بسم الله الرحمن الرحیم

گزارش پروژه سوم مبانی یادگیری ماشین

محمد رضا ضیالاری (97222057)

تمرین 1)

کرنل ها برای تفکیک کردن داده ها کاربرد دارند و به ما برای این موضوع کمکگر هستند و با افزایش ابعاد فضا ی ویژگی تفکیک پذیری داده ها را برای ما راحت تر می کنند . در زیر به چند نمونه کرنل و موارد غالب استفاده آنها اشاره می کنیم .

الف) کرنل خطی : برای توابع خطی به کار می رود .

ب) کرنل های گاوسی : این کرنل ها از توزیع گاوسی پیروی می کنند و بسیار پرکاربردند .

ج) کرنل های RBF : مشابه کرنل های گاوسی هستند با این تفاوت که شعاعی را نیز برای تفکیک داده ها در نظر می گیرند . این نوع کرنل هم بسیار پرکاربرد است .

د) کرنل سیگموید : در شبکه های عصبی و مباحثث مربوط به یادگیری عمیق بیشتر کاربرد دارد .

ه) کرنل های چند جمله ای : این گونه از کرنل ها بیشتر در پردازش تصویر کاربرد دارند .

و) Linear splines kernel in one-dimension : این کرنل بیشتر برای برای داده هایی با پراکندگی بالا (مانند تشخیص متن) کاربرد دارد .

تمرین 2)

در این تمرین از ما خواسته شده که بر روی دیتای موبایل SVM اجرا کنیم که پس از بررسی عدم تهی بودن و بررسی پراکندگی و دیگر پیش پردازش های مقدماتی مدل SVM پیاده سازی و اجرا شد که به ما دقت خوب 97٪ را داد .

تمرین 3)

در این بخش از تمرین از ما خواسته شده بود که کرنل های مختلف را با پارامتر های مختلف روی دیتاست بررسی کنیم .

کرنل خطی : دقت در این حالت 96.67٪ بود .

کرنل rbf : دقت 97٪ بود .

کرنل سیگموید : دقت در حالت سیگموید بسیار پایین و حدود 17 درصد بود و با تغییر پارامتر coef0 به 0.2 دقت به حدود 14 درصد کاهش یافت .

کرنل چند جمله ای : در حالت های مختلف با پارامتر های مختلف برسی شد . نتیجه ان بود که با افزایش درجه چند جمله ای دقت کاهش پیدا می کند و به ترتیب دقت برای درجه های 2و3و4و10و100 برابر 96.67 و 96.33 و 96 و 94.67 و 69.67 درصد بود .

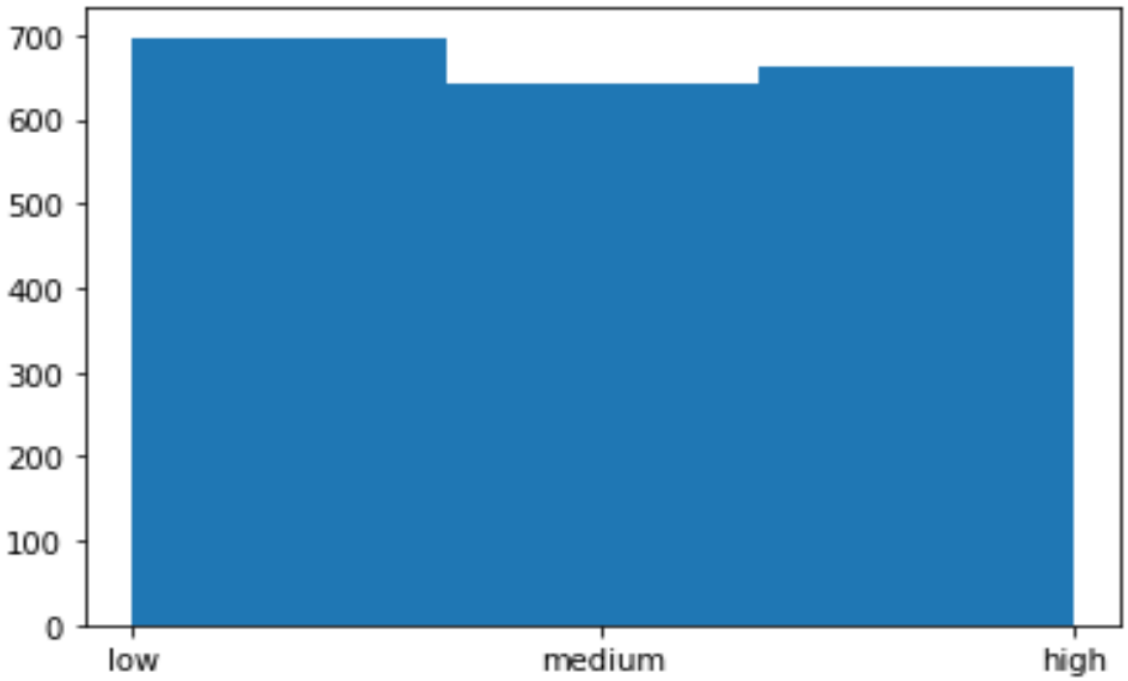
تمرین 4)

هنگامی که از هارد مارجین استفاده می کنیم به دلیل عدم miss classification دقت داده های آموزشی بالاست اما از طرفی چون به گونه ای داده ها اورفیت می شود ، دقت تست پایین تر از حالت سافت مارجین است .

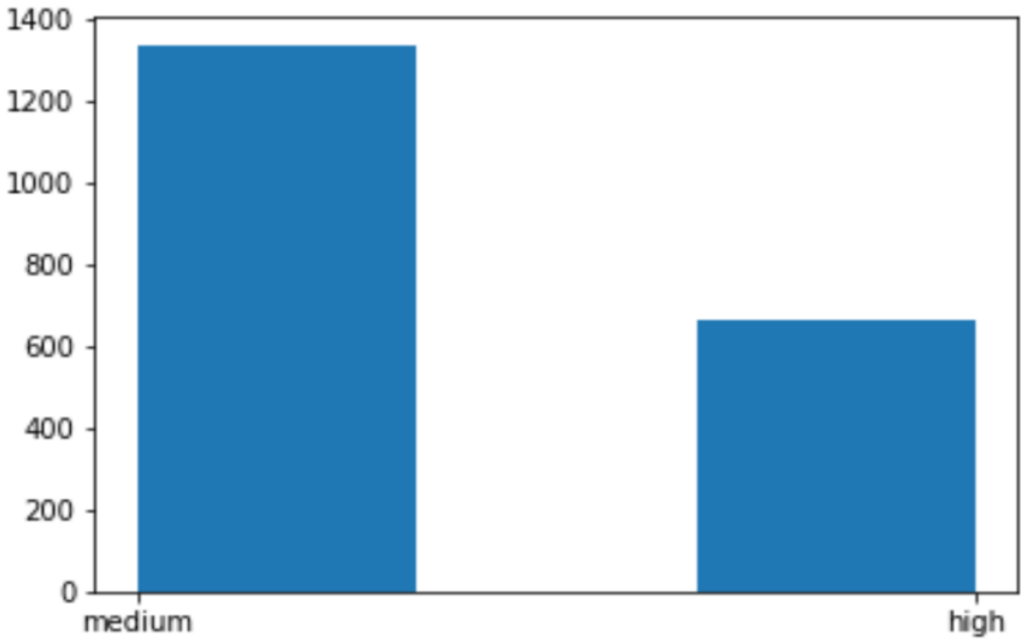
تمرین 5)الف)

سه اندازه مختلف ، binهای مختلف را در نظر می گیریم.

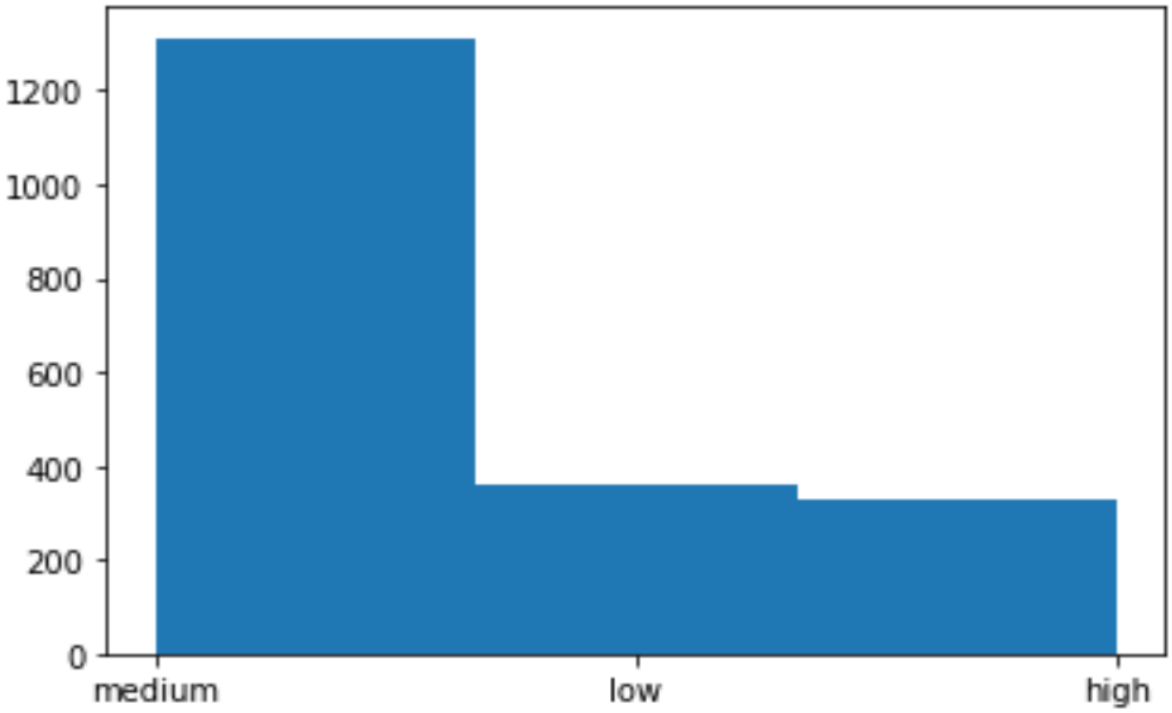
الف) سه قسمت مساوی در دامنه اعداد : که توزیع آن به شکل زیر است :



ب) bins = [0, 499, 1499, 2000] :



ج) bins = [0, 750, 1750, 2000] :



پس از اجرای SVM دقت برابر 83٪ شد .

تمرین 5)ب)

این دیتا ست داده ی کتگوریکال ندارد و تمامی داده ها عددی هستند .

این روش برای عددی کردن داده ها است و با جدا سازی همه متغیر های یک ستون به ستون های مجزا باعث میشود اگر مقداری در ستون مهم بود ، مدل ما راحت تر آن را کشف کند.

تمرین 5)ج)

استفاده از تبدیل ها باعث می شود که داده ها توزیع بهتری داشته باشند .

Log transform بیشتر مواقعی به کار میرود که داده ها گوناگونی زیادی دارند و از توزیع نرمال دور هستند (توزیع نمایی دارند) و توزیع آنها چولگی داشته باشد .پس از انجام این تبدیل

داده ها به توزیع نرمال نزدیک میشوند .همچنین میزان تاثیر داده های پرت را نیز با اینکار کم میکند و مدل را بهتر میکند.

در ابتدا مینیمم داده ها را برای هر ویژگی میگیریم تا چک کنیم ببینیم نا منفی نباشد .

که مشاهده می کنیم کمترین مقدار صفر است . به این دلیل همه ا یک واحد به راست شیفت میدهیم .

دقت مدل svm در این حالت برابر 89.25٪ بود .

تمرین 5)د)

با ضرب طول در عرض مساحت رابدست می آوریم و به جای این دوستون جایگذاری میکنیم و مدل را اجرا می کنیم . با این تغییر دقت به 31% کاهش یافت .

تمرین 6)

مدل SVM به طور جداگانه برای هر کدام از بخش های تمرین 5 اجرا شده است و گزارش آن نوشته شده است . حال یک مدل SVM کلی را پیاده سازی میکنیم .

دقت در این حالت حدود 87 درصد بود.

تمرین 7)

الگوریتم های مختلفی برای درخت تصمیم وجود دارد که برای مثال در برخی از آنها از هرس کردن استفاده میشود و... .

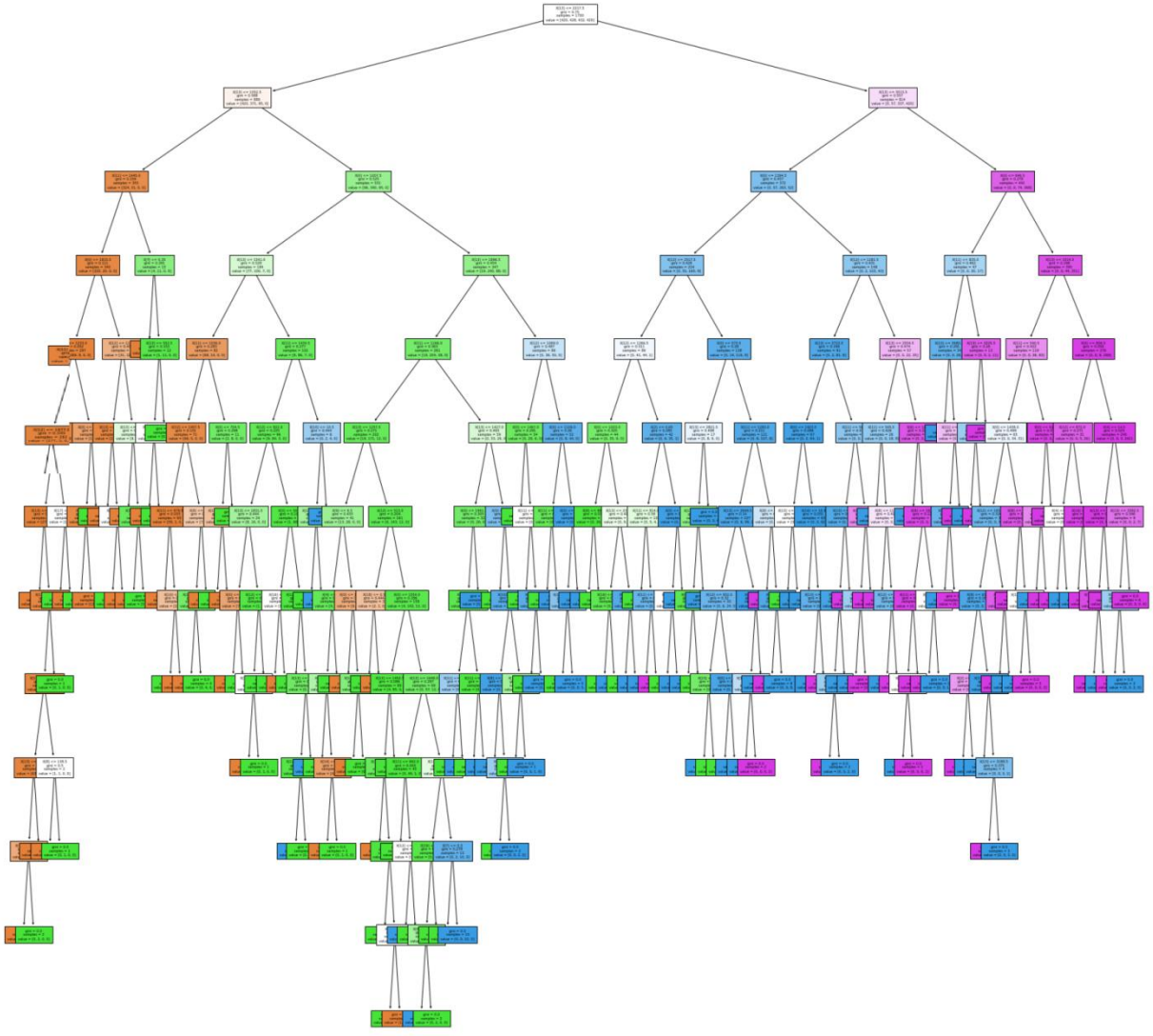
الگوریتم اول ID3است یا نام کامل آن Iterative Dichotomiser 3 که این الگوریتم از information gainبرای ساخت درخت استفاده میکند. این معیار تعیین میکند که کدام ویژگی ها اطلاعات بیشتری دارند و برای ساخت درخت لازم هستند.آنهایی که اطلاعات

بیشتری دارند در راس درخت قرار می گیرند و بدین ترتیب زیر درخت‌های دیگر نیز ساخته میشوند. این الگوریتم مخصوص داده های پیوسته است .

الگوریتم دیگر ، الگوریتم CHAIDاست. در این الگوریتم برای ساخت درخت، داده ها را متناوبا به زیر مجمعه های یکسان تقسیم میکند تا جایی که هر زیرمجموعه دارای تعداد مشخصی نمونه شود. این الگوریتم از آزمون Chi squaredبرای تصمیم گیری در هرتقسیم برای مشخص کردن زیر درخت ها استفاده می کند .

تمرین 8)

با استفاده از پکیج sklearn یک مدل درخت تصمیم می سازیم و آنرا پیاده سازی می کنیم .
دقت در این حالت برابر84.66% بود .



تمرین 9)

در ابتدا عمق درخت را زیاد میکنیم و مشاهده می کنیم با افزایش عمق درخت ، دقت نیز افزایش می یابد .

در حالتی که عمق درخت 5 بود ، دقت برابر 89.33٪ ولی هنگامی که عمق را به 10 و 50 افزایش دادیم دقت بع ترتیب برابر 98.33٪ و 98.67٪ شد .

همچنین با افزایش تعداد ویژگی نیز دقت بالا می رفت . برای مثال برای ماکسیمم تعداد ویژگی 2 و 5 و 10 ، دقت به ترتیب برابر 91.67٪ و 95.67٪ و 98.33٪ شد .

تمرین10)

هرس کردن ینی اینکه ما بخشی از درخت تصمیممان را در نظر نگیریم .

هرس کردن به گونه ای مانند drop out است و از پیچیدگی مدل ما کم میکند و مانع از اورفیت شدن می شود.

تمرین 12)

با پیاده سازی مدل جنگل تصادفی (بدون تغییر پارامتر) دقت باربر 98٪ شد که بسیار بیشتر از درخت تصمیم بود .

چون جنگل تصادفی به طور رندوم از زیر مجموعه ای از ویژگی ها استفاده می کند ، احتمال اورفیت شدن آن بسیار کمتر است و به سمت ویژگی خاصی بایاس نمی شود . پس این نتیجه منطقی است .

تمرین 13)

روش هایی مانند درخت تصمیم سرعت بیشتری نسبت به یادگیری عمیق دارد و پیاده سازی و تنظیم آن بسیار راحتتر می باشد . همچنین برای حالاتی که دیتا و فیچر تعداد نسبتا کمی دارند ، استفاده از ذرخت تصمیم بسیار منطقی تر و بهینه تر است .