به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس یادگیری ماشین

استاد ناظرفرد

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: پرسش‌های تشریحی

سوال ۱

الف) ؟

ب)‌ مانند تعیین مقدار هر ابرپارامتر دیگر می‌توان از یک مجموعه validation برای تعیین آن استفاده کرد. یعنی آنکه به ازای مقادیر k مختلف بررسی می‌کنیم که خطای مجموعه validation به چه شکلی بوده است. جایی که کمترین خطا را برای این مجموعه داشتیم احتمالا k مناسب داشته است.

سوال ۲

الگوریتم KNN‌ را می‌توان یک الگوریتم تمایز‌گر دانست؛ چراکه در این الگوریتم حساب می‌شود که یک داده به کدام کلاس تعلق دارد و مرز‌های کلاس‌ها مختلف تشکیل می‌شود (مانند سلول‌های voronoi برای 1NN) و نمی‌توان الگوی یک کلاس را محاسبه کرد و توسط آن داده‌های جدید ایجاد کرد.

الگوریتم درخت تصمیم را هم باید یک الگوریتم تمایزگر دانست؛ چراکه در این الگوریتم هم باز صرفا بررسی می‌شود که یک داده به کدام کلاس تعلق دارد و می‌توان مرز‌ها را برای کلاس‌های مختلف تشکیل داد ولی نمی‌توان توزیع هر کلاس را بدست آورد.[[1]](#footnote-1)[[2]](#footnote-2)

سوال ۳

الف)

سوال ۴

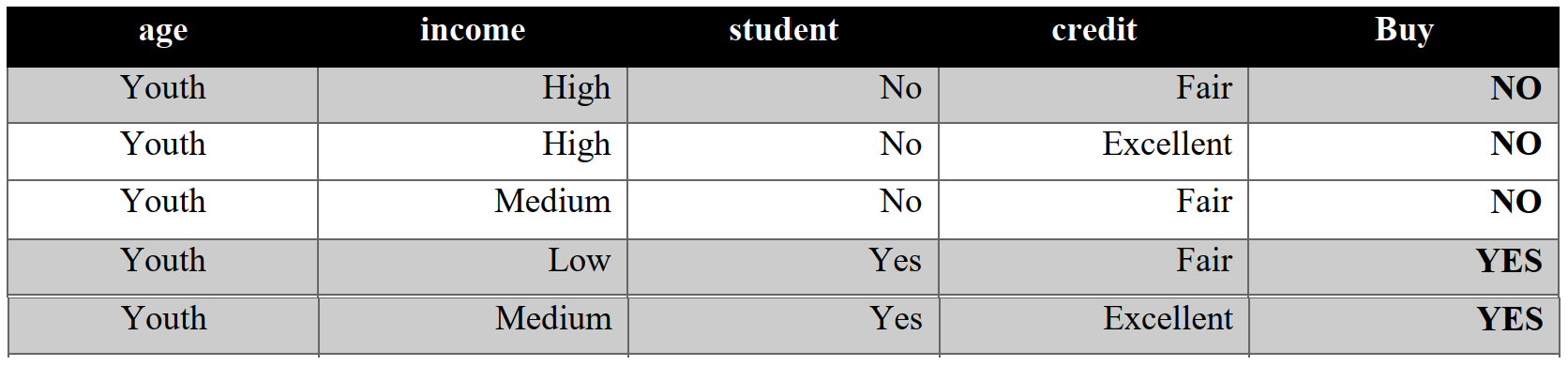
الف)

ب)

ابتدا باید بررسی کرد که برای ریشه درخت کدام ویژگی مناسب‌تر است.

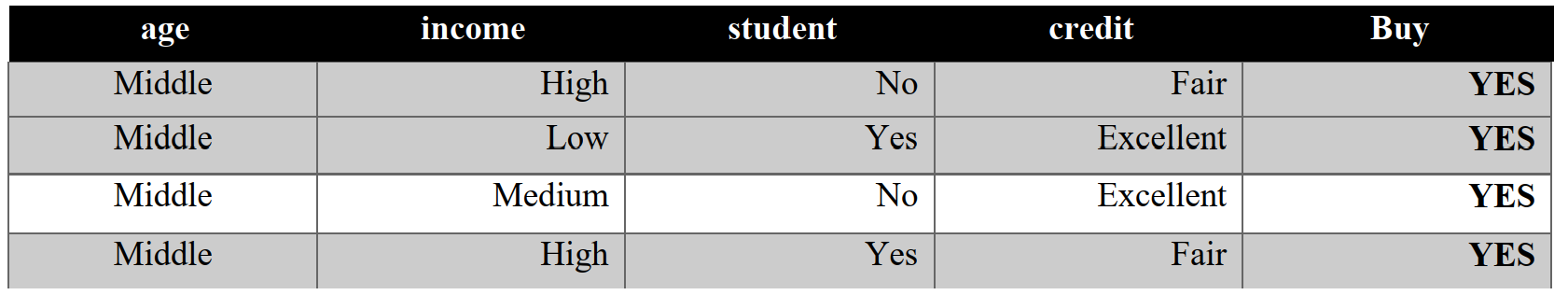
به نظر می‌رسد برای ریشه پرسش راجع به age بهترین گزینه است.

حال باید بررسی کرد که برای شاخه age = youth کدام ویژگی مناسب‌تر است.

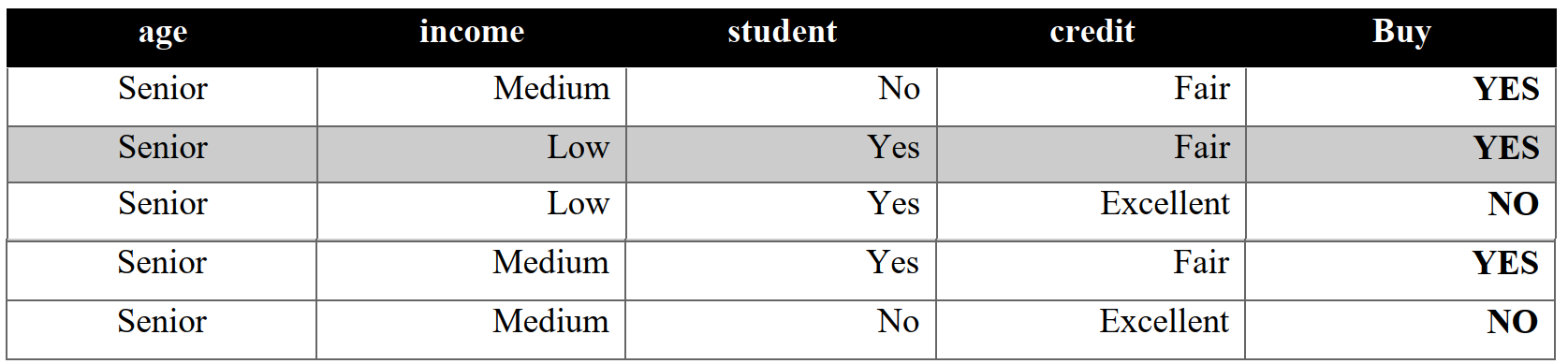


طبیعتا برای این شاخه هم می‌توان مشابه با محاسبات گره ریشه پیش رفت؛ اما پیش از آن توجه کنید که اگر student‌ مورد سوال قرار بگیرد قطعا دو گره خالص ایجاد خواهد شد این درحالی است که در مابقی ویژگی‌ها حداقل یک گره ناخالص باقی می‌ماند. ( برای income، مقدار medium؛ برای credit، مقدار fair) پس قطعا Gain ویژگی Student برابر بیشینه مقدار که انتروپی همین گره age= youth باشد است که از مابقی بیشتر است و نیازی به محاسبات بیشتر نیست.

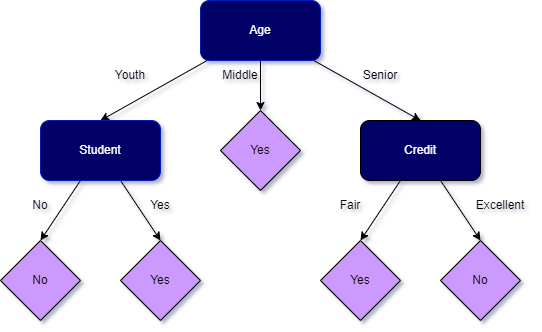
برای شاخه age = middle یک گره خالص باقی می‌ماند که نیاز به ادامه ندارد و می‌دانیم به پاسخ yes خواهیم رسید.



برای شاخه آخر یعنی age = senior، واضحا و مشابه با بحث گره age =youth انتخاب ویژگی credit بهترین گزینه است چراکه دو گره خالص حاصل می‌شود و Gain این ویژگی برابر با آنتروپی گره age = senior‌ است که بیشترین Gain ممکن است. برای دو ویژگی دیگر هم مشخص است که گره‌های حاصل تماما خالص نخواهند بود.

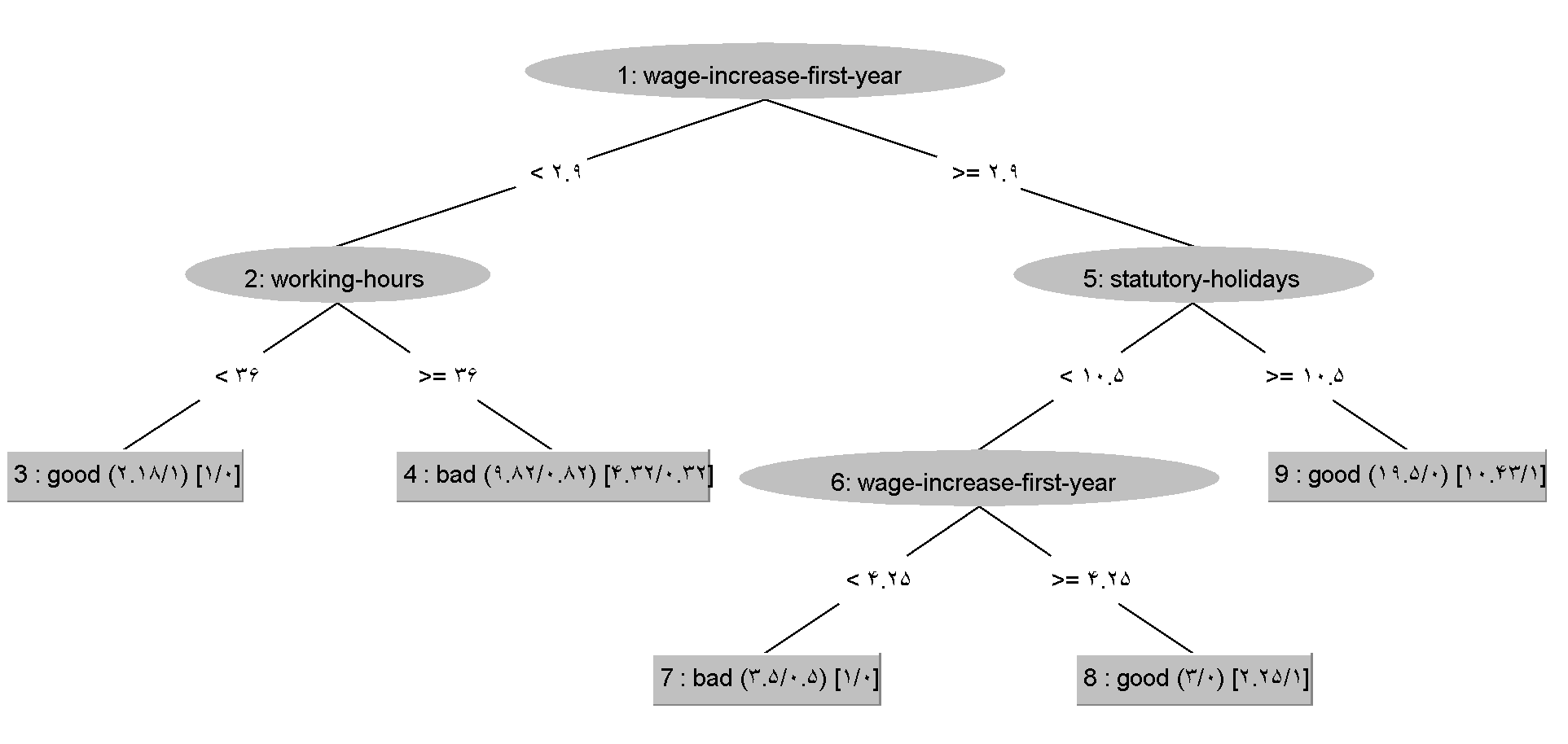


درخت نهایی برابر خواهد بود با:

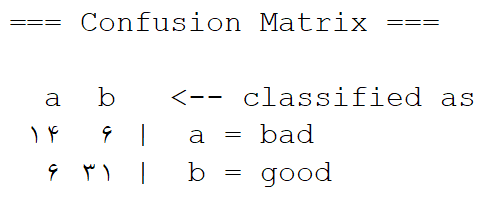


سوال ۵

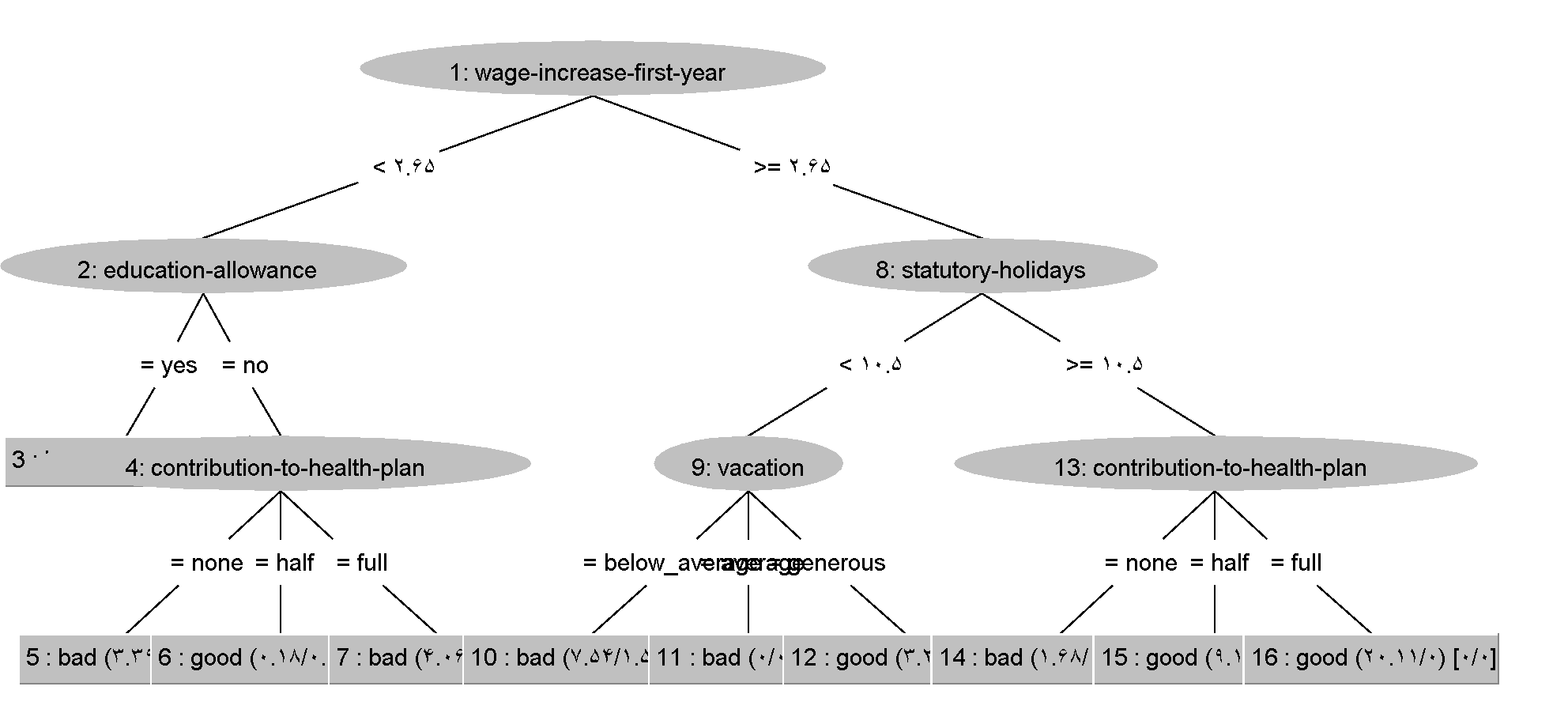
الف) برای این سوال از الگوریتم REPTree موجود در نرم‌افزار استفاده کردم و سایر پارامتر‌ها را تغییر ندادم.



ب)



ج)



د) مطابق انتظار درخت هرس‌نشده تعداد گره بیشتری دارد. همچنین می‌توان مشاهده کرد که درصد خلوص در گره‌های انتهایی درخت هرس‌نشده بیشتر است.

1. https://stats.stackexchange.com/questions/105979/is-knn-a-discriminative-learning-algorithm [↑](#footnote-ref-1)
2. https://stats.stackexchange.com/questions/12421/generative-vs-discriminative [↑](#footnote-ref-2)