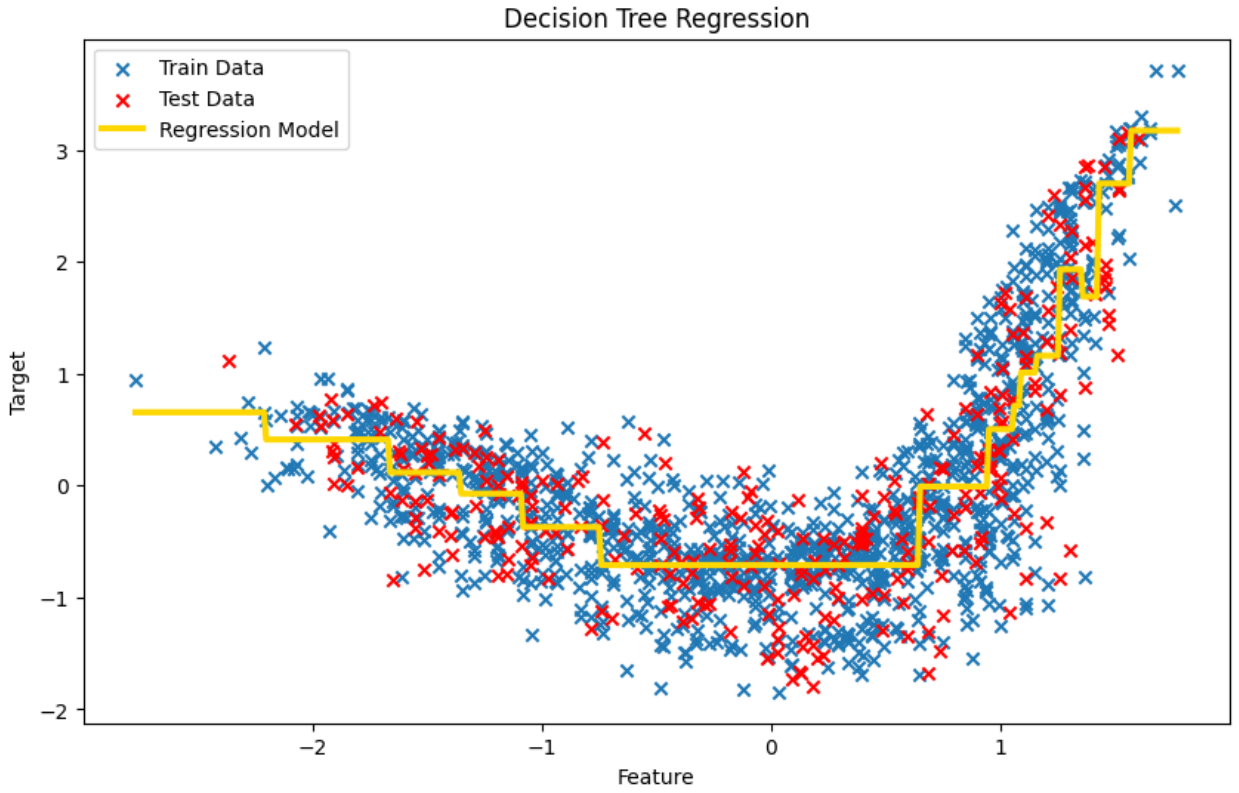
همان طور که مشاهده می‌شود،دقت به ازای مقادیر ccp_alpha بیشتر از ۰.۵ به صفر رسیده است و همان طور که گفته شد،هر چه مقدار این پارامتر کمتر باشد،عملکرد مدل بهتر خواهد بود.

حال همین کار را به ازای تغییرات ccp_alpha تا ۰.۱ انجام می‌دهیم.چون مقادیر بالاتر عملا بدتر هستند.داریم:



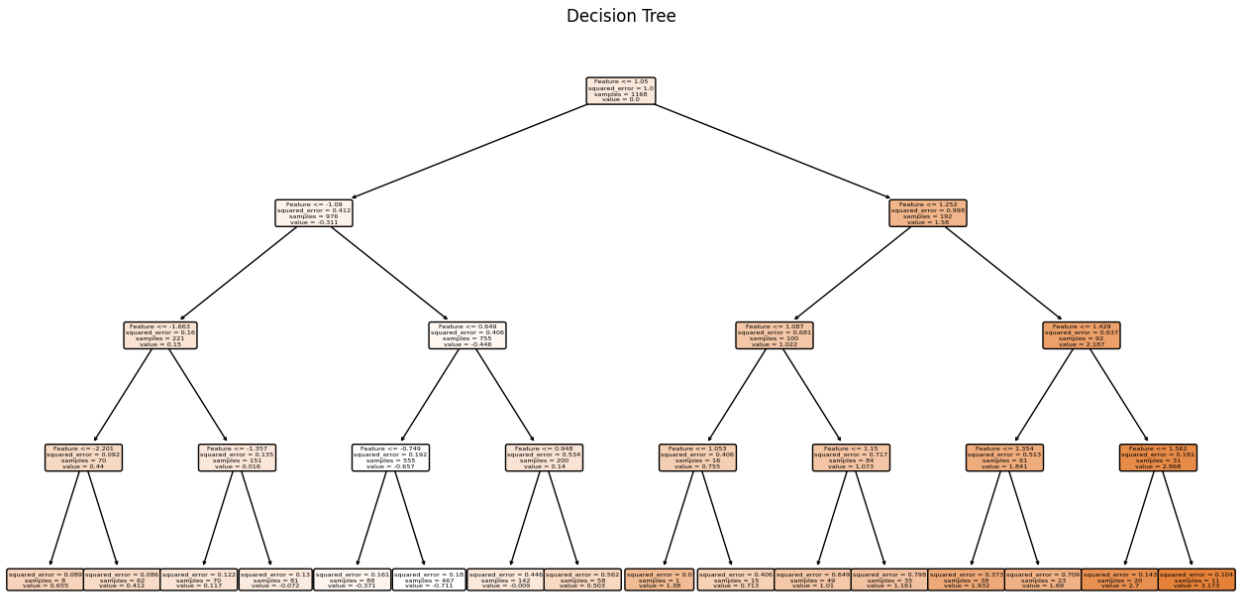
در این حالت هم می‌بینیم که بهترین عملکرد برای کمترین مقدار از ccp_alpha است.همچنین برای مقادیر عمق مختلف،مقدار $depth = 4$ تقریبا می‌توان گفت بهترین عملکرد را داشته است.

حال به ازای این مقادیر طبقه‌بندی انجام می‌شود.نتیجه به صورت زیر است:



همان‌طور که مشاهده می‌شود تقریبا تفکیک خوبی توسط مدل ما انجام شده است.خط تفکیک کننده تغییرات کوچکی دارد و به صورت خیلی آهسته تغییرات شیب انجام شده است.

همچنین درخت تصمیم به صورت زیر است:



همان طور که می‌بینیم عمق درخت همان طور که تعیین کرده بودیم برابر ۴ است.

همچنین دقت مدل به صورت زیر است:

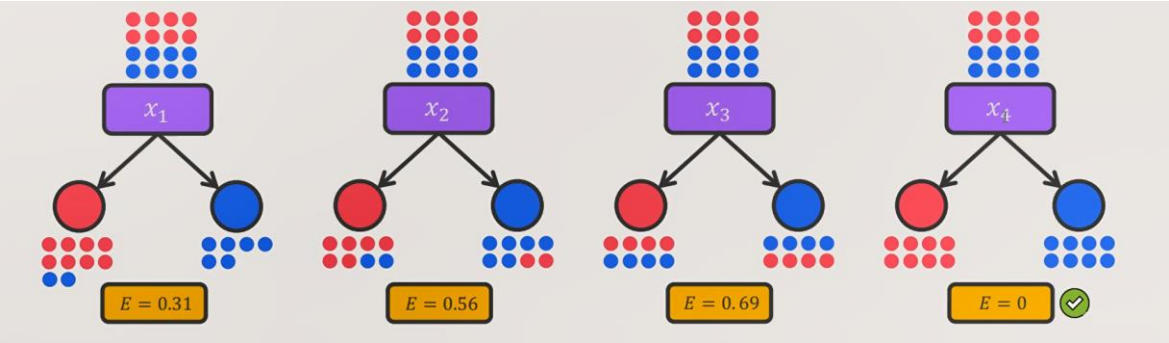
```
Train Accuracy: 0.7266
Test Accuracy: 0.6538
```

در قسمت آموزش به دقت ۷۳ درصد و در قسمت ارزیابی به دقت ۶۶ درصد رسیده‌ایم که تقریبا دقت خوبی است.

سوال ۴)

- کدام یک از ویژگی‌های نشان داده شده در شکل زیر برای decision node بهتر هستند و محاسبه‌ی IG برای ویژگی‌های ۲ و ۴ انجام شود.

شکل زیر کل ویژگی‌ها را نشان می‌دهد:



محاسبه برای ویژگی۲:

طبق شکل بالا عمل می‌کنیم:

$$IG = E(parent) - \sum_i w_i E_i(child)$$

$$E = -\sum_i p_i \log_2 p_i$$

$$E_r = -(p_0 log p_0 + p_1 log p_1)$$

$$E_L = -(p_0 log p_0 + p_1 log p_1)$$

$$E_t = w_r E_r + w_l E_l$$

$$E(parent) = 0.69$$

$$E_{t_{x_2}}(child) = 0.56$$

$$IG_{x_2} = 0.69 - 0.56 = 0.13$$

حال برای ویژگی۴ هم همین روند را انجام می‌دهیم:

$$IG = E(parent) - \sum_i w_i E_i(child)$$

$$E = -\sum_i p_i \log_2 p_i$$

$$E_r = -(p_0 log p_0 + p_1 log p_1)$$

$$E_L = -(p_0 log p_0 + p_1 log p_1)$$

$$E_t = w_r E_r + w_l E_l$$

$$E(parent) = 0.69$$

$$E_{t_{x_4}}(child) = 0.0$$

$$IG_{x_4} = 0.69 - 0.0 = 0.69$$

همان‌طور که مشاهده می‌شود مقدار IG برای ویژگی۴ بیشتر است پس این ویژگی برای decision node بهتر است.

پایان