گزارش پروژه پایانی درس استخراج ویژگی و انتخاب ویژگی

سهراب پیرهادی

۹۸۴۱۱۲

روش های feature selection به جای تعریف متغیر های جدید، یک زیر مجموعه از ویژگی های موجود در داده ها استفاده میکنند و به همین دلیل نگاه متفاوتی به مبحث کاهش بعد دارند. به همین دلیل مواقعی مناسب هستند که میخواهیم از ویژگی های موجود در داده ها استفاده کنیم و تفسیر بهتری از مساله در ابعاد کمتر داشته باشیم. چون دانش استخراج شده راحت تر قابل تفسیر است.

در این روش، زیر مجموعه ایی که دنبالش هستیم بایستی کمتر از تعداد ویژگی داده اصلی باشد تا بتوانیم کاهش بعد انجام دهیم.

روش های مختلفی برای انجام feature selection وجود دارد و در این پروژه از sequential forward selection (SFS) برای ایجاد مدلی برای برچسب زدن به داده های مساله استفاده کرده ایم.

این استراتژی ماهیت حریصانه دارد و مساله بهینه سازی را از یک مجموعه خالی شروع میکند و هر بار یک ویژگی اضافه می کند و مدل classification (Objective function) را بررسی میکند. از بین تمام مجموعه های یک عضوی، آن ویژگی که objective function را maximize (یا minimize) میکند، انتخاب میکند.

سپس دنبال این هستیم که کدام زیر مجموعه دو عضوی که عضو اول آن حتما این عضو انتخاب شده باشد، مناسب است. هر ویژگی که انتخاب میشود و در مجموعه جواب قرار میگیرد، دیگر از مجموعه جواب حذف نمیشود.

و به همین ترتیب به دنبال مجموعه های ۳ عضوی ... تا k عضو که زیر مجموعه مورد نظر ماست. پس جهت حرکت در این جستجو از زیر مجموعه خالی به سمت زیر مجموعه شامل تمام ویژگی هاست و با اعمال threshold، الگوریتم بهینه ازی را زودتر متوقف میکنیم.

معیاری که در این پروژه برای بررسی کیفیت مدل classification در نظر گرفته ایم، F1-score است و قصد داریم زیر مجموعه از ویژگی ها را طوری انتخاب کنیم که این معیار را maximize کند.

در هر لحظه میتواند یک ویژگی به مجموعه ویژگی های انتخاب شده، اضافه یا کم کند.

مجموعه داده این پروژه 2135 *102 است که مربوط به 102 نمونه آزمایشی است که برای هر نمونه 2135 ویژگی اندازه گیری شده است. پس 102 داده در فضای به ابعاد 2135 تعریف شده هستند. هر نمونه شامل یک برچسب سالم یا تومور است و قرار است به کمک SFS یک زیر مجموعه از ویژگی ها را طوری انتخاب کنیم که مدل بتواند برچسب ها را برای داده های test به بهترین شکل پیش بینی کند. معیار سنجش عملکرد مدل به کمک F1-score انجام می شود. از مدل KNN برای classification استفاده کردیم.

برنامه را برای انتخاب 5 ویژگی اجرا می کنیم. از ویژگی سوم به بعد، عملکرد مدل classification بهتر نمی شود و مقدار f1-score از 0.9677 تغییری نمی کند و دلیل این موضوع این است که از آن جایی که تعداد S ample ها در DataSet بسیار کمتر از تعداد feature هاست، مدل به سرعت overfit میشود. پس به ای ن نتیجه میرسیم که مدل classification با 3 ویژگی میتواند بهترین پیش بینی را برای برچسب داده ها ا نجام دهد. این سه ویژگی: `41706_at` ، `1944_f_at` و `40755_at` هستند.