بنام خدا

موضوع پروژه:کتابخانه sklearn در پایتون

کتابخانه sk-learn در پایتون یکی از محبوب‌ترین کتابخانه‌های ماشین در پایتون می‌باشد که برای آموزش و پیاده‌سازی الگوریتم‌های یادگیری ماشین استفاده می‌شود

این کتابخانه ابزار های کاربردی زیادی برای یادگیری ماشین و مدل سازی داده ها مثل طبقه بندی.خوشه بندی.رگرسیون و کاهش ابعاد در خود دارد.

کتابخانه sklearn بر پایه دیگر کتابخانه ها مثل scipy,numpy,matplotlib است که عمدتا از اززبان برنامه نویسی پایتون استفاده میکنند

برای نصب کتابخانه sklearn ما پیشنیازهایی را باید روی سیستم نصب کنیم مثل:

Python,numpy,scipy,matplotlib,pandas

طریقه نصب ماژول کتابخانه scikit-learn در cmd ویندوز به شکل زیر است:

Pip install scikit-learn

و برای فراخوانی کتابخانه کدزیر را مینویسیم:

Import sklearn as sk

ویژگی هایی که در این کتابخانه وجود دارد عبارت اند از:

1.الگوریتم های یادگیری نظارت شده:

تقریبا تمام الگوریتم های معروف رگرسیون خطی.درخت تصمیم گیری وماشین بردار پشتیبان در کتابخانه sklearn موجود هستند.

2.الگوریتم های یادگیری نظارت نشده

در این ویژگی از روش های خوشه بندی.pca.شبکه های عصبی نظارت نشده در این کتابخانه موجود هستند.

3.انتخاب ویژگی

روشی برای انتخاب ویژگی برای ساخت مدل هایی با دقت بیشتر یا افزونگی کمتر است.

4.استخراج ویژگی feature extraction

برای تعریف ویژگی های جدید از روی ویژگی اصلی برای استخراج داده های بکار میرود

5.کاهش ابعاد داده ها

برای خلاصه کردن.تصورسازی و انتخاب ویژگی مورد نظر استفاده میشود.

6.روش گروه بندی ensemble methods

از این روش برای ترکیب کردن مدل های یادگیری نظارت شده برای پیش بینی داده ها استفاده میشود

7.اعتبارسنجی

برای کنترل دقت نظارتی روی داده های تست استفاده میشود.

فرایند مدل سازی در sklearn

مدل سازی شامل بارگذاری داده.تقسیم داده . اموزش داده و تست ان است.

بارگذاری داده

در بعضی مواقع به پیش بینی کننده. صفت و یا ورودی گفته میشود .

ماتریکس ویژگی:مجموعه ای از ویژگی ها که طی اتفاق خاصی بیشتر از یک ویژگی دارند

تقسیم مجموع داده :ما میتوانیم تقسیم داده را به دو صورت :

1.مجموعه اموزشی:

از داده اموزشی برای اموزش یادگیری ماشین و استخراج مدل های اموزشی استفاده میشود.مدل استخراجی روی مجموعه تست اعمال میشود تا دقت مدل ان کاملا مورد بررسی قرار گیرد.

2.مجموعه ازمایشی

اموزش مدل:

بعد از تقسیم داده ها میتوانیم داده ها را برای یادگیری بهکار بگیریم

SkLearn الگوریتم‌های یادگیری ماشین گسترده‌ای دارد که رابط ثابتی برای مدل کردن، پیش‌بینی دقت و فراخوانی برای تمام الگوریتم‌ها ارائه میکند

پایداری مدل:

وقتی با اعمال الگوریتم روی مجموعه داده‌های آموزشی مدلی را استخراج می‌کنیم، انتظار میرود این مدل برای استفاده‌های بعدی نیز پایدار بماند. ازاین جهت مدل را چندین مرتبه بازآموزی (retrain) می‌کنیم. میتوانیم این کار را به کمک ویژگی‌های dump و load از مجموعه joblib انجام دهیم.

پیش پردازنده داده:

ما در فرم های خام داده ها با مقادیر زیادی از داده رو به رو هستیم.باید قبل از بکار گرفتن داده ها به عنوان ورودی ماشین انها را اماده سازی و به داده های مناسب و با معنی تبدیل کنیم. به این فرایند پیش پردازش داده میگویند.

کتابخانه sklearn از مجموعه preprocessing برای این کار استفاده میکند.این پکیج تعداد زیادی کلاس وتابع تبدیل کننده برای تبدیل بردار ویژگی و داده خام به داده های مفید و قابل استفاده تر توسط تخمین گر ها فراهم میکند.

الگوریتم های یادگیری با استفاده از استاندارد سازی داده ها می‌توانند عملکرد مفیدتری داشته باشند. اگر برخی داده‌های پرت در مجموعه داده وجود داشته باشد، مقیاس‌دهی داده‌ها و استفاده از توابع تبدیل کننده‌ی قوی تر مناسب است.

معمولا شکل و نحوه‌ی توزیع داده‌ها زیاد مورد توجه قرار نمیگیرند و اول مقدار میانگین داده ها را از انها کم میکنیم و بعد حاصل را بر انحراف معیار تقسیم میکنیم و به این روش داده ها را متمرکز میکنیم

تکنیک هایی که در این روش وجود دارند عبارت اند از :

1.باینری کردن

این تکنیک زمانی بکار میرود که میخواهیم مقدار های عددی را به منطقی تبدیل کنیم

2.حذف میانگین

این روش برای حذف میانگین داده ها از بردار ویژگی است که باعث نرمال سازی ویژگی ها با محدودیت0 شود.

3.مقیاس دهی

تابع scale روشی راحت و سریع برای انجام عملیات مقیاس دهی بکار میرود.

4.نرمال سازی

نرمال سازی برای این بکار میرود که ویژگی ها در مقیاس یکسانی قرار بگیرند و ما در اینجا با دو روش نرمال سازی رو به رو هستیم که شامل:

1.نرمال سازی L1

به این روش حداقل انحراف مطلق نیز می‌گویند، مقادیر ویژگی‌ها را به شکلی تغییر می‌دهیم که مجموع مقادیر مطلق در هر سطر حداکثر 1 باقی بماند.

2. نرمال سازی L2

به این روش حداقل مربعات نیز گفته می‌شود، مقادیر را به گونه‌ای تغییر می‌دهد که مجموع مربعات در هر سطر حداکثر 1 باقی بماند.

پیاده‌سازی SVM در sklearn

SVMمخفف Support Vector Machineاست که از متد های قدرتمند در زمینه طبقه بندی است.

Svm در ابعاد بزرگ دارای انعطاف پذیری است و در طبقه بندی داده بصورت گسترده مورد استفاده قرار میگیرد.مهم ترین هدف svm تقسیم مجموعه داده به تعدادی از کلاس ها برای بدست اوردن ابر صفحه و با بیشترین حاشیه است که در دو مرحله صورت میگیرد

که شامل: 1.ماشین بردار پشتیبان در ابتدا ابرصفحاتی را تولید می‌کند که نقاط را به شکل صحیح تقسیم می‌کند.

2. میان ابرصفحات آن ابرصفحه‌ای را انتخاب میکند که نقاط را به بهترین شکل ممکن جدا می‌کند.

ویژگی های مهم svm :

1. بردار های پشتیبان:

نقاطی هستند که به ابرصفحه نزدیک هستند.بردار های پشتیبان در شناسایی خطوط جدا کننده هستند

2. ابر صفحه:

صفحه ای که نقاط را به کلاس‌های مختلف تقسیم می‌کند.

3. حاشیه:

فاصله ی خالی بین نقاط مرزی دو کلاس یا همان بردارهای پشتیبان، را حاشیه یا لبه یا همان Margin گویند.

https://7learn.com/blog/scikit-learn-library-tutorail