تمرین پیاده سازی

پیاده سازی درخت تصمیم با استفاده از نرم افزار weka

ابتدا از دستورات زیر استفاده کنید و دیتاستی به اندازه ۶۰۰۰ داده دارای دو صفت x1 و x2 دارای مقادیری بین 5- و 5 و دو کلاس 1- و 1 فراهم کنید:

```
a=-5;
b=5;
matrix=[];
 for i=1 : 6000
    x1= a + (b-a)* rand (1);
    x2= a + (b-a)* rand (1);
    class=sign (-2+x1+2 * x2);
    matrix=[matrix; x1 x2 class];
 end
save -ascii data.dat matrix
```

از میان کلیه داده‌های تولید شده به ترتیب ۵۰، ۱۰۰ ، ۲۵۰ و ... رکورد از آنها را انتخاب کرده و درخت J48 را روی آن اجرا کنید. از آنجا که دامنه ورودی یک دامنه پیوسته است لازم است یک گسسته سازی توسط خود الگوریتم روی داده‌های آن انجام دهید. از طرفی می‌دانیم در درخت تصمیم استاندارد لازم است از هر مسیر ریشه به برگ هر یک از صفات حتما یکبار ملاقات گردد. ولی درخت J48 به این صورت عمل نمی‌کند و در هر مسیر از ریشه به برگ می‌توان چندین بار یک صفت (x1, x2) را ملاقات کرد. این الگوریتم صفحه را طوری بخش‌بندی می‌کند که حداکثر خلوص در هر بخش وجود داشته باشد و در عین حال اندازه درخت نیز کمینه باشد. که این یک trade off بین اندازه درخت و درجه خلوص در هر ند برگ در درخت می‌باشد. چیزی که واضح است این درخت نمی‌تواند معادله این خط مورب را به صورت کامل پیدا کند چرا که درخت فقط می‌تواند روی x1 و x2 شرط بگذارد و همواره قدری خطا روی داده‌های دسته بندی شده با این درخت وجود دارد. در صورتی خطا به صفر کاهش می‌یابد که اندازه درخت نیز به سمت بی‌نهایت میل کند. ولی واضح است هر قدر تعداد داده‌های آموزشی افزایش یابد درخت محاسبه شده بهتر می‌تواند معادله خط را تقریب بزند. بنابراین به شرط وجود داده‌های آموزشی کافی و دارای توزیع یکنواخت می‌توان معادله تقریبی خط را به کمک درخت تصمیم محاسبه نمود و هیچگاه نمی‌توان معادله دقیق خط را فراگرفت.

برای پیاده‌سازی این الگوریتم مطابق جدول زیر ابتدا داده‌های خود را فراهم کنید. از میان کل داده‌ها، هربار نیمی از آن‌ها را به عنوان داده های آموزشی و نیمی را به عنوان داده‌های تست انتخاب کرده و عملکرد الگوریتم را را روی داده‌های تست بررسی کنید.

اندازه درخت								
درصد خطا روی داده آموزشی								
درصد خطا روی داده تست								
اندازه داده آموزشی	۲۵	۵۰	۱۲۵	۲۷۵	۵۵۰	۸۲۵	۱۲۷۵	۳۰۰۰
اندازه داده تست	۲۵	۵۰	۱۲۵	۲۷۵	۵۵۰	۸۲۵	۱۲۷۵	۳۰۰۰

جدول ۱- نمایش افزایش دقت الگوریتم J48 با افزایش تعداد مثال‌های آموزشی برای یادگیری معادله $0<2x_2+x_1-2$.

خروجی روی دادهای آموزشی:

خروجی روی داده‌های تست: