Machin learning assignment #2

Decision tree algorithm for classification and regression

Shirin Mohebbi - Razie Masoudi

بخش اول:

برای ساختن decision tree ما احتیاج به ساختار درخت داریم. به این منظور کلاس Node را می‌سازیم. که با توجه به نوع مسئله دارای attribute های زیر هست.

.

Value: مقدارattribute ی که باعث ایجاد این شاخه شده

attribute: آنattribute ی که برای دسته بندی این شاخه انتخاب شده

childs: فرزندان این شاخه

label: مقدار label برگ

Cal_entropy:

در این تابع میزان آنتروپی محاسبه می‌شود.

Cal_gain:

در این تابع gain محاسبه می شود.

Select attribute:

در این تابع بهترین attribute برای دسته بندی این شاخه از درخت انتخاب می‌شود. به این صورت که information gain تمام attribute های باقی مانده را حساب میکنیم. و آن attributeی را که gain بیشتری دارد انتخاب میکنیم. سپس آن attributt، تمامی value های اون attribute و دیتاهای جدا سازی شده بر اساس value هایش را برمیگردانیم.

Make tree:

در این تابع ما باید درخت را بسازیم. به این صورت که چک میکنیم که اگر به یک node pure رسید آنرا بازگردانی کند. و در غیر اینصورت ساختن درخت را باید ادامه دهد. به این طور که ابتدا attribute بعدی برای جداسازی با استفاده از تابع select attribute انتخاب می‌کند. طبق آن برای node فعلی ارایه ای از فرزندانش را می‌سازد. و برای کامل کردن و ساخته شدن هر فرزند دوباره تابع ریکرسیو make tree را صدا می‌زند تا فرزندان node به صورت کامل ساخته شوند و سپس آن node را برمی‌گرداند.

Classifier:

در این تابع از طریق پیمایش درخت، و بر اساس attribute ها و value های هر node و مقدار attribute های دیتای تست، درخت پیمایش شده تا زمانی که به برگ برسیم، و label برگ را بر می‌گردانیم.

K10fold

در این تابع دیتاست ما به 10 بخش تقسیم شده، هر بار یک بخش را به عنوان داده تست و نه بخش دیگر را به عنوان ترین در نظر میگیریم. در نهایت مقدار میانگین و انحراف معیار این 10 بار را باز می گردانیم.

```
level: 0 , value: root , attr: 5 , label: None
level: 1 , value: a , attr: None , label: e
level: 1 , value: c , attr: None , label: p
level: 1 , value: f , attr: None , label: p
level: 1 , value: l , attr: None , label: e
level: 1 , value: m , attr: None , label: p
level: 1 , value: n , attr: 20 , label: None
level: 2 , value: b , attr: None , label: e
level: 2 , value: h , attr: None , label: e
level: 2 , value: k , attr: None , label: e
level: 2 , value: n , attr: None , label: e
level: 2 , value: o , attr: None , label: e
level: 2 , value: r , attr: None , label: p
level: 2 , value: w , attr: 22 , label: None
level: 3 , value: d , attr: 21 , label: None
level: 4 , value: v , attr: None , label: p
level: 4 , value: y , attr: None , label: e
level: 3 , value: g , attr: None , label: e
level: 3 , value: l , attr: 15 , label: None
level: 4 , value: n , attr: None , label: e
level: 4 , value: w , attr: None , label: p
level: 4 , value: y , attr: None , label: p
level: 3 , value: p , attr: None , label: e
level: 3 , value: w , attr: None , label: e
level: 2 , value: y , attr: None , label: e
level: 1 , value: p , attr: None , label: p
level: 1 , value: s , attr: None , label: p
level: 1 , value: y , attr: None , label: p
```

Result:

```
100.0 +/- 0.0
```

بخش دوم:

توابع این بخش مانند بخش قبلی است با این تفاوت که به جای آنتروپی از standard deviation دیتاها استفاده می‌کنیم. به دلیل اینکه regression است.

Enjoy sport tree:

```
level: 0 , value: root , attr: 1 , label: None
level: 1 , value: Overcast , attr: None , label: 46.25
level: 1 , value: Rainy , attr: 2 , label: None
level: 2 , value: Cool , attr: None , label: 38.0
level: 2 , value: Hot , attr: None , label: 27.5
level: 2 , value: Mild , attr: None , label: 41.5
level: 1 , value: Sunny , attr: 4 , label: None
level: 2 , value: False , attr: None , label: 47.666666666666664
level: 2 , value: True , attr: None , label: 26.5
```

Result:

```
mse: 14.208333333333334
```

Automobile tree:

```
level: 0 , value: root , attr: 2 , label: None
level: 1 , value: alfa-romero , attr: None , label: 16500.0
level: 1 , value: audi , attr: None , label: 16370.0
level: 1 , value: bmw , attr: None , label: 27910.0
level: 1 , value: chevrolet , attr: None , label: 6007.0
level: 1 , value: dodge , attr: None , label: 7837.166666666667
level: 1 , value: honda , attr: None , label: 8403.57142857143
level: 1 , value: isuzu , attr: None , label: 6785.0
level: 1 , value: jaguar , attr: None , label: 34125.0
level: 1 , value: mazda , attr: 11 , label: None
level: 2 , value: 2bbl , attr: None , label: 8495.0
level: 2 , value: 4bbl , attr: None , label: 11395.0
level: 2 , value: idi , attr: None , label: 18344.0
level: 2 , value: mpfi , attr: None , label: 16962.5
level: 1 , value: mercedes-benz , attr: None , label: 33647.0
level: 1 , value: mercury , attr: None , label: 16503.0
level: 1 , value: mitsubishi , attr: 1 , label: None
level: 2 , value: -1 , attr: None , label: 9279.0
level: 2 , value: 1 , attr: None , label: 8036.5
level: 2 , value: 2 , attr: None , label: 5789.0
level: 2 , value: 3 , attr: None , label: 12621.5
level: 1 , value: nissan , attr: 10 , label: None
level: 2 , value: four , attr: None , label: 7299.0
level: 2 , value: six , attr: None , label: 16639.0
level: 1 , value: peugot , attr: None , label: 14815.625
level: 1 , value: plymouth , attr: None , label: 8880.75
level: 1 , value: porsche , attr: None , label: 34528.0
level: 1 , value: renault , attr: None , label: 9595.0
level: 1 , value: saab , attr: None , label: 15260.0
level: 1 , value: subaru , attr: None , label: 8719.875
level: 1 , value: toyota , attr: 7 , label: None
level: 2 , value: 4wd , attr: None , label: 8778.0
level: 2 , value: fwd , attr: None , label: 7189.428571428572
level: 2 , value: rwd , attr: None , label: 12262.5
level: 1 , value: volkswagen , attr: None , label: 9918.888888888889
level: 1 , value: volvo , attr: None , label: 17447.222222222223
```

Result:

```
mse: 1.364672e+07 +/- 2.848938e+06
```