مقدمه ای بر الگوریتم های داده کاوی

تمرین درس داده کاوی – استاد مهردوست

کیان رضایی

چرا داده کاوی؟

- فرض کنید شخصی از حساب بانکی خود مبلغ 50 میلیون تومان برداشت کند شما به عنوان رئیس بانک چه کار میکنید؟ در بانک های بزرگ در مواردی که نسبت به برداشت یک وجه مشکوک هستند سریعا با مشتری خود تماس میگیرند تا از صحت این تراکنش مطمئن شوند. یا فرض کنید شما به عنوان یک پزشک تصاویری از چشم مراجعه کنندگان در مقابل شما قرار دهند و از شما بخواهند با نگاه کردن به عکس ها و معلوماتی که دارید تشخیص دهید که آیا مراجعه کنندگان شما دارای بیماری چشمی هستند یا خیر!
- اما اگر کامپیوتر بتواند تشخیص دهد یک برداشت با توجه به اطلاعات مربوط به برداشت های قبلی یک حساب و سوابق شخص مشکوک است یا بتواند با دقت بالا تر از ماهر ترین پزشکان بیماری چشمی را تشخیص دهد چه اتفاقی ممکن است بیوفتد؟!

- امروزه داده کاوی و الگوریتم های آن این امکان را به ما میدهد تا در بسیاری از زمینه ها بتوانیم وظیفه هایمان را با سرعت و دقت بالا انجام دهیم. این الگوریتم ها با توجه به داده های موجود (داده آموزشی) یاد میگیرد و سپس اگر داده های جدیدی به آن دهیم میتوانند خروجی آن را پیشبینی کنند.

مراحل داده کاوی

- تمیز کردن داده
- انتخاب الگوریتم مناسب
- آموزش مدل
- پیشبینی یک کیس جدید

در ادامه راجع نحوه چگونگی انتخاب یک الگوریتم توضیح داده میشود.

مثال ها

- میزان CO_2 تولید شده توسط یک ماشین (رگرسیون)
- آیا سلول سرطانی است یا خیر؟ (دسته بندی)
- دسته بندی مشتریان در بانک (خوشه بندی)
- سیستم توصیه گر فیلم شرکت Netflix (توصیه گرها)

پایتون

در این سری از تمرین ها الگوریتم های داده کاوی را با زبان پایتون پیاده سازی میکنیم.

چرا؟

- بسیار آسان است.
- دارای کتابخانه هایی مانند Numpy and Pandas + SciKit
- مجانی و رایگان در اختیار است.

یادگیری ماشین نظارت شده و نظارت نشده

- یادگیری نظارت شده: الگوریتم در هنگام یادگیری دائما جواب درست را میگیرد و در این صورت میتواند که مدلی درست کند تا مثال های ناشناخته در آینده را پیشبینی کند. در واقع این مدل ماشین با استفاده از داده های برچسب گذاری شده و داشتن جواب های درست یاد میگیرند.

مثال های مختلفی از یادگیری ماشین با نظارت :

- یکی از مثال های مرسوم در یادگیری با نظارت تشخیص و فیلتر کردن اسپم ها میان پیام ها است. ابتدا تمامی داده ها به دو کلاس سالم و اسپم تقسیم می شوند، سپس ماشین آن ها را با مثال های موجود می آموزد در نهایت از او امتحان گرفته می شود و امتحان به این منظور تلقی می شود که شما ایمیل جدیدی که تا به حال ندیده است را به آن بدهید، سپس آن تشخیص دهد که سالم یا اسپم است.

مثال های مختلفی از یادگیری ماشین با نظارت :

- نمونه دیگری از این دست یادگیری می توان زد پیشبینی مقدار عددی می باشد، به عنوان مثال قیمت یک ماشین با مجموعه ویژگی هایی مثل (مسافت طی شده، برند، سن ماشین و ...). از این دست مثال ها با عنوان regression نامیده می شوند.

- برای آموزش سیستم، شما باید تعداد زیادی نمونه یا به عبارتی داده، در اختیار سیستم بگذارید که شامل label و predictor ها باشد.

- نکته: دقت کنید بعضی از الگوریتم های regression را می توانند در classification استفاده شوند و برعکس.

- برای مثال، رگرسیون منطقی (Logistic Regression) معمولا برای طبقه بندی استفاده می شود، زیرا می تواند یک مقدار را که مربوط به احتمال متعلق به یک کلاس داده شده است، تولید کند.

یادگیری بدون نظارت

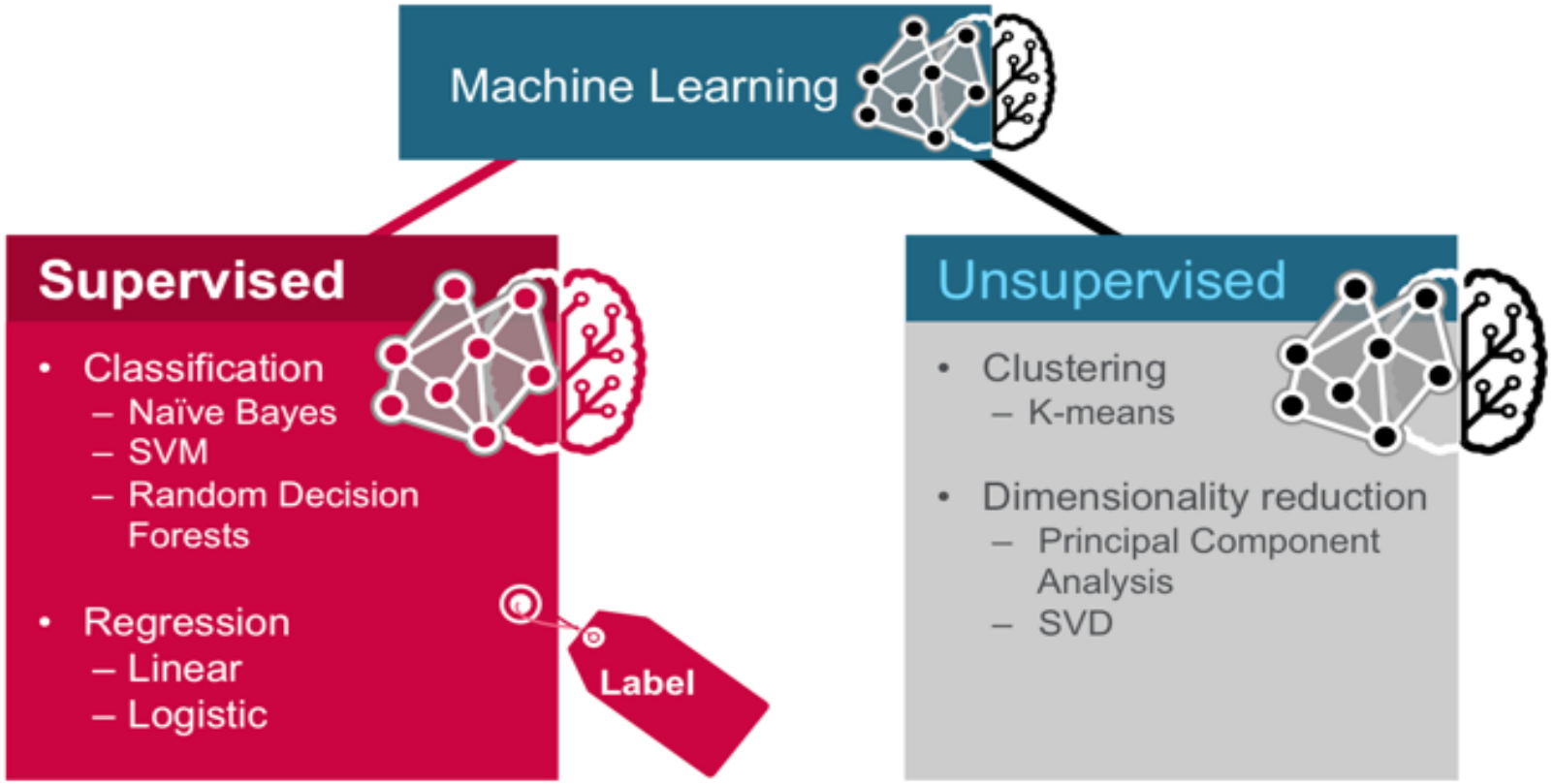
- این مدل ماشین بدون استفاده از داده های برچسب گذاری شده و بدون هیچ معلمی می آموزد که به آن Unsupervised Learning می گویند.
- درواقع به حالت ساده تر می توان گفت که در ابتدا تمامی نمونه هایی که به آن داده می شوند، هیچ برچسبی ندارند در صورتی که در یادگیری نظارتی تمامی داده ها برچسب داشتند. به عنوان مثال، ایمیل های اسپم و غیر اسپم. در یادگیری بدون نظارت برچسبی بر روی داده ها وجود ندارد.

مثال هایی از یادگیری بدون نظارت

- به عنوان مثال فرض کنید شما اطلاعات زیادی در رابطه با کاربران بلاگ شما دارید. شما ممکن است بخواهید یک الگوریتم خوشه بندی را اجرا کنید تا بتوانید تمامی ملاقات کننده های مشابه هم را در یک خوشه نگاه داری کنید.
- درواقع در مثال بالا شما بدون هیچ کمکی از سیستم می خواهید تشخیص دهد هر کاربر متعلق خوشه ای است و آن بدون کمک شما ارتباطات را پیدا می کند.
- فرض کنید که خروجی ماشین شما اینگونه است. 40 درصد ملاقات کنندگان شما مردهایی عاشق کتاب های علمی تخیلی هستند و عموما پست های بلاگ شما را غروب ها می خوانند. درحالی که 20 درصد آن ها جوان هایی علاقه مند به کتاب های داستانی و رمان هستند و بلاگ شما را تنها آخر هفته ها ملاقات می کنند. حتی شما می توانید با استفاده از hierarchical clustering algorithm در هر گروه آن ها را به گروه های کوچک تری تقسیم کنید و این کار ممکن به شما کمک بیشتری کند تا به هدفتان برسید.

- الگوریتم های مصور سازی یکی از مثال های دیگری هستند که برای یادگیری بدون نظارت می توان زد. به عنوان مثال شما تعداد بسیار زیادی به آن عکس می دهید و هیچ برچسبی هم بر روی عکس ها نمی گذارید. (مثلا عکس ماشین به آن می دهید و نمی گویید که این ماشین است). الگوریتم مورد نظر تلاش می کند تا ساختاری میان آن ها پیدا کند و آن ها را خوشه بندی کند. درنهایت متوجه می شوند که چطور داده را سازماندهی کنند.

نمای کلی


Machine Learning
Supervised
• Classification
– Naïve Bayes
– SVM
– Random Decision Forests
• Regression
– Linear
– Logistic
Label
Unsupervised
• Clustering
– K-means
• Dimensionality reduction
– Principal Component Analysis
– SVD