درخت تصمیم دودویی با MATLAB

ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیش بینی کننده می باشد که حقایق مشاهده شده در مورد یک پدیده را به استنتاج هایی در مورد مقدار هدف آن پدیده نقش می کند. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده ها، یادگیری درخت تصمیم نامیده می شود که یکی از رایج ترین روش های داده کاوی است.

هر گره داخلی متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. یک گرهء برگ، با داشتن مقادیر متغیرها که با مسیری از ریشه درخت تا آن گره برگ بازنمایی می شود، مقدار پیش بینی شده متغیر هدف را نشان می دهد.

یک درخت تصمیم ساختاری را نشان می دهد که برگ ها نشان دهنده دسته بندی و شاخه ها ترکیبات فصلی صفاتی که منتج به این دسته بندی ها را بازنمایی می کنند. یادگیری یک درخت می تواند با تفکیک کردن یک مجموعه منبع به زیرمجموعه هایی براساس یک تست مقدار صفت انجام شود. این فرآیند به شکل بازگشتی در هر زیرمجموعه حاصل از تفکیک تکرار می شود. عمل بازگشت زمانی کامل می شود که تفکیک بیشتر سودمند نباشد یا بتوان یک دسته بندی را به همه نمونه های موجود در زیرمجموعهء بدست آمده اعمال کرد.

درختان تصمیم قادر به تولید توصیفات قابل درک برای انسان، از روابط موجود در یک مجموعه داده ای هستند و می توانند برای وظایف دسته بندی و پیش بینی بکار روند. این تکنیک به شکل گسترده ای در زمینه های مختلف همچون تشخیص بیماری دسته بندی گیاهان و استراتژی های بازاریابی مشتری بکار رفته است.

این ساختار تصمیم گیری می تواند به شکل تکنیک های ریاضی و محاسباتی که به توصیف، دسته بندی و عام سازی یک مجموعه از داده ها کمک می کنند نیز معرفی شوند. داده ها در رکوردهایی به شکل $(x, y) = (x_{1,1}, x_{2,2}, x_{3...}, x_k, y)$ داده می شوند. با استفاده از متغیرهای $x_{1,2,...,k}$ سعی در درک، دسته بندی یا عام سازی متغیر وابستهء Y داریم.

انواع صفات در درخت تصمیم به دو نوع صفات دسته ای و صفات حقیقی بوده که صفات دسته ای، صفاتی هستند که دو یا چند مقدار گسسته می پذیرند (یا صفات سمبلیک) درحالی که صفات حقیقی مقادیر خود را از مجموعهء اعداد حقیقی می گیرند.

روش تقسیم صفات دسته ای به شکل دودویی: در این روش بعد از انتخاب یک صفت برای عمل تصمیم گیری در هر گره، فقط یک تقسیم دودویی انجام می دهیم که دو فرزند با عناوین "صفت تصمیم برابر یک مقدار خاص می باشد" و "صفت تصمیم برابر آن مقدار نمی باشد" تولید می کند.

دو مجموعه داده ای داریم به نام های TestingData.mat جهت تست داده ها و TrainingData.mat جهت آموزش داده ها. می خواهیم داده ها را بعد از آموزش، طبقه بندی نماییم و آن را تست کنیم. یک عمق

برای درخت در نظر می گیریم که مقدار آن را ۵ قرار می دهیم. یک مقدار برای آستانه های کاندید در هر گره داریم که مقدار آن ۱۰۰ است و حداقل اندازه گره هایی که برگ نیستند را هم با مقدار ۱۰ در نظر می گیریم. سپس داده های آموزشی را بارگذاری می کنیم. یک درخت با مقادیر عمق، آستانه های کاندید در هر گره و حداقل اندازه گره های غیر برگ می سازیم. ماتریس داده های آموزشی به صورت M×N:X خواهد بود که هر سطر یک نمونه از داده است. برچسب داده های آموزشی به صورت N×۱ است که هر ورودی مقدار بین ۰ و ۱ را دارد.

همینطور در نظر گرفتن یک مقدار جهت نتایج درخت تصمیم دودویی لازم است که در برنامه با متغیر T نشان داده می شود. X یک نمونه از داده در بردار سطر است. Y برچسب نتیجه است. P نتیجه احتمال این که برچسب ۱ باشد خواهد بود. یک گره شاخص داریم (k) که مقدارش ۱ است. عمق مسافتی (d) برابر ۱ است و یک ستون (c) داریم. نیاز است که عمق و ستون ها را به گره شاخص تبدیل نماییم.

در داده های آموزش وارد شده به برنامه مقدار X یک ماتریس ۲۳×۳۰۰۰ و مقدار Y یک ماتریس ۱×۳۰۰۰ است. داده های تست نیز برای X یک ماتریس ۲۳×۳۹۱ و برای Y برابر ۱×۳۹۱ است.

لازم است که گره شاخص را به عمق و ستون تبدیل نماییم که رابطه آن به صورت ذیل است:

```
d=floor(log(k)/log(2))+1;
```

```
c=k-2.^(d-1)+1;
```

در فایل get01Tree.m مقدار k همان گره جاری است که در حال کار کردن است. T درخت تصمیم جاری است. M×N:X ماتریس داده آموزشی است که هر سطر یک نمونه داده است. در این فایل آموزش در گره k انجام می شود. سپس برگ و غیربرگ ها به دست می آید و به عنوان فرزندان سمت چپ و راست درخت، به وجود می آیند. K×2=K برای فرزندان چپ و K×2=K+۱ برای فرزندان سمت راست می باشد. در نهایت لازم است که انتروپی مجموعه برچسب ها را به دست آوریم که از بردار یک برچسب، می توان نتیجه انتروپی را به دست آورد. هدف به دست آوردن نرخ خطا می باشد.

نتایج به دست آمده در خط فرمان در محیط MATLAB به صورت جدول ۱ است.

جدول ۱ – خطای به دست آمده و زمان اجرای آموزش و تست

نرخ خطا	۰,۱۰۲
زمان آموزش	۲,۶۱۸۱ ثانیه
زمان تست	۰,۰۱۶۹ ثانیه