گزارش کار

یادگیری ماشین

تمرین دوم

آرمین خیاطی

9931153

پویان انصاری راد

9960565

فهرست

مقدمه

در این تمرین در بخش اول به پیاده سازی الگوریتم Decision Tree ID3 برای کلاس بندی داده های دیتاست قارچ ها، و در بخش دوم به پیاده سازی یک الگوریتم درخت تصمیم مبتنی بر ID3 برای رگرسیون و پیشبینی مقادیر پیوسته خواهیم پرداخت. برای این تمرین از هیچ کتابخانه ای به جز Pandas و Numpy برای پیاده سازی الگوریتم ها استفاده نشده است.

بخش اول

در این بخش با استفاده از توابع Entropy و Information Gain به محاسبه و یافتن بهترین ویژگی برای تقسیم داده ها میپردازیم و درخت را مرحله به مرحله تا وقتی که ویژگی ای باقی نماند یا همه داده ها یک دست شده باشند یعنی به برگ برسیم، پیش میبریم.

بعد از پیاده سازی درخت، و جایگذاری مقادیر خالی با Mode آن ستون داده، با استفاده از روش آماری K-Fold Cross Validation با K=10 به بررسی عملکرد آن میپردازیم که میانگین دقت و واریانس زیر حاصل میشود.

Accuracy mean is 1.0, and STD is 0.0

بخش دوم

در این بخش با استفاده از درخت تصمیم به پیش بینی مقادیر پیوسته برای دو دیتاست Enjoy Sport و Automobile میپردازیم. الگوریتم درخت، مشابه قبل هست با این تفاوت که، بجای استفاده از انتروپی و Information Gain برای انتخاب بهترین ویژگی از واریانس میان دو ویژگی، و STD Reduction استفاده میکنیم که فرمول آن ها در مقاله ای که همراه تمرین فرستاده شده وجود دارد. شرط توقف ساخت گره یعنی ایجاد گره برگ نیز نداشتن ویژگی دیگری برای ادامه، کمتر بودن داده های باقی مانده از یک حد خاص و کمتر بودن CV داده ها از یک حد خاص که فرمول محاسبه CV نیز در مقاله آورده شده است. برای بررسی دقت مدل برای دیتاست ماشین ها نیز، داده ها را ده بار بصورت رندم به دسته های ترین و تست به نسبت 70 به 30 تقسیم میکنیم و میانگین خطا را بدست می آوریم.

برای مدیریت وضعیت هایی که مقادیر بعضی ویژگی ها در داده های تست، هنگام آموزش و ساخت درخت دیده نشده است، نیز از راه حل زیر استفاده میکنیم:

دیتاست را به تعداد ده بار تکرار (کپی) میکنیم، با این کار مطمئن میشویم هر مقدار حداقل ده بار تکرار شده است و مقدار میانگین خطای MSE کار با این عمل از 17 میلیون به 2 میلیون خواهد رسید.

نتیجه حاصله برای دیتاست Enjoy Sport بصورت زیر

Enjoy sport MSE: 14.208333333333334

و برای دیتاست Automobile نیز مقادیر زیر بدست می آید :

مقدار خطای میانگین سیزده میلیون قبل از ده بار تکرار داده ها:

Average + STD = 13447774.749307003 ± 2371353.7775489474

مقدار خطای میانگین دو میلیون بعد از ده بار تکرار داده ها:

Average + STD = 2781598.6418455904 ± 325501.93123312865

خطای روی داده های ماشین ها با استفاده از ماژول DecisionTreeRegressor کتابخانه Sklearn نیز مقدار هفده میلیون بدست می آید که نشان دهنده قوی تر بودن مدل و تکنیک ما در پیاده سازی این تمرین است.

Average + STD = 17746774.93513889 ± 7131258.144507702

خلاصه نتایج:

الگوریتم	دیتاست	میانگین خطا	واریانس خطا
ID3 Decision Tree	قارچ ها	1	0
Decision Tree Regression - From Scratch	Enjoy Sport	14.2	-
	ماشین ها بدون تکرار داده ها	13,447,774.7	2,371,353.7
	ماشین ها با ده بار تکرار داده ها	2,781,598.6	325,501.9
Decision Tree Regressor - SkLearn	ماشین ها بدون تکرار داده ها	17,746,774.9	7,131,258.1