بسم الله الرحمن الرحیم

درخت تصمیم گیری

- درس هوش مصنوعی
- دکتر مهدی قاسمی ورنامخواستی

امید امیدی – دانشجوی دکتری مهندسی مکانیک ماشین های کشاورزی – دانشگاه شهرکرد

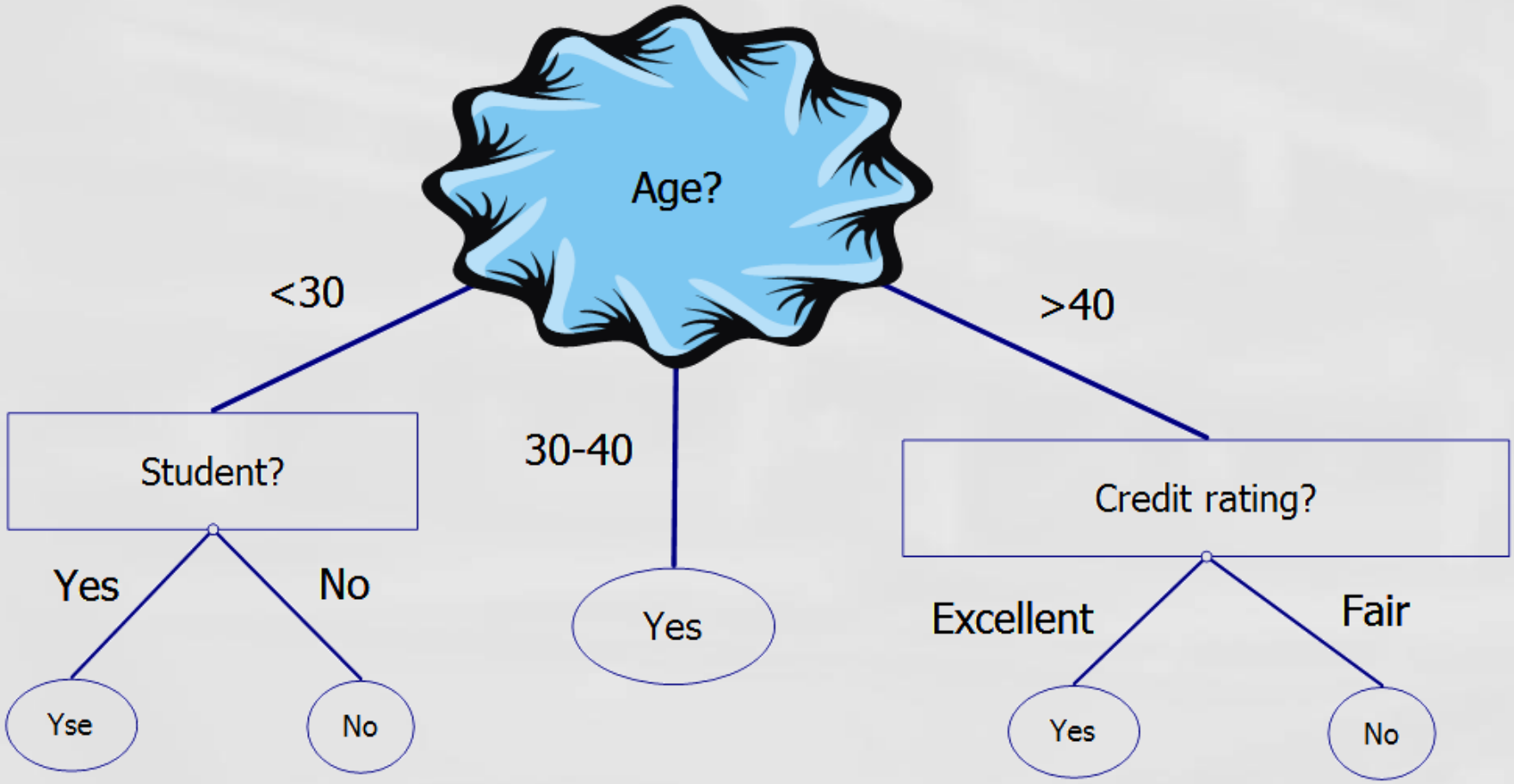
تعریف

- درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته بندی و پیش بینی است.
- درخت تصمیم گیری یک ساختار درختی شبیه فلو چارت است.
- در این ساختار هر گره داخلی آزمونی را بر روی یک ویژگی مشخص میکند.
- گره های برگ، کلاس ها یا توزیع کلاس ها را ارائه میکنند.
- بالاترین گره در درخت گره ریشه است.

تعریف

- درخت ها در هوش مصنوعی برای نمایش مفاهیم مختلفی نظیر ساختار جملات، معادلات، حالات بازی، و غیره استفاده میشود.
- یادگیری درخت تصمیم روشی برای تقریب توابع هدف با مقادیر گسسته است. این روش نسبت به نویز داده ها مقاوم بوده و قادر است ترکیب فصلی گزاره های عطفی را یاد بگیرد.
- این روش ID3 مشهورترین الگوریتم های یادگیری استقرایی است که به صورت موفقیت آمیزی در کاربردهای مختلف بکار گرفته شده است.

ساختار درخت تصمیم گیری



خرید کامپیوتر

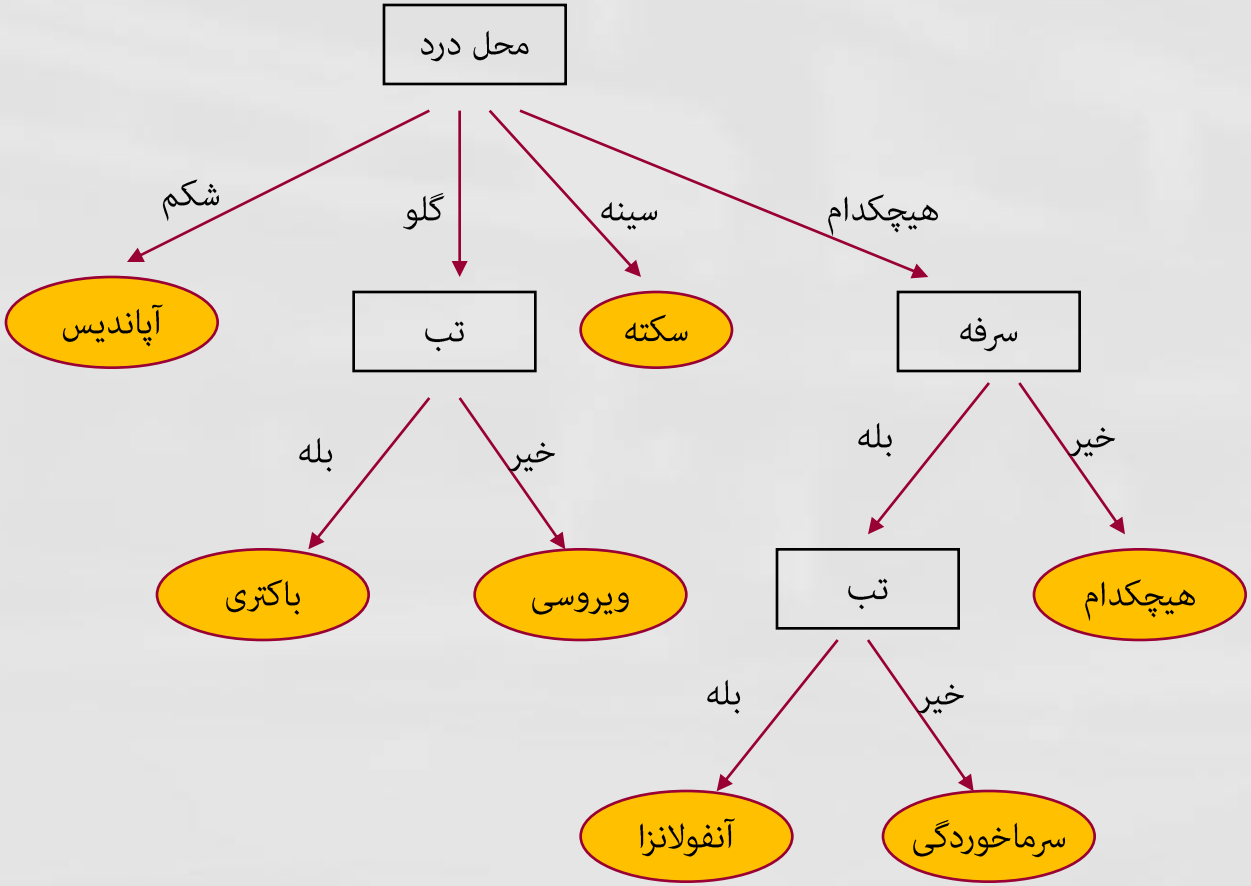
تعریف

- درخت تصمیم درختی است که در آن نمونه ها را به نحوی دسته بندی میکند که از ریشه به سمت پائین رشد میکنند و در نهایت به گره های برگ میرسد.
- هر گره داخلی یا غیر برگ non leaf با یک ویژگی attribute مشخص میشود. این ویژگی سوالی را در رابطه با مثال ورودی مطرح میکند.
- در هر گره داخلی به تعداد جواب های ممکن با این سوال شاخه branch وجود دارد که هر یک با مقدار آن جواب مشخص میشوند.

تعریف

- برگهای این درخت با یک کلاس و یا یک دسته از جوابها مشخص میشوند.
- علت نامگذاری آن با درخت تصمیم این است که این درخت فرایند تصمیم گیری برای تعیین دسته یک مثال ورودی را نشان میدهد.

نمونه ای از درخت تصمیم گیری



خصوصیات درخت تصمیم گیری

- هر گره داخلی یک ویژگی را تست میکند.
- هر شاخه متناسب با مقدار یک ویژگی است.
- هر گره برگ نشان از یک کلاس است.

خصوصیات درخت تصمیم گیری

- درخت تصمیم پیش بینی خود را در قالب یک سری قوانین توضیح می دهد در حالیکه در شبکه عصبی تنها پیش بینی بیان می شود و چگونگی آن در خود شبکه پنهان است.

- در درخت تصمیم گیری بر خلاف شبکه عصبی لزومی ندارد که داده ها به صورت عددی باشند.

درخت تصمیم گیری چگونه کار میکند؟

- در درخت تصمیم گیری یک سری سؤال وجود دارد و با مشخص شدن پاسخ هر سؤال یک سؤال دیگر پرسیده می شود. اگر سؤال ها درست و خوب پرسیده شوند یک سری کوتاه از سؤالات برای پیش بینی دسته رکورد جدید کافی است.

قوانین نمایش درخت تصمیم گیری

- XOR
- ∧ یا
- ∨ و

X	Y	O
0	0	1
0	1	0
1	0	0
1	1	1

© CODERSOURCE.NET

XNOR

X	Y	O
0	0	0
0	1	1
1	0	1
1	1	0

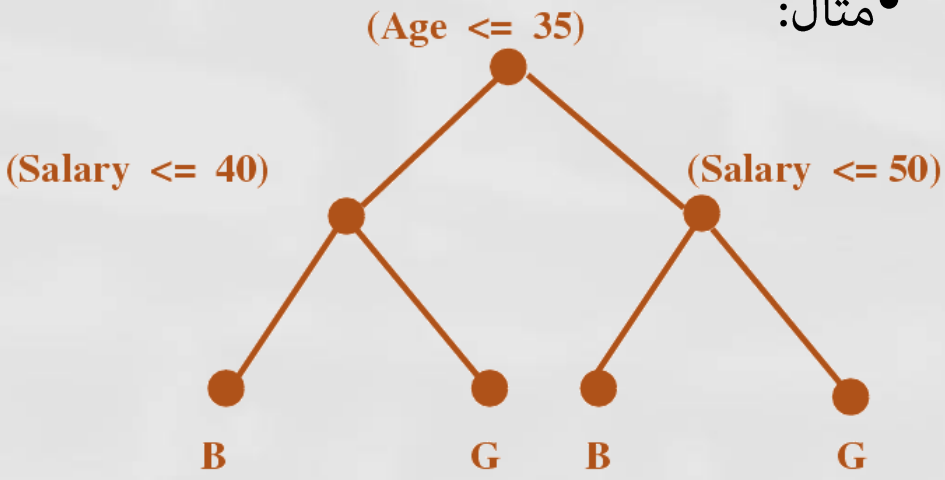
© CODERSOURCE.NET

XOR

ساختار درخت تصمیم گیری

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

- مثال:



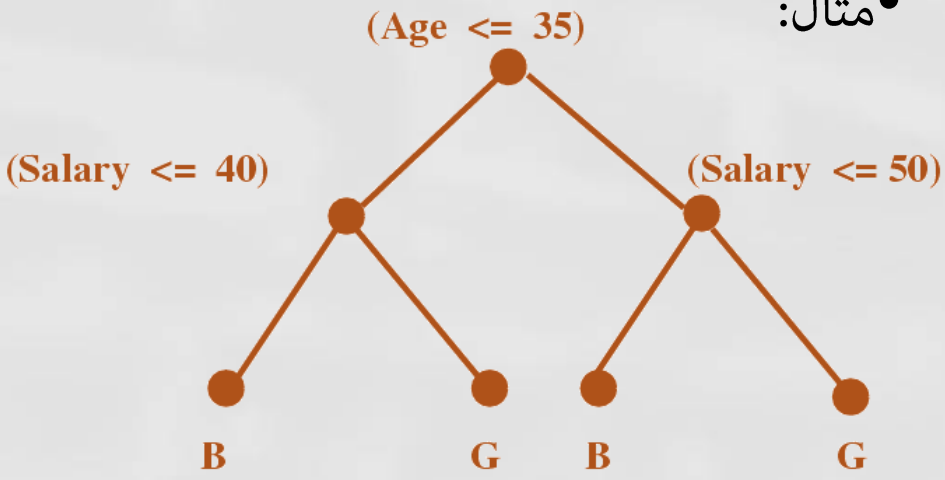
- قوانین دسته بندی:

Class B: (Age <= 35 AND Salary <= 40) OR (Age > 35 AND Salary <= 50)
Class G: (Age <= 35 AND Salary > 40) OR (Age > 35 AND Salary > 50)

ساختار درخت تصمیم گیری

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

- مثال:



- تست:

Test data: Age = 25 AND Salary = 50
Class = G

اثر بخشی یک درخت تصمیم گیری

- درصد داده هایی که درست دسته بندی می شوند و دسته پیش بینی شده با دسته واقعی آنها یکسان است.
- کیفیت شاخه های ایجاد شده مهم است؛ هر راه ایجاد شده از ریشه به یک برگ معادل یک قانون است که برخی قانون ها بهتر از سایر قانون ها هستند. در برخی اوقات بریدن برخی شاخه های ضعیف تر درخت باعث بهبود قدرت پیش بینی درخت می شود.

انواع متغیر ها در درخت تصمیم گیری

- متغیر های عددی مثل سن، قد، وزن و...
- متغیر های رده ای مثل نوع، جنس، کیفیت و ...
- متغیر های مستقل مثل متغیر های CAR و AGE در مثال بعد، که با گره نشان داده می شوند.
- متغیر های وابسته مثل متغیر RISK در مثال بعد که با برگ نشان می دهند.
- اگر متغیر وابسته عددی بود مسئله یک مسئله Regression می شود و اگر رده ای بود مسئله یک مسئله Clasification می شود.

مراحل ایجاد درخت تصمیم گیری

- مرحله رشد و ایجاد درخت
- مرحله هرس درخت به منظور کاهش خطاها

الگوریتم های متفاوتی برای ایجاد درخت وجود دارد.

مراحل ایجاد درخت تصمیم گیری

- روش ID3
- روش C4.5
- روش CART (Classification And Regression Trees)
- در دو روش آخر ابتدا درخت تصمیم گیری با روش بالا به پایین ساخته میشود. و در مرحله بعد با استفاده از یک الگوریتم هرس شاخه های اضافی درخت حذف می شوند.

مراحل ایجاد درخت تصمیم گیری

- روش ID3 یا Iterative Dichotomiser 3
- برای انتخاب بهترین ویژگی ها برای کلاس بندی از مقدار آماری بهره اطلاعات یا Information Gain استفاده میکند.

روش ID3یا Iterative Dichotomiser 3

- در این روش تمام ویژگی های نمونه های آزمایش که مورد بررسی قرار نگرفتند را به همراه مقدار انتروپی شان بررسی میکند.
- ویژگی ای که مقدار انتروپی حداقل دارد را بر میگزیند.
- گره مربوط به آن ویژگی را ایجاد میکند.

روش ID3یا Iterative Dichotomiser 3

- انتروپی مقدار خلوص یا ناخالصی، همگنی یا ناهمگنی یک مجموعه نمونه دلخواه را بیان میکند. در واقع چگونگی اختلال را بیان میکند.
- انتروپی یک مجموعه همگن صفر است.
- نمونه های با تقسیم مساوی انتروپی ۱ دارند.
- رابطه انتروپی:

$$Info(D) = -\sum_{i=1}^{m} p(C_i)\log_2(p(C_i)),$$

روش ID3یا Iterative Dichotomiser 3

- رابطه انتروپی:

$$Info(D) = -\sum_{i=1}^{m} p(C_i)\log_2(p(C_i)),$$

- مقدار p، میزان فراوانی نمونه خارجی در یک کلاس به کل جمعیت نمونه ها است.
- این رابطه به رابطه انتروپی شانون هم معروف است. (Shannon)

روش ID3یا Iterative Dichotomiser 3

- فرض کنیم ویژگی A برای شکستن D به v جزء یا زیر مجموعه استفاده شده باشد.

(مجموعه های D1 تا Dv)

- مقدار اطلاعاتی که ما برای کلاس بندی نیاز داریم از رابطه زیر حاصل میشود:

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \mathrm{Info}(D_j).$$

روش ID3یا Iterative Dichotomiser 3

- فرض کنیم ویژگی A برای شکستن D به v جزء یا زیر مجموعه استفاده شده باشد.

(مجموعه های D1 تا Dv)

- مقدار اطلاعاتی که ما برای کلاس بندی نیاز داریم از رابطه زیر حاصل میشود:

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \mathrm{Info}(D_j).$$

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D)$$

- بزرگترین بهره حاصل را انتخاب میکنیم.

روش ID3یا Iterative Dichotomiser 3

- بهره اطلاعات بر اساس کاهش در انتروپی پایه ریزی شده است بعد از تقسیم و شکستن در یک ویژگی.
- در مرحله اول انتروپی کل داده ها محاسبه می شود.
- داده ها به ویژگی های متفاوت شکسته می شود.
- انتروپی برای هر شاخه محاسبه میشود. سپس به تناسب هر کدام انتروپی کل شکست محاسبه می شود.
- انتروپی حاصل از انتروپی قبل از شکست کم می شود. (نتیجه همان بهره اطلاعاتی است و یا همان کاهش انتروپی.
- ویژگی ای که محصول بزرگترین بهره اطلاعاتی است برای تصمیم گره انتخاب می شود.

روش ID3یا Iterative Dichotomiser 3

- یک شاخه با انتروپی صفر یک گره برگ است.
- از طرف دیگر، شاخه نیاز به شکست بیشتری برای کلاس بندی داده ها دارد.
- الگوریتم ID3به صورت بازگشتی روی شاخه های بدون برگ تا زمانی که همه داده ها کلاس بندی شوند اجرا می شود.

روش ID3 و ایجاد Overfitting

- الگوریتم ID3 هر شاخه از درخت را آنقدر به عمق میبرد که بتواند بطور کامل مثال های آموزشی را دسته بندی کند. این امر میتواند منجر به Overfitting شود. دلایل بروز Overfitting عبارتند از:
  - وجود نویز در داده های آموزشی
  - تعداد کم مثال های آموزشی
- برای مثال اگر فقط دو بار پرتاب سکه داشته باشیم و هر دو بار شیر آمده باشد چه نتیجه ای در مورد این آزمایش میتوان گرفت؟

پرهیز از Overfitting

- جلوگیری از رشد درخت قبل از رسیدن به مرحله ای که بطور کامل داده های آموزشی را دسته بندی نماید.
- اجازه به رشد کامل درخت و سپس هرس کردن شاخه هایی که مفید نیستند. (Post Pruning)
- در عمل روش دوم بیشتر استفاده شده است زیرا تخمین اندازه صحیح درخت کار ساده ای نیست.

نمونه از هرس کردن به روش Reduced Error Pruning

- ابتدا به درخت اجازه داده میشود تا به اندازه کافی رشد کند. سپس گره هایی را که باعث افزایش دقت دسته بندی نمیشوند هرس میگردند.

  - داده ها به دو مجموعه تست و آموزشی تقسیم میشوند.
  - درخت با داده های آموزشی مطابق روش قبل یاد گرفته میشود.
  - سپس برای یک گره داخلی غیر برگ n
  - زیر شاخه n حذف میگردد. این زیر شاخه با یک برگ جایگزین میشود. به این برگ دسته مثال های اکثریت یعنی دسته بندی اکثر مثال های قرار گرفته تحت این شاخه نسبت داده میشود.

- عملکرد درخت برروی مثال های تست بررسی میشود: اگر درخت هرس شده عملکرد بهتر و یا مساوی با درخت فعلی داشت از درخت هرس شده استفاده میشود.
- هرس کردن آنقدر ادامه می یابد تا هرس بیشتر، سودی نداشته باشد.

مثال بدست آوردن یک درخت تصمیم گیری

- یک جدول با چند متغیر در انواع مختلف

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

ایجاد درخت تصمیم گیری

- یک روش استفاده از روش شاخص جینی یا GINI Index است. در روش شاخص جینی همه متغیر ها در گره ها را امتحان می کنیم و آن متغیری که از همه کوچکتر باشد را انتخاب میکنیم.
- بهترین انتخاب برای تقسیم مجموعه S به دو مجموعه S1 و S2 اینست که رابطه زیر را ماکزیمم کنیم:

$$I(S)=|S1|/|S|*I(S1)+|S2|/|S|*I(S2)$$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

ایجاد درخت تصمیم گیری

- ابتدا جدول را بر اساس متغیر AGE به صورت صعودی مرتب میکنیم.

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

ایجاد درخت تصمیم گیری

- می خواهیم از روش شاخص جینی برای شاخه بندی استفاده کنیم.
- متغیر AGE یک متغیر مستقل عددی و متغیر CAR TYPE یک متغیر مستقل رده ای است.
- رابطه شاخص جینی:

$$Gini(T)=1-\sum pj^2$$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

- Age<=17

H: High

L: Low

R: Right Child

L: Left Child

	H	L
L	1	0
R	3	2

$I(S1): 1-(1/1)^2-(0/1)^2 = 1-1-0=0$

$I(S2): 1-(3/5)^2-(2/5)^2 = 1-9/25-4/25=0.48$

$|s1|=1, |s2|=5, |s|=6 \Rightarrow I(S): |1|/|6|*0+ |5|/|6|*0.48 = 0+ 5/6*0.48=0.4$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

- Age<=20

H: High

L: Low

R: Right Child

L: Left Child

	H	L
L	2	0
R	2	2

$I(S1): 1-(2/2)^2-(0/2)^2 = 1-1-0=0$

$I(S2): 1-(2/4)^2-(2/4)^2 = 1-4/16-4/16=0.5$

$|s1|=2, |s2|=4, |s|=6 \Rightarrow I(S): |2|/|6|*0+|4|/|6|*0.5= 4/6*0.5=0.33$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

- Age<=23

H: High

L: Low

R: Right Child

L: Left Child

	H	L
L	3	0
R	1	2

$I(S1): 1-(3/3)^2-(0/3)^2 = 1-1-0=0$

$I(S2): 1-(1/3)^2-(2/3)^2 = 1-1/9-4/9=0.444$

$|s1|=3, |s2|=3, |s|=6 \Rightarrow I(S): |3|/|6|*0+ |3|/|6|*0.444= 3/6*0.444=0.222$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

- Age<=32

H: High

L: Low

R: Right Child

L: Left Child

	H	L
L	3	1
R	1	1

$I(S1): 1-(3/4)^2-(1/4)^2 = 1-9/16-1/16=0.375$

$I(S2): 1-(1/2)^2-(1/2)^2 = 1-1/4-1/4=0.5$

$|s1|=4, \quad |s2|=2, |s|=6 \Rightarrow \quad I(S): \quad |4|/|6|*0.375 \quad + \quad |2|/|6|*0.5 \quad =$

$4/6*0.375+2/6*0.5=0.4166$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

- Age<=43

H: High

L: Low

R: Right Child

L: Left Child

	H	L
L	4	1
R	0	1

$I(S1): 1-(4/5)^2-(1/5)^2 = 1-16/25-1/25=0.32$

$I(S2): 1-(0/1)^2-(1/1)^2 = 1-0-1=0$

$|s1|=5, |s2|=1, |s|=6 \Rightarrow I(S): |5|/|6|*0.32 + |1|/|6|*0 = 5/6*0.32 +0=0.266$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High
32	Truck	Low
43	Sports	High
68	Family	Low

	H	L
L	4	2
R	0	0

- Age<=68

H: High

L: Low

R: Right Child

L: Left Child

$I(S1): 1-(4/6)^2-(2/6)^2 = 1-16/36-4/36=0.444$

$I(S2): 1-(0/0)^2-(0/0)^2 = 1-0-0=1$

$|s1|=6, |s2|=0, |s|=6 \Rightarrow I(S): |6|/|6|*0.32 + |0|/|6|*1 = 6/6*0.444 +0=0.444$

ایجاد درخت تصمیم گیری

- برای بررسی متغیر رده ای و برای آسانی کار جدول فراوانی هر کلاس را از روی همان جدول اولیه برای متغیر های رده ای تشکیل داده و سپس محاسبات را مشابه روش قبل انجام می دهیم.

Car Type	High	Low
sports	2	0
Family	2	1
Truck	0	1

ایجاد درخت تصمیم گیری

Car Type	High	Low
sports	2	0
Family	2	1
Truck	0	1

Car type={sports}

H: High

L: Low

R: Right Child

L: Left Child

	H	L
Car type = sports	2	0
Car type # sports	2	2

$I(S1): 1-(2/2)^2-(0/2)^2 = 1-1-0=0$

$I(S2): 1-(2/4)^2-(2/4)^2 = 1-4/16-4/16=0.5$

$|s1|=2, |s2|=4, |s|=6 \Rightarrow I(S): |2|/|6|*0 + |4|/|6|*0.5 = 4/6*0.5 +0=0.333$

ایجاد درخت تصمیم گیری

Car Type	High	Low
sports	2	0
Family	2	1
Truck	0	1

Car type={family}

H: High

L: Low

R: Right Child

L: Left Child

	H	L
Car type = family	2	1
Car type # family	2	1

$I(S1): 1-(2/3)^2-(1/3)^2 = 1-4/9-1/9=0.444$

$I(S2): 1-(2/3)^2-(1/3)^2 = 1-4/9-1/9=0.444$

$|s1|=3, |s2|=3, |s|=6 \Rightarrow I(S): |3|/|6|*0.444 + |3|/|6|*0.444 = 0.444$

ایجاد درخت تصمیم گیری

Car Type	High	Low
sports	2	0
Family	2	1
Truck	0	1

Car type={truck}

H: High

L: Low

R: Right Child

L: Left Child

	H	L
Car type = truck	0	1
Car type # truck	4	1

$I(S1): 1-(0/1)^2-(1/1)^2 = 1-0-1=0$

$I(S2): 1-(4/5)^2-(1/5)^2 = 1-16/25-1/25=0.32$

$|s1|=1, |s2|=5, |s|=6 \Rightarrow I(S): |1|/|6|*0 + |5|/|6|*0.32 = 5/6*0.32 +0=0.266$

ایجاد درخت تصمیم گیری

- پس از بررسی کلیه حالات می بینیم که حداقل مقدار I(S) بدست آمد:

Min{0.4,0.33,0.222,0.4166,0.266,0.444,0.303, 0.266,0.444}=0.222

- پس معیار AGE<=23 را به عنوان نقطه انشعاب انتخاب میکنیم.

Age<=23 = {17,20,23}

ایجاد درخت تصمیم گیری

Age<=23={17,20,23}

Age	Car Type	Risk
17	Sports	High
20	Family	High
23	Family	High

- کلاس در این شرایط برای همه حالات یکسان است.

ایجاد درخت تصمیم گیری

- حال جدول بالاتر از ۲۳ سال را تشکیل میدهیم و مثل روش قبل مثال را ادامه می دهیم تا معیار انشعاب تعیین شود:

Age	Car Type	Risk
32	Truck	Low
43	Sports	High
68	Family	Low

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
32	Truck	Low
43	Sports	High
68	Family	Low

Car type ={sports}

H: High

L: Low

R: Right Child

L: Left Child

	H	L
Car type = sports	1	0
Car type # sports	0	2

$$I(S1): 1-(1/1)^2-(0/1)^2 = 1\text{-}0\text{-}1=0$$

$$I(S2): 1-(0/2)^2-(2/2)^2 = 1\text{-}0\text{-}1=0$$

$$|s1|=1,\ |s2|=2, |s|=3 \Rightarrow I(S): |1|/|3|*0 + |2|/|3|*0 = 0$$

ایجاد درخت تصمیم گیری

Age	Car Type	Risk
32	Truck	Low
43	Sports	High
68	Family	Low

Car type ={truck}

H: High

L: Low

R: Right Child

L: Left Child

	H	L
Car type = truck	0	1
Car type # truck	1	1

$I(S1): 1-(0/1)^2-(1/1)^2 = 1-0-1=0$

$I(S2): 1-(1/2)^2-(1/2)^2 = 1-1/4-1/4=0.5$

$|s1|=1, |s2|=2, |s|=3 \Rightarrow I(S): |1|/|3|*0 + |2|/|3|*0.5 = 0.333$

ساختار درخت تصمیم گیری

Age	Car Type	Risk
32	Truck	Low
43	Sports	High
68	Family	Low

Car type={family}

H: High

L: Low

R: Right Child

L: Left Child

	H	L
Car type = family	0	1
Car type # family	1	1

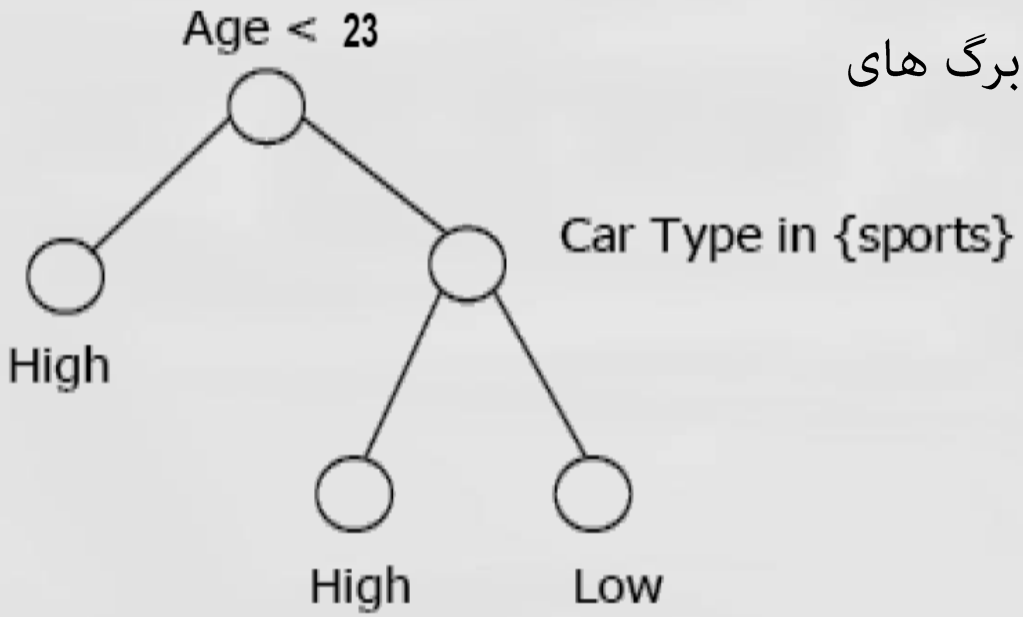
$I(S1): 1-(0/1)^2-(1/1)^2 = 1-0-1=0$

$I(S2): 1-(1/2)^2-(1/2)^2 = 1-1/4-1/4=0.5$

$|s1|=1, |s2|=2, |s|=3 \Rightarrow I(S): |1|/|3|*0 + |2|/|3|*0.5 = 0.333$

تکمیل درخت تصمیم گیری

- با بررسی تمامی حالات I(S) میبینیم که مینیمم آنها مقدار صفر است. پس درخت به شکل زیر تکمیل می گردد.
- این درخت نهایی است؛ چون تمام برگ های آن به کلاس ها ختم شده اند.



تمرین: در جدول زیر یک دسته داده های استاندارد از یک فروشگاه ***Mp3 Player*** جمع آوری شده است. هر نمونه توسط سه ویژگی نمایش داده شده است. عمل طبقه بندی بر اساس فیلد هدف ***"Customer Satisfaction"*** در دو دسته 'yes' و 'no' صورت می گیرد. درخت تصمیم گیری را جهت طبقه این دسته از داده ها به صورت دستی و بر اساس ***Information Gain*** و ***Gain Ratio*** ایجاد کنید. تشریح جزییات در هر مرحله الزامی است.

Memory	Battery life	Price	Customer satisfaction
<=4	long	<=150	yes
>4	long	>150	yes
>4	long	<=150	yes
<=4	long	>150	yes
>4	long	>150	yes
>4	low	>150	yes
<=4	low	>150	no
<=4	low	>150	no
>4	low	<=150	yes
<=4	low	<=150	no
<=4	medium	<=150	no
>4	medium	<=150	no
<=4	medium	>150	yes
>4	medium	>150	yes
>4	medium	<=150	no

مثال درخت تصمیم گیری

الف): ایجاد درخت تصمیم بر اساس Information Gain :

برای هر عضو درمجموعه S ، Gain را محاسبه کرده و با هم مقایسه می‌کنیم، هر عضویی که مقدار بیشتری داشت باشد به عنوان نود ریشه انتخاب می‌شود.

محاسبه گین memory:

$$s = [+9\,, -6]$$

$$value(memory) = \{\leq 4, > 4\}\ ,\ s_{\leq 4} = [3, -4]\ ,\ \ s_{>4} = [6, -2]$$

$$Entropy(s) = -(+)P\log_2 +p - (-)P\log_2 -p$$

$$Entropy(S) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15}$$

$$Entropy(S) = 0.442 + 0.528 = 0.97$$

$$Entropy(S) = 0.97$$

مثال درخت تصمیم گیری

$$Gain(S, Memory) = Entropy(S) - \sum_{v=\{\leq 4, >4\}}^{2} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Memory_{\leq 4}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$$

$$Entropy(Memory_{>4}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$$

$$Gain(S, Memory) = 0.97 - \frac{7}{15}\,0.985 - \frac{8}{15}0.811 = 0.0.87$$

$$\mathrm{Gain}(S, \mathrm{Memory}) = 0.087$$

مثال درخت تصمیم گیری

محاسبه گین battery life :

$$value(battery\ life) = \{long, low, medium\} \quad , \quad s_{long} = [5,0] \quad , s_{low} = [2,-3] \quad , s_{medium} = [2,-3]$$

$$Gain(S, Batterylife) = Entropy(S) - \sum_{v=\{long,medium,low\}}^{3} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Batterylife_{long}) = -\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5} = 0$$

$$Entropy(Batterylife_{medium}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

$$Entropy(Batterylife_{low}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

$$Gain(S, Batterylife) = 0.97 - 0 - \frac{5}{15}0.97 - \frac{5}{15}0.97 = 0.323$$

$$Gain(S, Batterylife) = 0.323$$

مثال درخت تصمیم گیری

محاسبه گین price :

$$value(price) = \{\leq 150, > 150\} \quad , s_{\leq 150} = [3, -4] \quad , s_{>150} = [6, -2]$$

$$Gain(S, Price) = Entropy(S) - \sum_{v=\{\leq 150, >150\}}^{2} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Price_{\leq 150}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$$

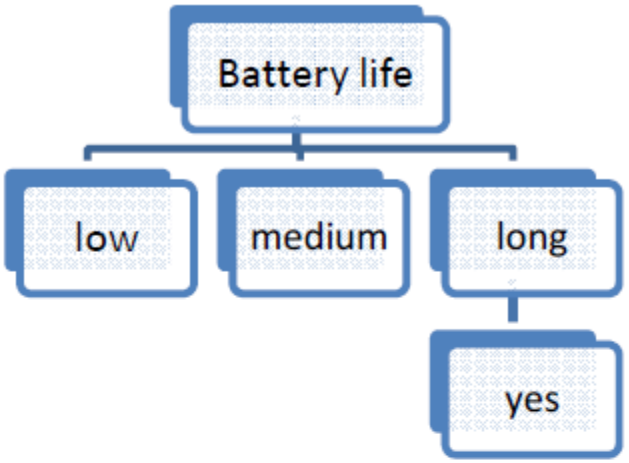
$$Entropy(Price_{>150}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$$

$$Gain(S, Price) = 0.97 - 0 - \frac{7}{15}0.97 - \frac{8}{15}0.97 = 0.087$$

$$Gain(S, Price) = 0.087$$

با توجه به سه فوق می‌توان battery life **به عنوان ریشه در نظر گرفت.**

مثال درخت تصمیم گیری



$$Gain(Low, Memory) = Entropy(Low) - \sum_{v=\{\leq 4, >4\}}^{2} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Memory_{\leq 4}) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$$Entropy(Memory_{>4}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

$$Gain(Low, Memory) = 0.97 - 0 - 0$$

$$Gain(Low, Memory) = 0.97$$

مثال درخت تصمیم گیری

$$Gain(Low, Price) = Entropy(Low) - \sum_{v=\{\leq 150, >150\}}^{2} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Price_{\leq 150}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

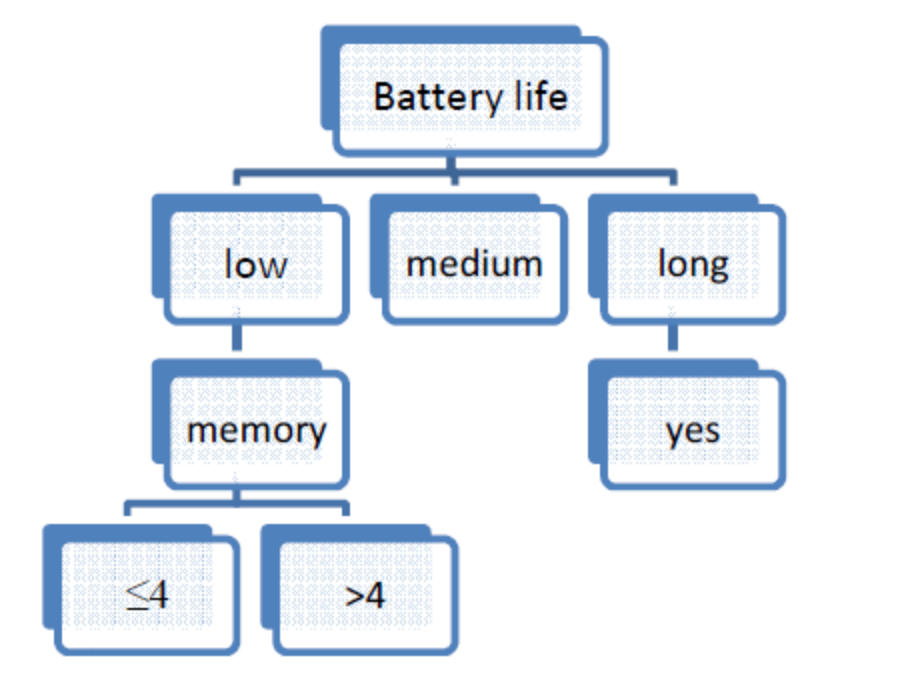
$$Entropy(Price_{>150}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{1}{3} = 0.922$$

$$Gain(Low, Price) = 0.97 - 0 - \frac{2}{5}1 - \frac{3}{5}0.922$$

$$gain(low, price) = 0.017$$

Memory را به عنوان فرزند Low انتخاب می‌کنیم.

مثال درخت تصمیم گیری



مثال درخت تصمیم گیری

$$Gain(Medium, Price) = Entropy(Medium) - \sum_{v=\{\leq 150, >150\}}^{2} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Price_{\leq 150}) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$$

$$Entropy(Price_{>150}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

$$Gain(Medium, Price) = 0.97 - 0$$

$$gain(Medium, price) = 0.97$$

$$Gain(Medium, Memory) = Entropy(Medium) - \sum_{v=\{\leq 4, >4\}}^{2} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(Memory_{\leq 4}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

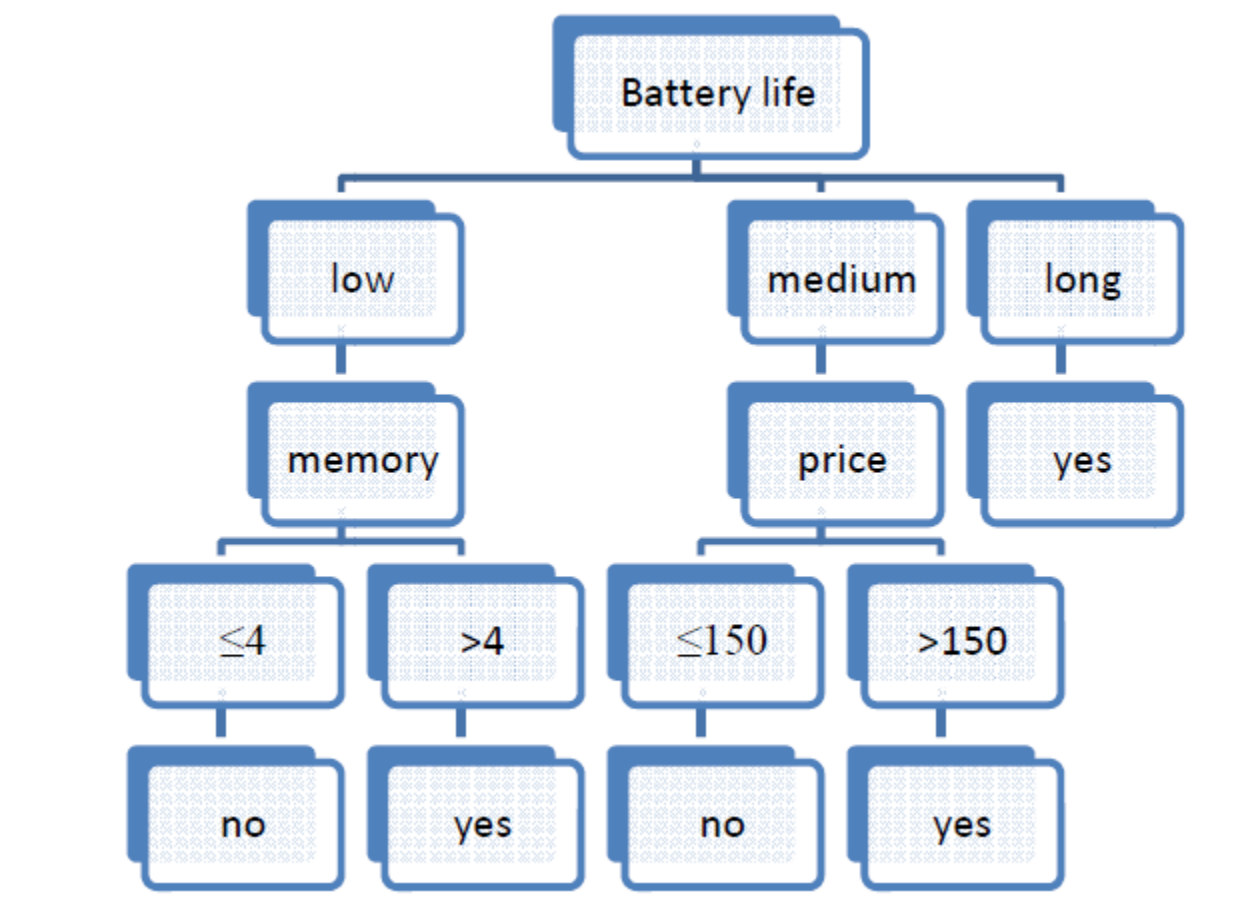
$$Entropy(Memory_{>4}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.922$$

$$Gain(Medium, Memory) = 0.97 - \frac{2}{5}1 - \frac{3}{5}0.922$$

$$gain(Medium, Memory) = 0.017$$

Price را به عنوان فرزند Medium در نظر می‌گیریم.

مثال درخت تصمیم گیری



مثال درخت تصمیم گیری

ب): ایجاد درخت تصمیم بر اساس GainRatio

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Spilit\ information(S,A)}$$

$$Gain(S,Memory) = 0.087$$

$$Spilit\ information(S,Memory) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$S_I(S,Memory) = -(\frac{7}{15}\log_2\frac{7}{15} - \frac{8}{15}\log_2\frac{8}{15}) = 1$$

$$Gain\ Ratio(S,Memory) = \frac{0.087}{1} = 0.087$$

مثال درخت تصمیم گیری

$$Gain(S, Price) = 0.087$$

$$Spilit\ information(S, Price) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$S_I(S, Price) = -(\frac{7}{15} \log_2 \frac{7}{15} - \frac{8}{15} \log_2 \frac{8}{15}) = 1$$

$$Gain\ Ratio(S, Price) = \frac{0.087}{1} = 0.087$$

$$Gain(S, BatteryLife) = 0.323$$

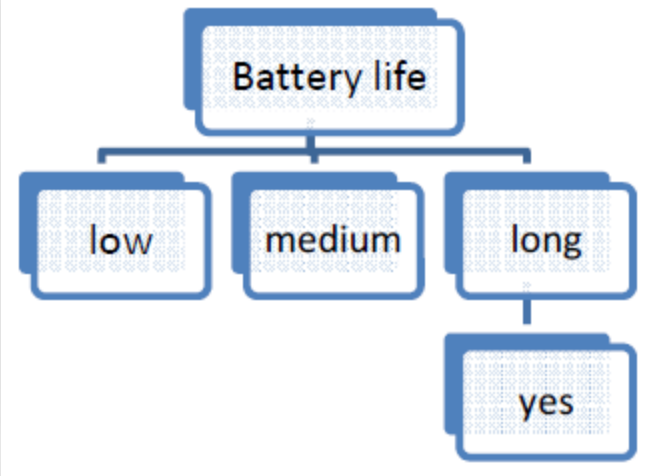
$$Spilit\ information(S, Price) = -\sum_{i=1}^{3} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$S_I(S, Price) = -3(\frac{5}{15} \log_2 \frac{5}{15}) = 1.58$$

$$Gain\ Ratio(S, BatteryLife) = \frac{0.087}{1.58} = 0.204$$

پس **battery life** به عنوان ریشه انتخاب می‌شود.

مثال درخت تصمیم گیری



مثال درخت تصمیم گیری

$$Gain(Low, Memory) = 0.97$$

$$Spilit\ information(Low, Memory) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$S_I(LOW, Memory) = 0.97$$

$$Gain\ Ratio(low, memory) = \frac{0.097}{0.97} = 1$$

$$Gain(Low, Price) = 0.017$$

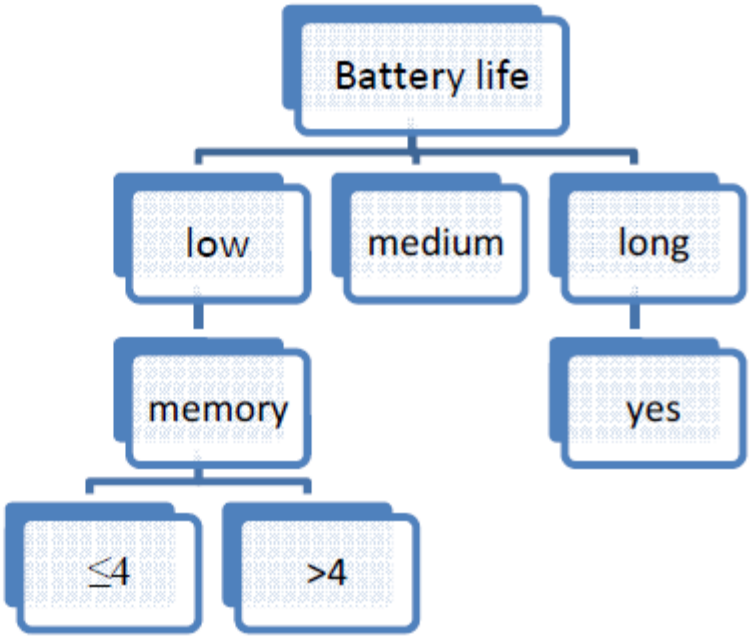
$$Spilit\ information(Low, Price) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$S_I(LOW, Price) = 0.97$$

$$Gain\ Ratio(low, price) = \frac{0.017}{0.97} = 0.0175$$

برای زیر شاخه **low** ، **memory** انتخاب می‌شود.

مثال درخت تصمیم گیری



$$Gain(Medium, Memory) = 0.017$$

$$Spilit\ information(Medium, Memory) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

مثال درخت تصمیم گیری

$$S_{I(Medium,Price)} = -\left(\frac{2}{5}\log_5\frac{2}{5} + \frac{3}{5}\log_5\frac{3}{5}\right) = 0.97$$

$$Gain\ Ratio(Medium, Memory) = \frac{0.017}{0.97} = 0.0175$$

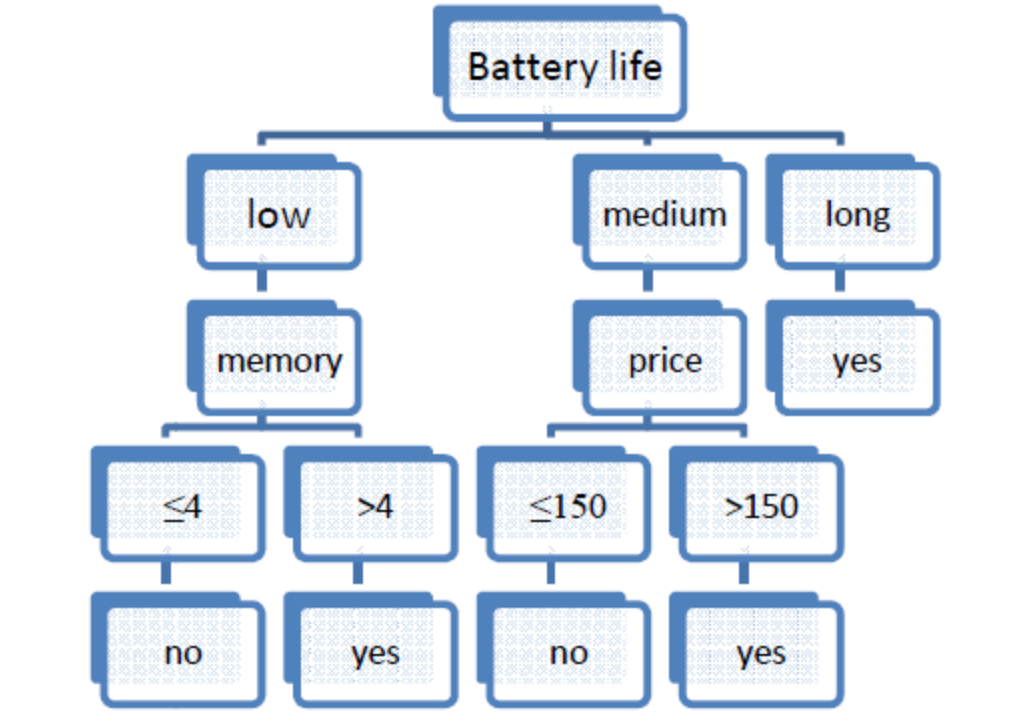
$$Gain(Medium, Price) = 0.97$$

$$Spilit\ information(Medium, Price) = -\sum_{i=1}^{2}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}$$

$$S_{I(Medium,Price)} = 0.97$$

$$Gain\ Ratio(Medium, price) = \frac{0.97}{0.97} = 1$$

مثال درخت تصمیم گیری



مثال درخت تصمیم گیری

مزایای درخت تصمیم گیری

- درخت تصمیم گیری به ما این قابلیت را می دهد که پیش بینی خود را بر اساس یک سری قوانین ارائه کنیم.
- درخت تصمیم گیری نیاز به محاسبات خیلی پیچیده ای برای دسته بندی داده ها ندارد.
- درخت تصمیم گیری برای انواع داده ها از جمله عددی و رده ای قابل استفاده است.
- درخت تصمیم گیری به ما نشان می دهد که کدام فیلد یا متغیر ها تأثیر مهمتری در پیش بینی و دسته بندی ما دارند.

معایب درخت تصمیم گیری

- برخی روش های دسته بندی در درخت تصمیم گیری برای کلاس های دوتایی و برخی دیگر برای کلاس های بیش از دوتا تنها جوابگو هستند.
- الگوریتم های درخت نیاز به حافظه بالایی دارند. وضعیت هر شاخه و هر فیلد به منظور بررسی نیاز است تا به خاطر سپرده شود.

نرم افزار هایی برای درخت تصمیم گیری

- WEKA, a free and open-source data mining suite, contains many decision tree algorithms
- Orange, a free data mining software suite, module orngTree
- KNIME
- Microsoft SQL Server

نرم افزار WEKA

- یک بستهء نرم افزاری است که حاوی تعداد زیادی از تکنیک های یادگیری ماشین و داده کاوی است که امکان مقایسهء تکنیک های یادگیری ماشین مختلف را می دهد.
- این نرم افزار دارای یک واسط کاربر گرافیکی می باشد که اجازه دسترسی به قابلیت هایی مثل تجسم و تحلیل بسیاری از الگوریتم های داده کاوی را می دهد.
- نرم افزار WEKA منبع باز بوده و به زبان جاوا است.

نرم افزار WEKA

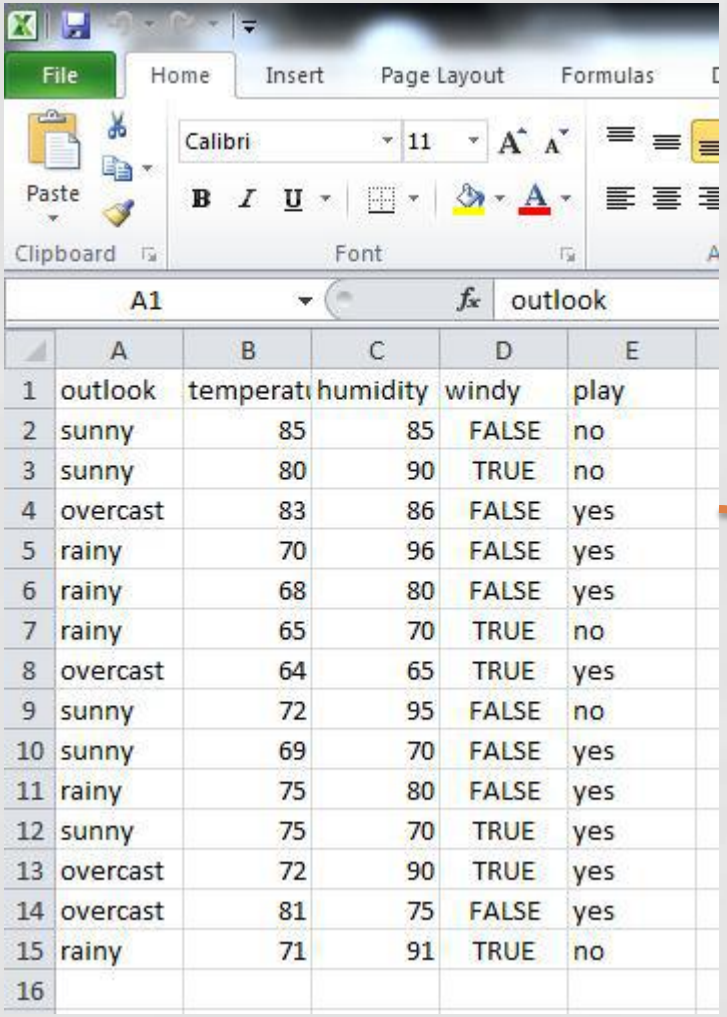
- بسته نرم افزاری WEKA حاوی پیاده سازی آخرین نسخه عمومی یاد گیرنده درخت تصمیم C4.5 بوده و درخت های تصمیم با تنظیم پارامترهای مشخص شده در آن به شکل خودکار ساخته شده و به شکل متن ASCII یا گرافیکی نمایش داده می شود.
- از WEKA میتوان در جاوا و VB.Net یا C# نیز استفاده کرد.

نرم افزار WEKA

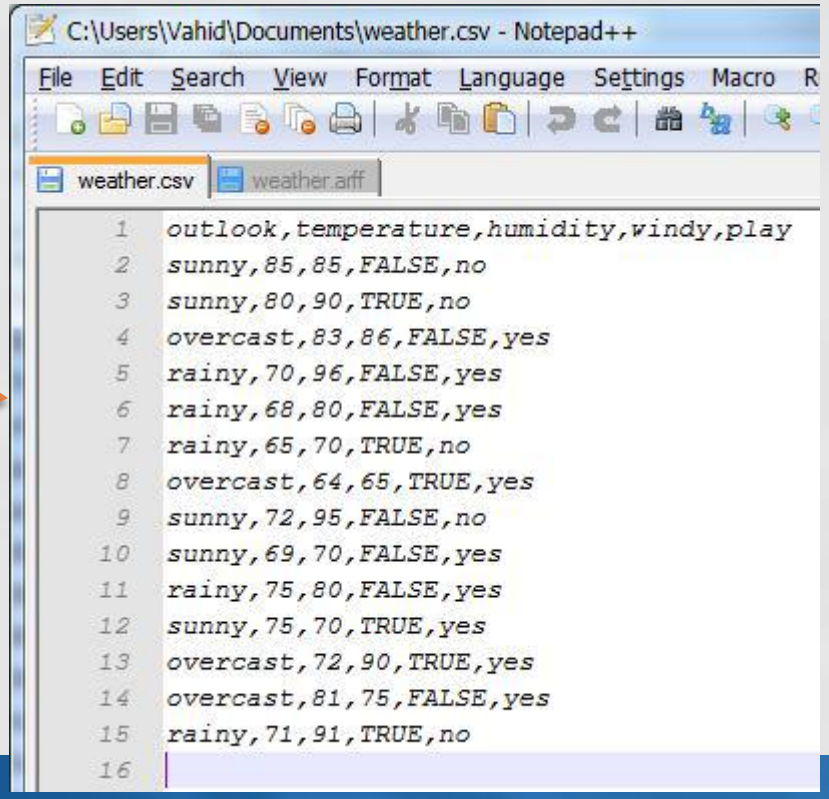
- بسته نرم افزاری WEKA حاوی پیاده سازی آخرین نسخه عمومی یاد گیرنده درخت تصمیم C4.5 بوده و درخت های تصمیم با تنظیم پارامترهای مشخص شده در آن به شکل خودکار ساخته شده و به شکل متن ASCII یا گرافیکی نمایش داده می شود.
- از WEKA میتوان در جاوا و VB.Net یا C# نیز استفاده کرد.

نرم افزار WEKA

- فرمت داده های قابل پذیرش
- CSV



	A	B	C	D	E
1	outlook	temperat	humidity	windy	play
2	sunny	85	85	FALSE	no
3	sunny	80	90	TRUE	no
4	overcast	83	86	FALSE	yes
5	rainy	70	96	FALSE	yes
6	rainy	68	80	FALSE	yes
7	rainy	65	70	TRUE	no
8	overcast	64	65	TRUE	yes
9	sunny	72	95	FALSE	no
10	sunny	69	70	FALSE	yes
11	rainy	75	80	FALSE	yes
12	sunny	75	70	TRUE	yes
13	overcast	72	90	TRUE	yes
14	overcast	81	75	FALSE	yes
15	rainy	71	91	TRUE	no
16					



```
outlook,temperature,humidity,windy,play
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```