مبانی داده کاوی علیخانی

-

جزوه دست نویس امیرمحمد خالقی فرید

درخت تصمیم گیری/ Decision Tree

درخت تصمیم‌گیری با نگاه کردن به مقادیر لیبل (نتیجه)، بررسی می‌کند که کدام ویژگی اگر مبنا و معیار دسته‌بندی قرار گیرد، باعث می‌شود داده‌ها به گروه‌هایی تقسیم شوند که در هر گروه، مقادیر لیبل تا حد ممکن مشابه و یکنواخت باشند.سپس بر اساس همان ویژگی و مقادیر آن، داده‌ها را دسته‌بندی می‌کند و این فرآیند را به‌صورت تکراری ادامه می‌دهد تا به نتیجه برسد.

درخت تصمیم گیری سعی میکنه با تعیین شروط مبتنی بر ویژگی های دیگه، گروه بندی هایی ارائه بده که مقدار لیبل هاشون تو یه مایس!

مثلا:

یه جدول داریم سه تا ستون داره:

- سابقه ورزش
- سن
- توانایی ددلیفت 100کیلو

توانایی ددلیفت 100 کیلو	سن(سال)	- سابقه ورزش (سال)
ندارد	70	0
ندارد	60	5
دارد	50	10
دارد	40	15
دارد	36	20

اگر لیبل را توانایی ددلیفت 100 کیلو در نظر بگیریم:

بر اساس معیار سابقه ورزش به شرط و دو شاخه زیر میرسیم:

اگر سابقه ورزش<= 5 توانایی ددلیفت 100 کیلو ندارد 100

و در غیر اینصورت دارد!

اینصورت دارد!

و بر اساس معیار سن به شرط و دو شاخه زیر میرسیم:

اگر سن >50 توانایی ددلیفت 100 کیلو ندارد و در غیر اینصورت دارد!

Information Gain

یه معیار ساده که فقط میاد چک میکنه ببینه مقادیر ویژگی اگر بر اساسش شرط نوشته بشه چقد یکنواخت میشه مقادیر لیبل هاشون بر اساس اون شرط. به کلام ساده در این متد اول میاد میزان تصادفی بودن داده ها رو قبل از دسته بندی بر اساس لیبل توسط درخت تصمیم گیری میبینه اگه گزینه های توی گروه رنج مقدار لیبلشون نزدیک تر یا دقیقا یک مقدار شده باشه، IG بالا میره.

Gain_ratio

بجای اینکه فقط یکنواختی رو چک کنه میاد ببینه داده ها شاخشون تکی نباشه(این نباشه که گروه ها هر کدوم تک عضوی باشن و صرفا بخاطر این یک نواختی و "مشابه بودن مقادیرشون" بالا رفته باشه) و اگر اینطور بود information Gain شون رو کم کنه تا داده های تک شاخه بیوفتن تو اولویت کمتر.(در نتیجه شروط و خصوصیاتی که خیلی جزئی هستند اینطوری میوفتن آخر تر و در شکل درخت اونایی که تعداد زیر گره های کمتری دارن بیوفتن پایین مثلا شاخه های تک برگ بیوفتن پایین پایین)