باز شناسی الگو

جلسه هشتم

کلاسیفایرهای معمول- درخت تصمیم

فهرست مطالب

- مقدمه و یادآوری موقعیت بحث
- ایده اساسی درخت تصمیم
- بررسی یک مثال ابتدایی
- عملیات اصلی در ساخت درخت
- مشخصه ها Features
  - نوع آنها
  - تقسیم بندی آنها
  - ترتیب بررسی
- معرفی الگوریتم های معروف

مقدمه

- دیدیم که در SPR سه کار اساسی وجود دارد:
  - بازنمایی الگو (انتخاب نوع،استخراج و انتخاب مشخصه ها)
  - طراحی کلاسیفایر
  - ارزیابی کلاسیفایر
- دیدیم که این سه مرحله کاملا به هم وابسته بوده و به اصطلاح به یکدیگر فید بک وارد می سازند.
- روش های زیادی برای طراحی کلاسیفایر ارائه شده است که معروفترین آنها در این درس معرفی می شوند.
- دیدیم بطور کلی دو رویکرد مستقیم و غیر مستقیم برای طراحی کلاسیفایر قابل تصور است

روش های معمول طراحی کلاسیفایر

- دیدیم که طراحی کلاسیفایر عبارت بود از ارائه مسیری برای دستکاری داده ها، به گونه ای که در شکل دستکاری شده، تصمیم گیری برای انتساب نمونه ها به کلاس ها، آسان باشد.
- روش های معمول
  - طراحی بر اساس داشتن توزیع (کلاسیفایر Bayesian)
  - کلاسیفایر های خطی، درجه ۲ و..
  - درخت تصمیم
  - ماشین بردارهای پشتیبان
  - شبکه های عصبی
  - ....

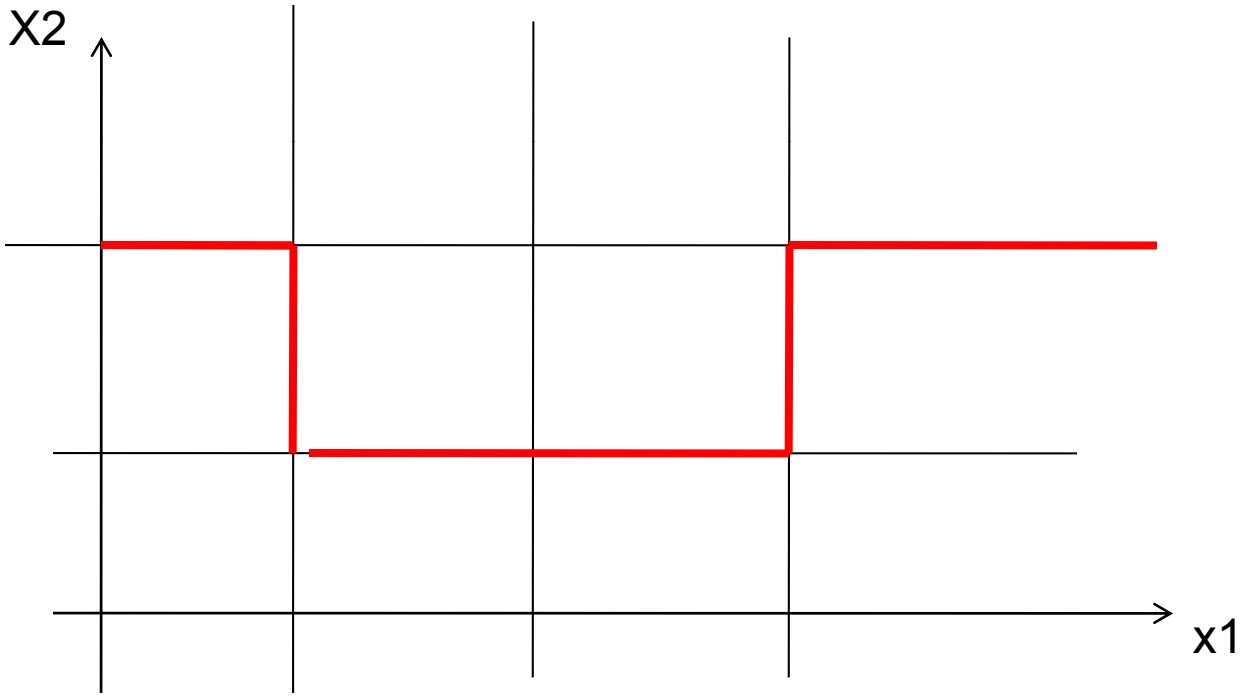
چگونگی دستکاری مشخصه ها

- قبلا گفته شد که در SPR عموما هر نمونه به صورت یک کلیت واحد در نظر گرفته شده و همه مشخصه های آن باید با هم و بطور همزمان مورد بررسی قرار گیرد.
- این امر، مشکلات نسبتا زیادی را بخصوص از نظر محاسبات و ملاحظات اجرایی بوجود میآورد.
- برای رفع این مشکلات، در روش های طراحی مختلف، پیش فرض های متفاوتی در نظر گرفته می شود.(مثل داشتن توزیع مشخص، جدائی پذیری خطی و ...)
- یک گزینه برای این کار، پیش فرض های مربوط به ارتباط و چگونگی مشخصه هاست.

پیش فرض های اساسی در درخت تصمیم

- حالتی را در نظر بگیریم که مشخصه های نمایش دهنده الگو دارای خواص زیر باشند:
  - اولا نسبت به هم مستقل باشند. به تعبیری که بتوان هر یک از مشخصه ها را با ثابت نگه داشتن بقیه، به دلخواه تغییر داد. به لحاظ هندسی، این مشخصه ها، موازی محور های اصلی فضا(که متعامد هستند) هستند.
  - مشخصه ها، گسسته هستند یعنی مقدیری که برای آنها در نظر گرفته می شود به اصطلاح کوانتیزه است.
- تحت این شرایط:
  - می توان مشخصه ها تک تک بررسی کرد
  - میتوان مبنای کلاسه بندی را به جای محاسبات ریاضی، به شکل رشته ای از تصمیم های منطقی در آورد.
- درخت تصمیم بر اساس ایده فوق شکل می گیرد.

فضای مشخصه ها



رویکرد ساختاری

- بدیهی است با تعدد تصمیم ها، این قابلیت وجود دارد که مرز تصمیم با هر دقتی بدست بیاید
- درخت های تصمیم، بطور بنیادی از رویکرد ساختاری استفاده می کنند. این امر به این معناست که یک مشخصه، مستقل از سایر مشخصه ها مورد بررسی قرار می گیرد.
- اما از آنجائیکه در بسیاری از دیگر موارد مانند انتخاب مشخصه و برآورد احتمال و سایر شاخص های آماری نیز کاربرد دارند، در SPR قرار می گیرند.

درخت تصمیم...

- این ایده ممکن است در هر یک از کاربرد ها با اصطلاحات متفاوتی تبیین شود اما از لحاظ مفهومی ، همه آنها یکی هستند .
- درخت تصمیم از انواع یادگیری بامربی برخوردار است.
- درخت تصمیم مسائل مختلف را با یک سری گرافها که دارای یک سری تصمیمها در بخشهای مختلف آن است مدل کرده و راه حلی برای این مسائل ارائه می کند.
- توجه به نکته ضروری است که قابلیت تبدیل دوباره مفاهیم منطقی به اعداد (ریاضی) وجود دارد. لذا حتی در مواردی که خروجی مورد نیاز چیزی غیر از نشانه lable کلاس هاست، باز هم این روش قابل استفاده است.

عملیات قابل تصور درخت بر روی اطلاعات

- **توصیف دیتاها** : کاهش حجم دیتاها با تبدیل کردن آنها به فرم فشرده تر تا حدی که مشخصات اصلی دیتاها حفظ شده و نتیجه دقیقی را بتوانیم داشته باشیم .
- **کلاسبندی** : یافتن اینکه هر یک از دیتاها در کدامیک از کلاسها قرار بگیرند.
- **آنالیزهای رگرسیون** : برای تقریب دیتاهایی که در آینده به سیستمی وارد خواهند شد.
- **تعمیم** : یافتن نگاشتی از یک سری از متغیرهای مستقل که برای پیش بینی متغیرهای وابسته بعدی از آنها استفاده شود .

حوزه کاربرد درخت تصمیم

- آمار
- بازشناسی الگوها
- تئوری تصمیم
- پردازش سیگنال
- یادگیری ماشین
- شبکه های عصبی

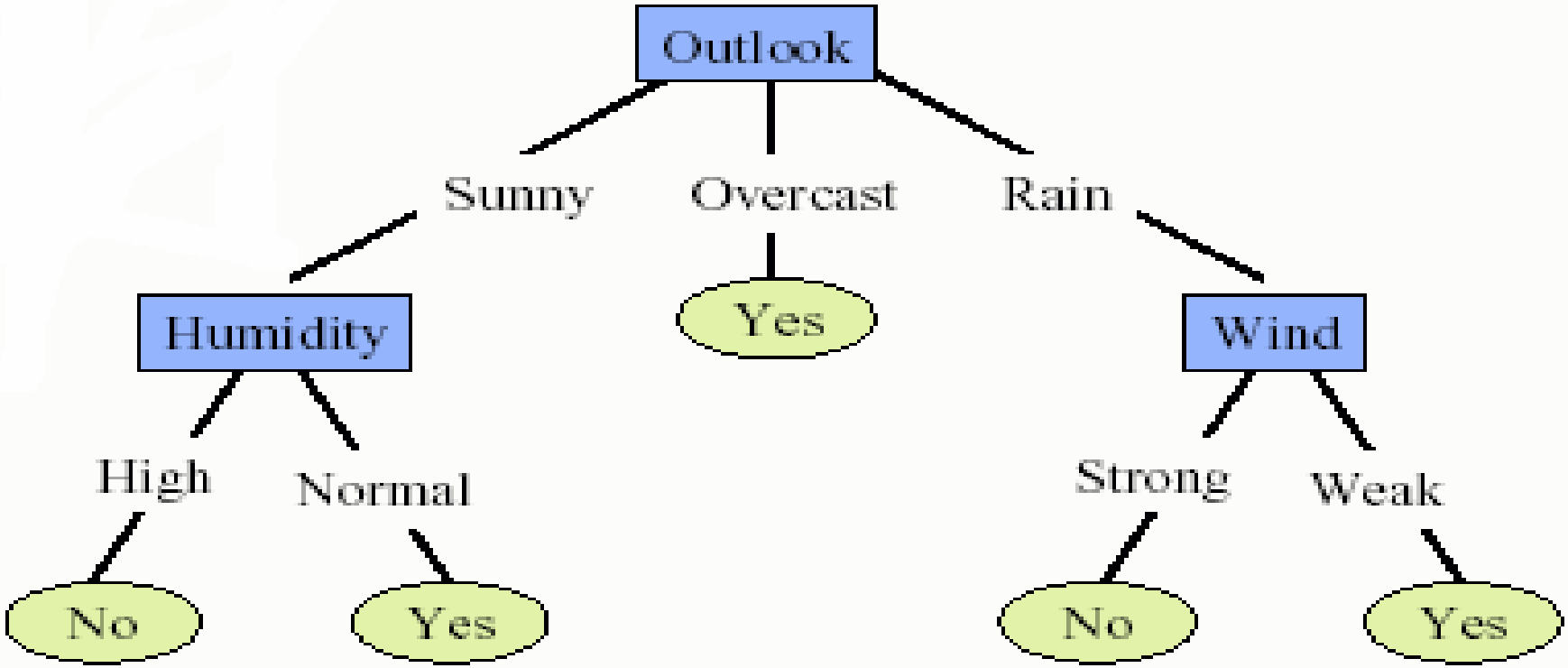
نمونه ای از یک درخت تصمیم

- فرض کنید قرار است بازی تنیس انجام شود و برای این کار شرایط جوی مهم است. ممکن است سوال این باشد که آیا بازی برگذار می شود یا نه
- پارامتر های تاثیر گذار بر پاسخ این سوال عبارتند از:
  - وضعیت هوا : آفتابی یا بارانی (دو حالت)
  - میزان رطوبت هوا: کم یا زیاد(دو حالت)
  - وزش باد: تند یا آرام (دو حالت)
  - دمای هوا :کم یا زیاد(دو حالت)
- به تعبیری در حالت کلی، برای ۴ فاکتور دو حالتی، ۱۶ وضعیت مختلف وجود دارد که در هر یک از این وضعیت ها به نحو معینی به سوال اولیه پاسخ داده می شود.
- ممکن است برخی موارد فاکتور هائی دارای چند حالت باشند.
- البته برخی از حالت ها به دلایل منطقی و موضوعی بایستی حذف شوند

نمونه داده ها برای ساخت یک درخت

Day-sample	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

نمونه ای از یک درخت تصمیم



درخت تصمیم

- این دیاگرامها مثل درخت وارونه هستند . از یک ریشه شروع می شوند و دارای شاخه هایی می شوند که به انتهای آن ختم می شوند .
- بیان درخت به صورت ترکیبی از پارامترهای منطقی است:
  - زمانی که هوا آفتابی و رطوبت هوا نرمال باشد بازی تنیس برگزار می شود.
- باید توجه داشت که:
  - انتهای شاخه ها منجر به تصمیمها(کلاس ها) می شود.
  - یک تصمیم (کلاس) خاص می تواند در انتهای چندین شاخه دیده شود

اساس درخت تصمیم

- طرح مجموعه ای از سوالات به صورت مرحله به مرحله و در هر مرحله بر اساس مشخصه ای خاص
- تقسیم شدن دیتاهای مختلف با توجه به پاسخ به این سوالات
- ادامه این سوالات تا جائیکه همه دیتاها به درستی تقسیم شوند.
- انتخاب یا تشخیص این سوالات، توسط بخشهای آموزش که شامل یک سری نمونه ها و دیتاها می باشند ساخته می شوند .
  - هر نمونه با یک سری توصیفها و برچسب یک کلاس مشخص می شود .

اساس درخت تصمیم

- نگاشت درست بین صفتها و کلاسها
- دارای حداقل یک گره ترمینال یا Leaf و تعدادی گره داخلی یا بدون گره داخلی
- تمام گره های داخلی دارای حداقل دو یا بیشتر بچه گره می باشند .

- کلاسبندی نمونه X طبق الگوریتمهای فوق

تجزیه و تحلیل نکات:

- درخت تصمیم در واقع از رویکرد ساختاری در بازشناسی استفاده می کند.
  - نمایش یک نمونه قادر به شکسته شدن به اجزائی است که مستقل از هم قابل تجزیه و تحلیل هستند.
  - مشخصه ها و اجزا از هم استقلال خطی دارند.
  - مرزهای تصمیم بوجود آمده توسط درخت تصمیم، قادر به ساخته شده توسط خطوط قائم و افقی هستند.

درخت تصمیم

- درخت تک متغیره : همه مشخصه ها را می توان به یک پارامتر وابسته نمود.
- درخت چند متغیره: در این ساختار پارامتر های مستقل از هم وجود دارد(نمی توان همه مشخصه ها را بر اساس یک پارامتر بیان کرد.
- فرآیند تشکیل درخت(سلسله ای از سوال ها) تقریبا مشخص است بنابر این مهمترین کار در ساخت یک درخت تصمیم عبارتست از: یافتن مشخصه (مشخصه های) مناسب برای تقسیم

تاریخچه الگوریتم های ساخت (طرح) درخت تصمیم

- درخت CLS : Hunt ; ۱۹۶۶
- درخت CART (درخت رگرسیون و طبقه بندی) : Beriman ;۱۹۸۴
- درخت ID3 : Quinlan ;۱۹۸۶
- درخت C4.5 : Quinlan ; ۱۹۹۳

- البته از لحاظ اصول کار ، همه این درختان شبیه به هم هستند.

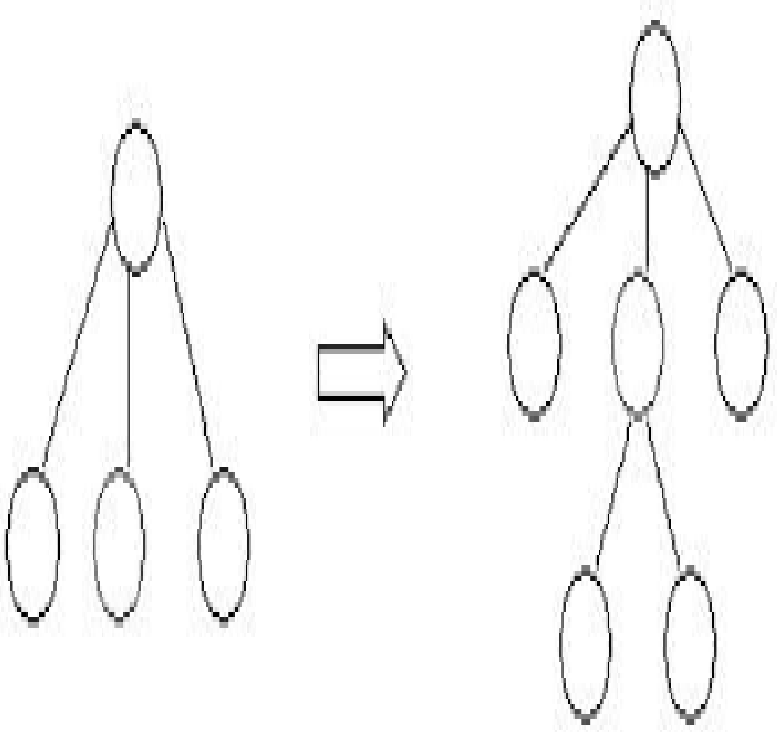
کارکرد های پایه در ساخت درختها

- در یک مسئله عملی، کلیه این عملیات بایستی با استفاده از تعدادی از نمونه ها(مثال ها) و با استفاده از کارکرد های زیر صورت گیرد:
  - انشعاب(عملیات رشد یک گره)
  - هرس کردن(حذف شاخه ها و گره ها)
  - تعیین برچسبهای گره های Leaf
  - انتخاب مشخصه مناسب

انشعاب (تقسیم)

- فرض کنیم یک گره از درخت، دارای یک سری خصوصیات Xj می باشد .
- فرض کنید رنج تغییرات این خصوصیات ، به زیرمجموعه های Lj تقسیم شود(که مجمو عه ای گسسته است).
- می توانیم توسط هر یک از این زیرمجموعه ها یک شاخه از این گره خارج کنیم و این شاخه به یک گره جدید ختم شود که به آن بچه گره نیز گفته می شود.
- بنابراین گره قبلی را به Lj گره جدید تقسیم کرده ایم .

انشعاب (تقسیم)

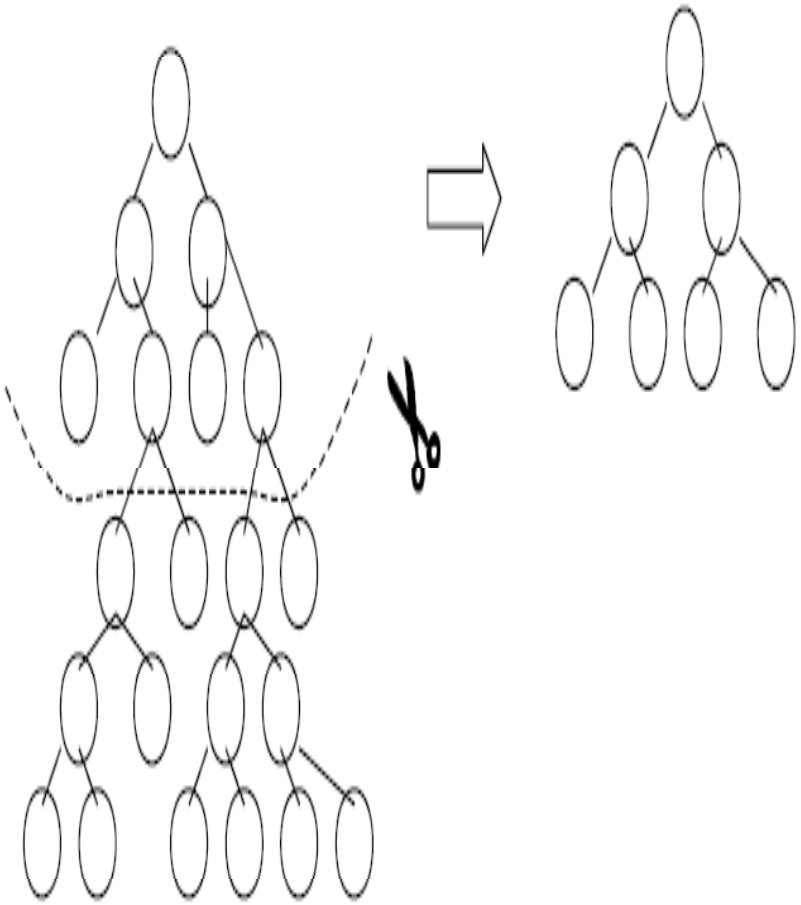


- تعداد Lj برابر با ۲:
- درخت باینری

- تعداد Lj برابر با ۳ :
- درخت ، Ternary

- تعداد Lj برابر با ۴ :
- درخت درجه دوم یا Quadratic

هرس کردن



- این ایده توسط Beriman ارائه شد.
- این عملیات دقیقا متضاد عمل رشد درخت می باشد .
- فلسفه اصلی آن است که برخی از گره های جدید عملا و یا منطقا یا وجود ندارند و یا ارزش بررسی ندارند.
- هرس کردن ، با کنترل میزان پیچیدگی درخت، برای حصول یک درخت با اندازه مناسب استفاده می شود .

هرس کردن

۱- استفاده از نمونه های تقسیم شده:

در این روش نمونه ها را به دو قسمت آموزش و ارزیابی تقسیم می کنیم و پس از ساختن درخت با نمونه های آموزش آن را با نمونه های ارزیابی هرس می کنیم.

۲-استفاده از تمام نمونه ها برای آموزش:

از تمام نمونه ها برای آموزش درخت استفاده می کنیم و سپس با استفاده از روشهای آماری تعیین می کنیم که کدام گره ها احتمالا سبب بهبود درخت می شوند و سایر گره ها را هرس می کنیم.

تعیین برچسبهای گره های Leaf

- آسانترین مرحله ساختن درخت
- زمانی که گره ها ، تا جای ممکن تقسیم شدند و هر گره Leaf ، به یک کلاس از الگوها تعلق داشت (به طور مستقیم به یک تصمیم منجر شد) ، این کلاس را به عنوان برچسب برای این گره Leaf ، در نظر می گیریم

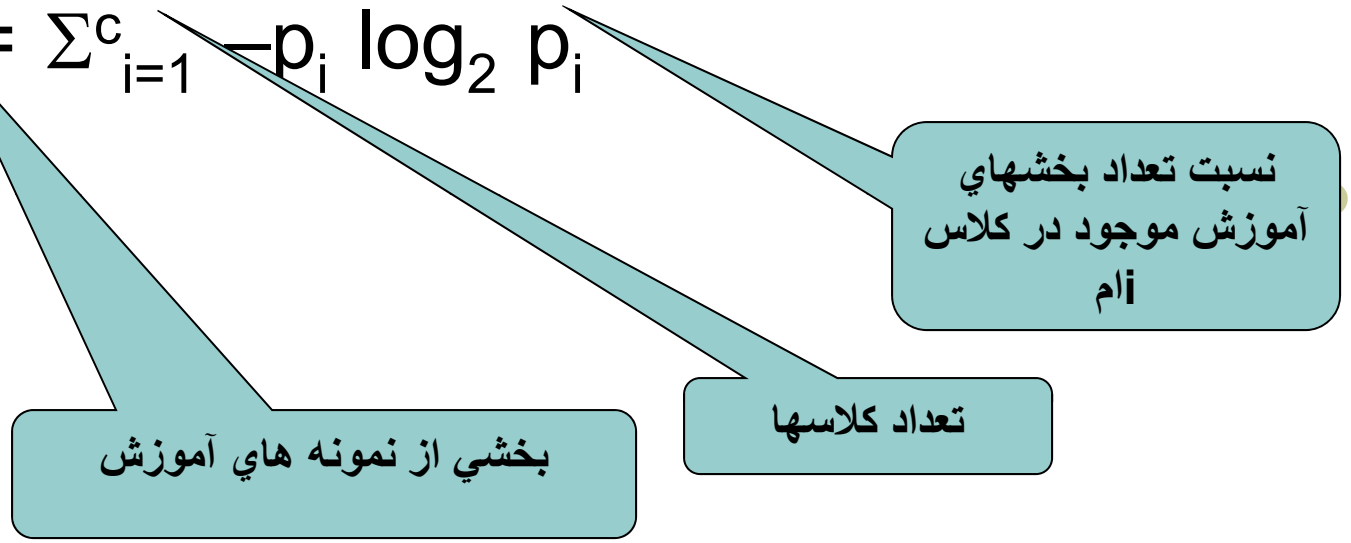
- در حالتهایی که بیشتر معمول است گره به دست آمده ناخالص می باشد.
  - در این حالت، عنوان برچسب را به کلاسی منسوب می کنیم که آن گره دارای نمونه های بیشتری در آن کلاس باشد
  - نزدیکی بیشتری به آن کلاس داشته باشد .

انتخاب مشخصه مناسب

- می خواهیم کوچکترین درخت را به دست آوریم.
- بنابراین باید مشخصه ای را انتخاب کنیم که توسط آن گره های خالصتری بوجود بیاید.
- می توانیم از معیاری به نام انتروپی برای این کاراستفاده کنیم.
  - پس از محاسبه مقدار انتروپی ، از پارامتری به نام Information Gain استفاده می شود.
- این دو پارامتر وابستگی زیادی با هم دارند: هر چقدر مقدار انتروپی بیشتر باشد به اطلاعات بیشتری برای تخمین آن مجموعه نیاز داریم.

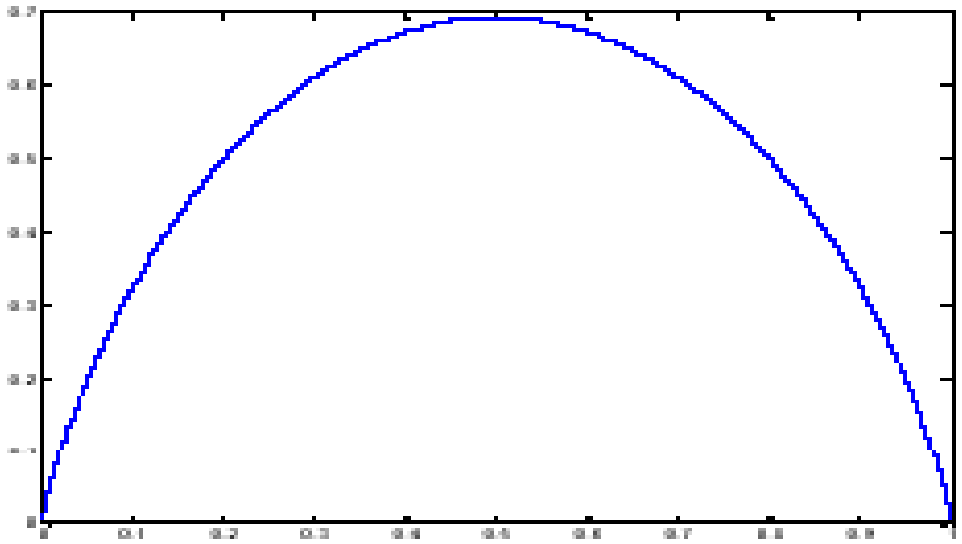
انتخاب مشخصه مناسب

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$



انتخاب مشخصه مناسب

Example for k=2



- Information Gain :
- $G(S,A) \equiv E(S) - \Sigma_{v \text{ in Values}(A)} (|S_v|/|S|) * E(Sv)$

انتخاب مشخصه مناسب

- برای درخت گفته شده در ابتدای این مبحث می توانیم مقدار انتروپی را به صورت زیر محاسبه نماییم:

$$E(s)=E(9+,5-)=(-9/14 \log_2 9/14) + (-5/14 \log_2 5/14)=0.94$$

تعداد Yes

تعداد No

انتخاب مشخصه مناسب

- حال می توانیم Information Gain را برای صفات این درخت یعنی {باد، دما ، وضعیت هوا و رطوبت} تعیین کنیم information gain برای وضعیت هوا به صورت زیر تعیین می شود:

G(S,Outlook) = E(S) – [5/14 * E(Outlook=sunny) + 4/14 * E(Outlook = overcast) + 5/14 * E(Outlook=rain)] =

E([9+,5-]) – [5/14*E(2+,3-) + 4/14*E([4+,0-]) + 5/14*E([3+,2-])]=

0.94 – [5/14*0.971 + 4/14*0.0 + 5/14*0.971]

G(S,Outlook) = 0.246

انتخاب مشخصه مناسب

- مقدار Information Gain برای سایر مشخصه ها به صورت زیر قابل محاسبه اند:

- **G(S,Temperature) = 0.029**
- **G(S,Humidity) = 0.1515**
- **G(S,Wind) = 0.048**
- **G(S,Outlook) = 0.246**

اولين مشخصه براي تقسيم

ساخت درخت

- انتخاب صفت مناسب
- یافتن انشعاب مناسب برای آن مشخصه
- تقسیم نمونه ها به زیرمجموعه هایی تقسیم که برای نود هر یک از این شاخه ها مناسب باشند.
- انتخاب بهترین صفت از بین صفات باقیمانده که بتواند بهترین تقسیم را بوجود آورد
- تکرار این کار برای نمونه هایی که به هر یک از نودها رسیده اند

ساخت درخت

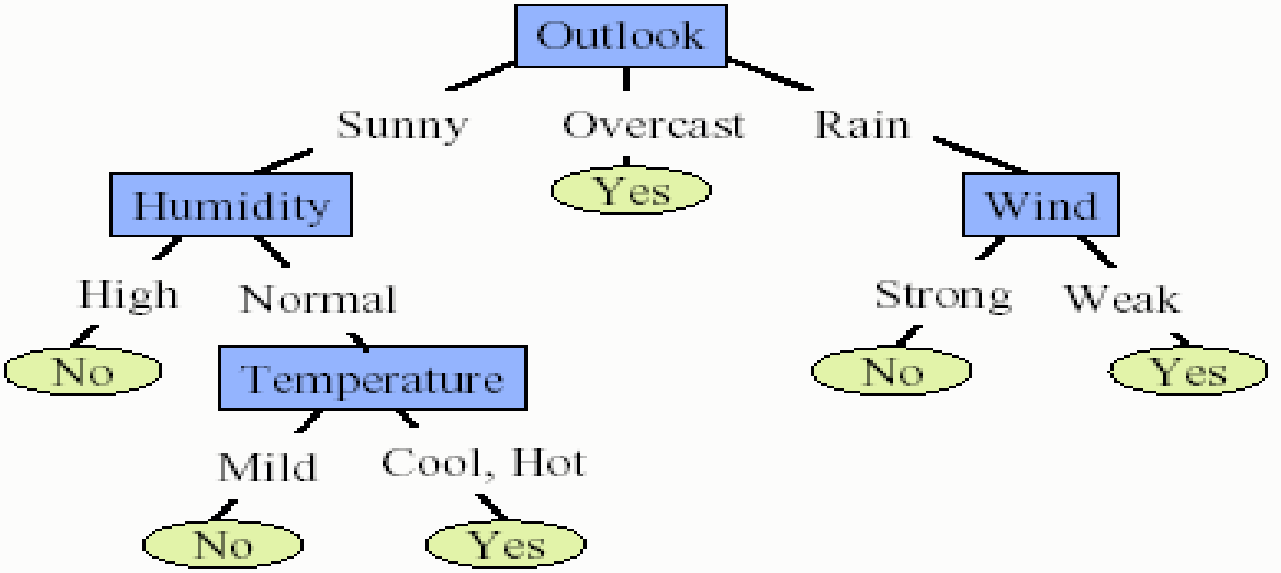
زمانی کار به پایان می رسد که:

۱-تمام نمونه ها مربوط به یک کلاس شوند

۲-هیچ مشخصه ای برای تقسیم بیشتر وجود نداشته باشد

۳-هیچ نمونه ای باقی نمانده باشد

تصمیم گیری از روی درخت



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D15	Sunny	Mild	Normal	Strong	No

انتخاب درخت بهینه

- باید راههای گوناگون مشخص کردن پارامترهای توصیف کننده کیفیت درخت را بررسی کنیم. می توانیم مقدار کارایی را روی نمونه های آموزش یا ارزیابی پیدا کنیم
- می توانیم میزان خطای مورد انتظار پیش بینی ها را به عنوان پارامتر اساسی در نظر بگیریم .
- اندازه گیری میزان این خطا : می توانیم با استفاده از مشاهدات ، کیفیت درخت را به صورت تقریبی تخمین بزنیم .

پارامترهای کیفیت یک درخت

- معمولا در ساختار یک درخت ، درختی را ترجیح می دهیم که ساده تر و فشرده تر با تعداد گره های کمتری باشد.

دو نوع پارامتر اساسی برای توصیف کیفیت یک درخت می توان در نظر گرفت.

- پارامترهای مربوط به دقت درخت
- پارامترهای مربوط به پیچیدگی درخت

پارامترهای دقت درخت

- در مسائل بازشناسی ، به کمک نمونه ها و مشخص کردن اینکه چقدر نمونه های کلاسهای مختلف به خوبی تقسیم شده اند .

- در مسائل مربوط به آنالیزهای رگرسیون ، با استفاده از این مفهوم که خطای پیشگویی چقدر است

پارامترهای پیچیدگی یک درخت

- این پارامتر ها، ناشی از شکل درخت بوده و به نمونه های مورد استفاده در آن ربطی ندارد .

پارامترهای پیچیدگی یک درخت ، عبارتند از :

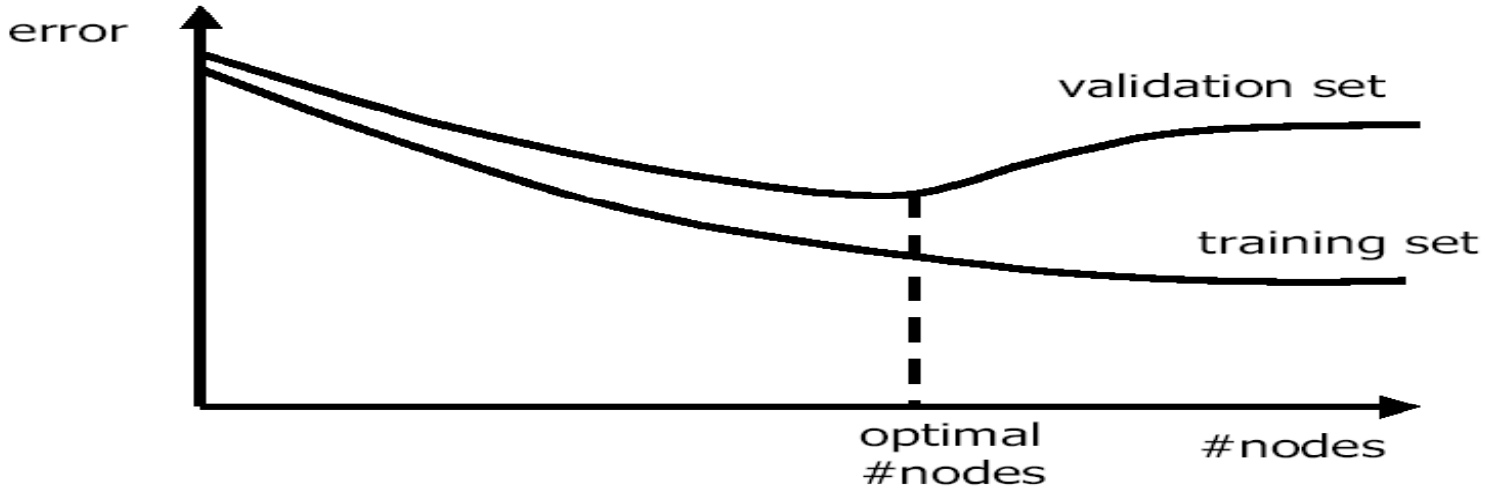
- تعداد Leafهای یک درخت
- تعداد گره های داخلی درخت
- ماکزیمم طول یک مسیر از ریشه تا Leaf

وابستگی پارامترهای دقت و پیچیدگی

- معمولا درختان خیلی پیچیده دارای دقت بسیار بالایی هستند . ماکزیمم دقت زمانی به دست می آید که برای هر یک از نمونه ها یک Leaf درنظر گرفته شود .
- اما درختان با پیچیدگی کمتر، ترجیح داده می شوند . چون مدل کردن این درختان و تفسیر آنها آسانتر و نیز قدرت تعمیم آنها بهتر می باشد .
- در نمونه های با تعداد کم در مقایسه با تعداد مشخصه ها ، یک درخت خیلی پیچیده دارای ناپایداری می شود یعنی برای مشاهدات جدید ، خطای آن خیلی بالا می رود. از طرف دیگر ، کاملا واضح است که درختان خیلی ساده نیز نمی توانند پیش بینی خوبی را از پدیده ها ارائه دهند .
- بنابراین در طراحی یک درخت خوب ، باید یک توافق بین دقت و پیچیدگی درخت ایجاد کرد.

Over fitting

- این مساله زمانی رخ می دهد که ما در زمان آموزش ، نمونه ها را به خوبی تقسیم و کلاسبندی کنیم اما در زمان آزمایش دارای خطای زیادی می شویم.



Over fitting

راه حل:

- ۱- بعضی از انشعابهای درخت ساخته شده در مرحله قبل را با کمک گرفتن از نمونه های تست (ارزیابی) هرس کنیم
- ۲-در زمان ساخت درخت آن را برای احتمال رخ دادن over fit تست کنیم و چنانچه این مساله به وجود آمد ، در همانجا ساخت درخت را متوقف کنیم.

تفاوت درخت ID3 ، CART و C4.5

- درخت CART و C4.5 و ID3تفاوتهایی با هم دارند که برخی از آنها عبارتند از :
  - تعداد تقسیمها در درخت CART همیشه باینری است اما در ID3 و C4.5 هر تعدادی می تواند باشد.
  - معیار انتخاب مشخصه مناسب در درخت CART ملاک Gini می باشد در حالی که در C4.5 ملاک gain Ratio و در ID3 ، Information Gain می باشد.
  - در روش هرس کردن نیز این درختان تفاوتهایی با هم دارند.

مزایا و معایب

- درخت تصمیم دارای این خاصیت است که تمام درختان را می توان به صورت درخت باینری تبدیل کرد.
- هم برای دیتاهای عددی و هم برای دیتاهای غیرعددی قابل پیاده سازی است.
- به راحتی می توان آنها را با ورودی تطبیق داد.
- فهم و پیاده سازی آنها راحت است.

اما

- یافتن مشخصه ها برای دیتاهایی که در دنیای واقعی موجودند کار مشکلی است.
- تغییر مختصری در ورودی می تواند تغییرات زیادی در آرایش درخت و در نتیجه به دست آمده ، بوجود آورد.

خلاصه

- درخت تصمیم از انواع روش های مستقیم طبقه بندی با استفاده از داده هاست.
- این روش اساسا یک روش ساختاری است.
- عملیات مهم آن عبارتند از: انتخاب مشخصه ها، انتخاب برگ هاؤ عملیات تقسیم و عملیات هرس کردن
- الگوریتم های متفاوتی برای طرح درخت تصمیم از روی داده های با مربی وجود دارد. **ID3 ، CART و C4.5**
- معمولترین روش های طبقه بندی درخت ها از روی شیوه تقسیم است(باینری و غیر باینری)