**مفهوم درخت تصمیم**  
  
  
درخت تصمیم یکی از ابزارها و تکنیک‌هایی است که در مهارت‌های داده‌ کاوی بسیار پر کاربرد است. زمانی که حجم داده‌ها بسیار بالا باشد این تکنیک می‌تواند به کمک شما بیاید .   
یکی از روش های کاربردی در یادگیری ماشین می باشد  
کاری که برای ما انجام می دهد این است که از روی مثال های آموزشی به ما یک فرضیه ای را می دهد که ما انتظار داریم این فرضیه روی بقیه داده ها نیز درست عمل کند  
در درخت تصمیم اهمیت و ویژگی ها و جایگاه ویژگی ها بسیار مهم است.  
در نود های درخت ویژگی ها می آید و در برگ های آن نیز تصمیم ها است.  
در درخت تصمیم، تصمیم گیری ارزش ویژگی ها خیلی مهم است علی الخصوص برای فیلتر کردن و به نتیجه رسیدن.   
  
**ویژگی ها**  
  
1.حساسیت درخت تصمیم به نویز زیاد می باشد  
2.کسب دانش در درخت تصمیم به صورت تدریجی نیست

یکی از روش های شناخته شده و کاربردی تو یادگیری ماشین است. کاری که در درخت تصمیم برای ما میکند این هست که از روی مثالای آموزشی فرضیه میسازد که انتظار داریم این فرضیه نه تنها روی مثالای اموزشی درست کار کند بلکه روی داده هایی که بعدا میبیند نیز درست جواب دهد یعنی جواب نزدیک به تابع هدف بدهد. این فرضیه با طرز فکر انسان نزدیک است و با منطق انسان سازگار هست. فرضیه در این روش به شکل درخت است که نودها ویژگی ها و برگ ها نتیجه گیری میشیند. ویژگی ای که توی ریشه درخت مینشیند اهمیتش از همه بالاتر است و بر اساس اهمیت کم کم درختو کامل میکنیم. جایگاه ویژگی ها در درخت در موفقیت فرضیه ها تاثیر دارد. درواقع با دسته بندی ویژگی ها بر اساس مثال ها درخت ساخته میشود. در نهایت انتخاب ویژگی ها بر اساس اهمیت باعث میشود عمق درخت کم شود. اگر ویژگی ها حالت باینری داشته باشد ویژگی بهتر بر اساس کمترین خطا انتخاب میشود و دو فرضیه برای محاسبه خطای هر ویژگی در نظر گرفته میشود و فرضیه با خطای کمتر فرضیه بهتری بوده و خطای ان به عنوان خطای ویژگی تلقی میشود. در نهایت خطای دو ویژگی محاسبه و مقایسه شده و بهتره در ریشه درخت مینشیند و اگه برای حالتی از ویژگی خطا صفر شود برای ان حالت نتیجه نوشته میشود به عنوان برگ و سپس از سایر ویژگی ها به همین ترتیب در صورت عدم قطعیت یا صفر نشدن خطا برای حالتی از ویژگی، کم ترین خطایشان به ازای ان حالت محاسبه و مجددا کم خطا ترین ویژگی به عنوان گره بعدی انتخاب میشود. در نهایت ممکنه ۲ فرضیه مثلا ایجاد شود که هر دو فرضیه روی مثالای اموزشی را پوشش میدهند اما روی داده های جدید ممکن است جوابای متفاوت دهند. جواب ها در صورت یکی شدن ما را مطمین میکند ولی اگر جواب فرضیه ها یکسان نبود متوجه میشویم باید ویژگی بهتری را در کل دخیل کنیم. اگر هم فقط یک فرضیه بدست بیایذ که جواب قطعی است و البته اگر فرضیه ها چندتا باشن باید بینشون رای گیری کرد و بر اساس اکثریت انتخاب شود.

نکات: درخت یا همون فرضیه لزوما متقارن نمیشه. اگر خطا صفر بشه برای حالت باینری درختو برای ان حالت ادامه نمیدهیم وگرنه سراغ محاسبه سایر ویژگی ها برای ادامه کار میرویم. درخت تصمیم باینری به نویز حساس هست، اگر نویز باشد باید از اول درخت را بسازیم. کسب دانش در درخت تصمیم تدریججی نیست که ویژگی بدی است. به همین دلیل باید با امدن مثالای جدید درخت را از اول ساخت. درخت تصمیم برای حالات پیوسته مناسب نیست مثل قد و کلی حالت دارد که مثال به قدر کافی یافت نمیشود. فرضیه زیاد بشه بد نیست اما ممکنه قطعیت رو بگیره. ساخت درخت تصمیم بر اساس انتخاب ویگی مهمتر یا خطای کمتر در حالت باینری با فرض هزینه مساوی برای ویژگی ها هستچون ممکنه ویژگی ای هزینه بالا داشته باشد اما کمترین خطا را بدهد و دقت خوبی داشته باشد اما به کار ما نمیخورد. دانش کم و مثالای کم ویژگی های مهمتر انتخاب میشومد و شیوه گذاشتن ویژگی ها به ترتیب اولویت مانند حالت باینری است فقط معیار انتخاب متفاوته. کم میشود فرضسه خوبی ساخت که روی مثالای جدید هم خوب جواب دهد و جواب نزدیکی با تارگت دهد. اگر ویژگی ای یک حالت داشت از مجموعه ویژگی ها حذف کرده و باقی را محاسبه میکنیم تا کمترین خطا یافت شود. اگر تابعی باینری نباشد و مثلا تصمیم گیری سه گانه باشد میتوان ان را به تصمیم گیری های دوگانه شکست.

در صورتی که حتی یک ویژگی بیش از دو حالت داشت از روش محاسبهgain ویژگی مهمتر انتخاب میشود و باقی مراحل ساخت درخت مانند حالت باینری بوده و فقط تفاوت در معیار انتخاب ویژگی مهمتر است که با بهره بیشتر و یا انتروپی کمتر ویژگی بهتر انتخاب میشود. البته با این روش درخت تصمیم id3 نامیده میشود. در این روشم درختای با عمق کمتر پیشفرض شکل فرضیه inductive biase هست چون قابل تعمیم تر هست و ساده تر و عمومی تر است. البته در این روش درخت کمتر به نویز حساسه چون اماری تر شده.

از معایب درخت تصمیم میتوان حساسیت زیاد نسبت به نویز و incremental (افزایشی) نبودن کسب دانش را نام برد. به عبارتی یادگیری تدریجی نخواهد بود یعنی اگر داده ای به دادگان آموزشی اضافه شود از ابتدا باید درخت و فرضیه ساخته شود.   
فرض مهمی که ما در زمان انتخاب ویژگی در نظر می گیریم برابر بودن و یکسان بودن هزینه برای تمام ویژگی هاست.   
درخت تصمیم برای ویژگی هایی مناسب است که دارای مقادیر گسسته هستند و برای مقادیر پیوسته مناسب نیست مگر اینکه با روش های گسسته سازی مقادیر پیوسته را به گسسته تبدیل کنیم.