بسم الله الرحمن الرحیم

نویسنده : محمد صادق ایلاقی

شماره دانشجویی : ۹۸۱۴۳۰۰۷

ورودی ۹۸ دانشگاه ولی عصر

مقطع کارشناسی علوم کامپیوتر

لینک مخزن پروژه :

https://github.com/mohammad-IR/DataMiningProject_Final_Collage

مقدمه

دیتاستی که انتخاب کردیم به نام adultمی باشد که برای بررسی مقدار درامد ها در کشورهای مباشد که چک میکند که آیا درآمد بالایی پنجا هزارتا میباشد یا خیر ولی بنده تصمیم گرفتم که بر اساس بقیه داده ها بیام سطح تحصیلی رو پیش بینی کنم اول بار به دلیل عددی نبودن عداد و اشنایی نداشتن کامل با داده ها اول سعی در شناخت انهاcorrelation بین انها کردم و بعد هم داده ها prepareکردم **بعد از مدل های مختلف مثل** svc و liniearsvc **با نرمال کردن داده ها انجام دادم و انواع درخت های تصمیم رو انجام**

دادم ولی یه نکته جالب این بود که حتی با cross_vla_scoreکه امتحان کردم و 10 مدل رو ساختم از روی decision_tree همه دقت صد در صد رو دادن و حتی extra_tree و random_forestهم دقت بالای داشت و ensembleها رو هم روی svc و linearsvc و انواع درخت تصمیم پیاده سازی کردم و نتجه خوبی داشت در اخر بعضی مدل ها خطای converge رو دادن که میگفت مدل بهینه رو نمیتونه پیدا کنه که مشکل رو با دادن پارامتر های خوب میشد حل کرد ولی در کل عملکرد های مدل ها به صورت زیر شرح می شود.

SVC با 05.regularizes

مدل افتضاح کار میکنه دلیل عمده اون مال استاندارد نکردن داده ها میباشد و با هر پارامتری با مدل دقت پایینی دارد اما بعد از استاندارد سازی و با کرنل های مختلف از ploy و rbf و regularizesهای مختلف به این نتیجه رسیدم که نباید regularizes کمتر از 0.4 باشد و با کرنل poly روی requlriesعملکرد خیلی خوبی دارد 96 درصد دقت.

LinrearSVC

با سازوکار ovrخیلی خوب کار نکرد در کل با خطای convergeرو میداد که مدل نمیتونه بهینه رو خوب پیدا کنه اما یه نکته که داره روی multi_classبا پارامتر crammer_singer خیلی خوب کار کرد ولی همچنان ارور رو داره ولی این سازو کار چیزی که من در کتابخونه sickit-learn خواندم مثل همون ovoهست که خیلی خوب کار میکنه و دقت نود و پنج درصد رو دارد.

فایل دوم

دلیل ساخت فایل دوم اجرای طولانی مدل های svc و linearsvc بود که دوباره داده ها رو اماده کردم از قبیل عددی کردن داده ها و split کردن اون ها میباشد و روی مدل

های مثل درخت ها از درخت تصمیم تا randomforest تا ensembleها رو اجرا کردم که نتیجه به شرح زیر است.

ما از SGDclassifierاستفاده کردیم که هم standard regularizer و elasticnet استفاده کردیم که فرقی نداشت و مدل نمیتوانست به خوبی تعمیمدهد البته بهتر بود mini-batch و batch-GD رو هم استفاده کنیم ولی چون کتابخونه هاش رو ندیدم استفاده نکردم جالبی اسنجاست که بدون استاندارد سازی داده دقت به 20 درصد کاهش پیدا کرد و خیلی بد عمل کرد.

ما از vottingClassifier نیز استفاده کردیم که مدل های SVC و SGDClassifier و LinearSVC نیز مدل ها بود استفاده کردیم با استاندارد سازی داده در piplineکه تعمیم 60 درصد داد و خوب کار نکرد.

Bagginنیز برای مدل SVCاستفاده کردیم که دقت پایین داشت و خوب عمل نکرد.

از adoboostنیز استفاده کردیم که اون هم به همین شکل بود احتمالا یه جاهای بد داده رو ست کردم یاد پارامتر ها رو باید بیشتر امتحان میکردم چون معمولا ensemble ها فقط انتخاب کننده مدل هستن ولی در کل خوب عمل نکردن .

درخت های تصمیم خیلی خوب عمل کردن extraclassifier بالای 80 درصد دقت و randomforestclassifierنیز خیلی خوب کار کرد ولی بهترین decisiontree بودچون روی ده تا مدل همه 100 در صد بود.

نتیجه

با کار کردن بر روی مدل های مختلف تصمیم گرفتم درخت تصمیم رو انتخاب کنم ولی معمولا درخت های تصمیم چون بر اساس پارامتر های که مقایسه میکند و جدا میکند میشه انتظار این دقت ها رو داشت