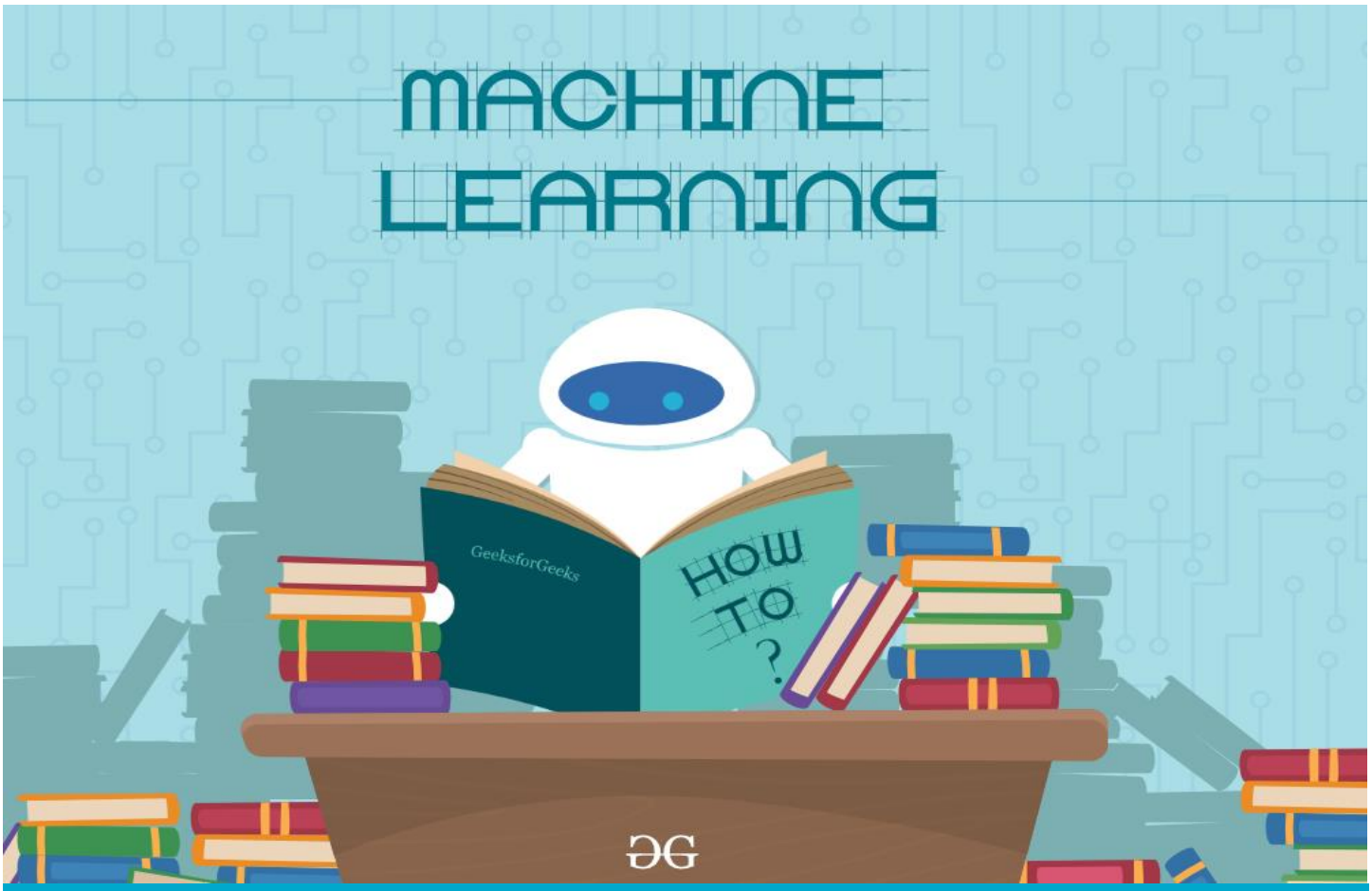


دوره جامع یادگیری ماشین

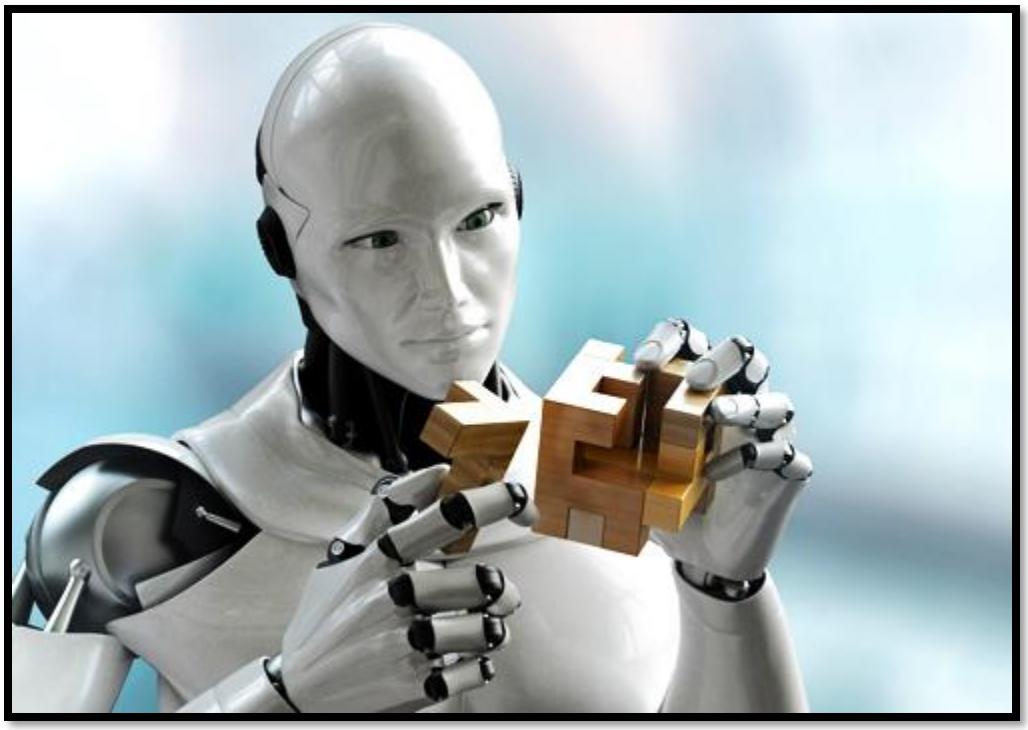
قسمت هشتم، تعداد کم داده ها، داده های غیر مرتبط، کیفت پایین داده ها، ابعاد غیر مرتبط

چالش های یادگیری ماشین ۱

Datalk.ir
Created by : Ali Arabshahi
Contact us : Linkedin.com/in/mrAliArabshahi

چالش های اصلی در یادگیری ماشین (۱)

از آن جایی که کار اصلی ما انتخاب الگوریتم یادگیری و همچنین آموزش (training) داده ها به ماشین بر اساس آن الگوریم می باشد، دو چالش مهم که با آن سر و کار خواهیم داشت، الگوریتم بد و همچنین داده های به درد نخور خواهد بود. بیاین **از دیتا های بد** شروع کنیم.



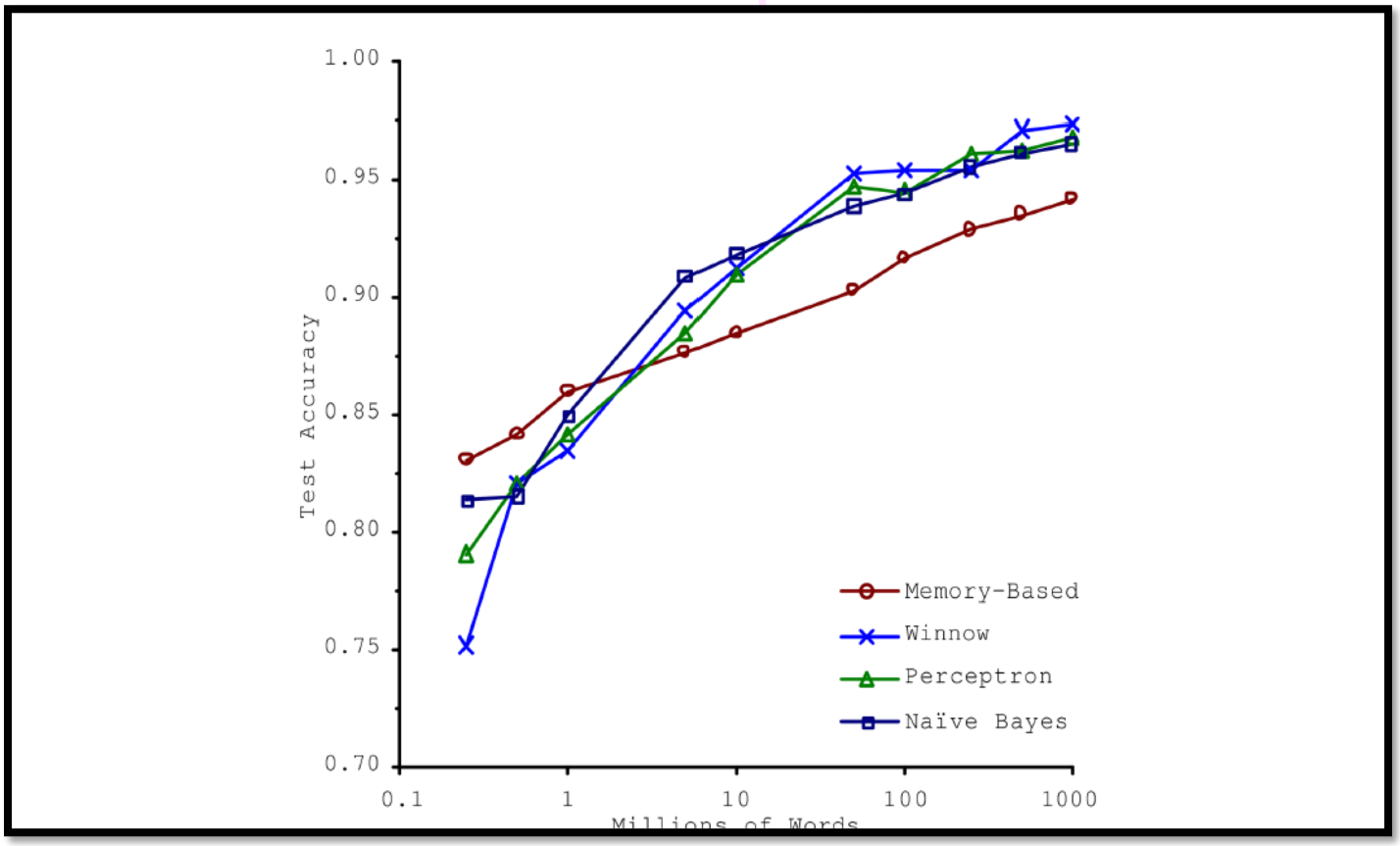
تعداد کم دیتاها

وقتی می خواین به یک کودک نوپای کوچولو موچولو! یاد بدین که سیب چیه، به یه سیب اشاره می کنین و میگین این سیبه 😊 و با چند بار تکرار کردن این کار، کوچولومون میتونه انواع و اقسام سیب ها رو در اندازه مختلف تشخیص بده.

اما یادگیری ماشین به این راحتی ها هم نیست. معمولا ماشین برای این که دقت مناسبی داشته باشه، حتی واسه ساده ترین مسائل هم باید کلی دیتا دریافت کنه و برای مسائل پیچیده تر (مثل پردازش تصویر و پردازش صوت) که چه بسی نیاز به دریافت میلیون ها داده داشته باشه!(مگر این که درون اون مدل ها از مدل های از پیش طراحی شده استفاده کنیم).

سال ۲۰۰۱ یک مقاله ای منتشر شد که خروجی اش این ایده بود که شاید حتی به جای این که به نوع مدلمون و جزییات خاص اون توجه کنیم شاید بهتر باشه تمرکزمون رو بزاریم روی جمع آوری دیتا!

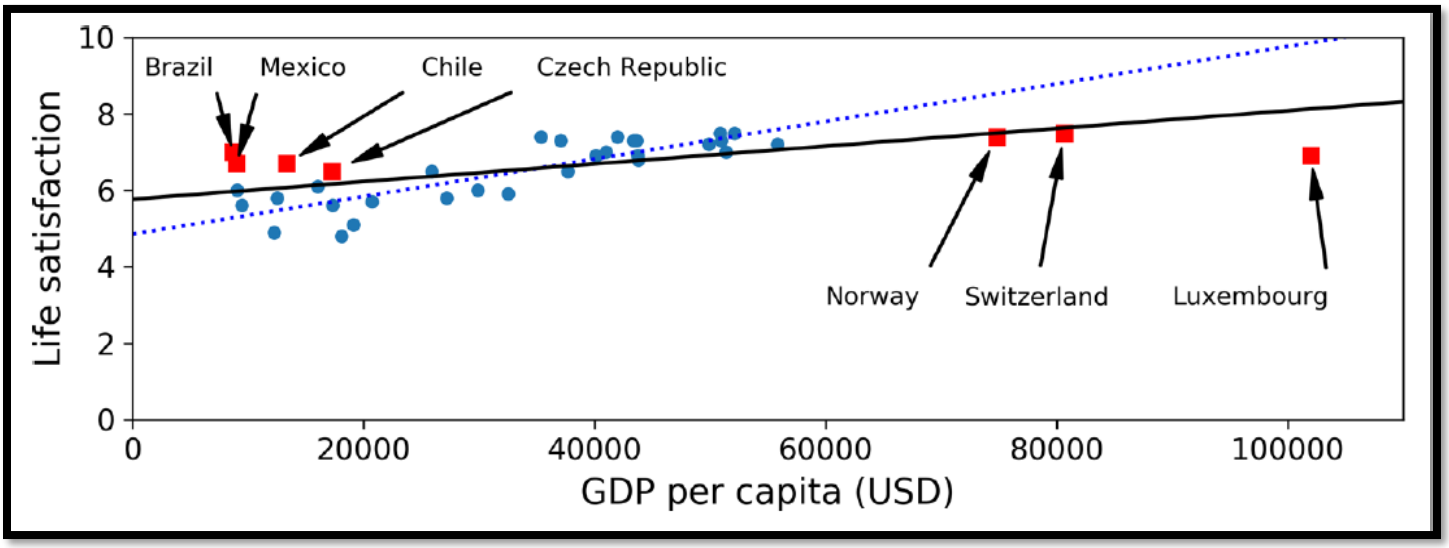
در تصویر زیر مشاهده می کنین که عملکرد مدل های مختلف با افزایش حجم دیتا چه قدر می تونه متفاوت و خارق العاده تر بشه. پس اینو یادتون باشه، همیشه برگ برنده دیتاست.



داده های غیر مرتبط

برای این که ماشینمون پیشبینی خوبی رو به ما ارایه بده، توجه به این نکته خیلی مهمه که داده هایی که به ماشین آموزش می دیم نماینده ای کامل باشن از همه ی داده هایی که در آینده به ماشین خواهیم داد تا مدل برامون پیشبینی شون کنه. (good generalization)

تصویر زیر رو در نظر بگیرین؛ یادتون هست در جلسات قبل به مدلی برای پی بردن به نحوه ی ارتباط نرخ رضایت با ثروت در کشور های مختلف رسیدیم. (خط نقطه چین)



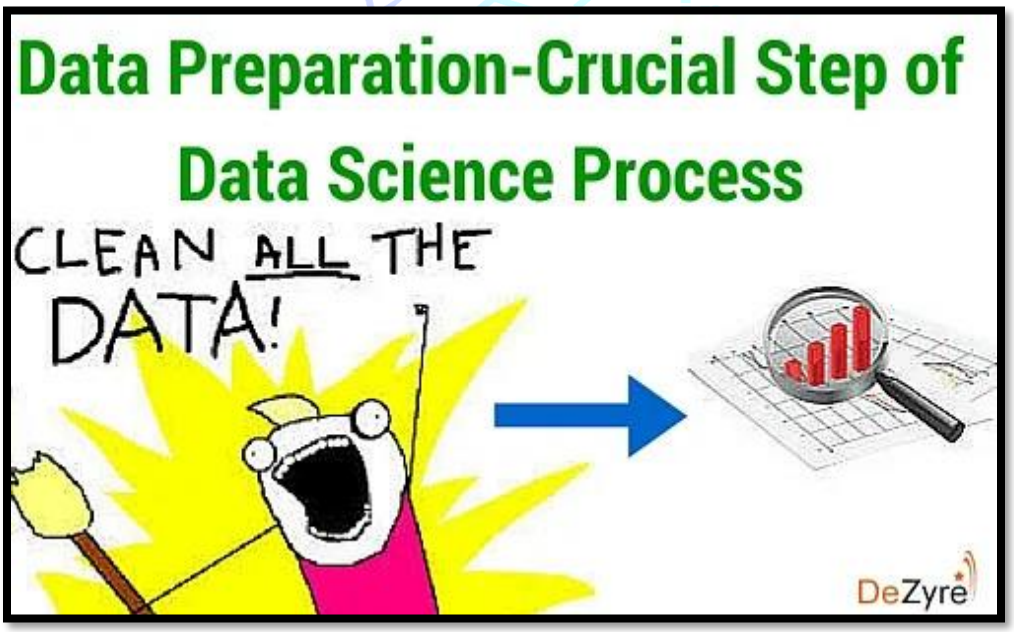
بعدها فهمیدیم که گویا یک سری دیتا رو فراموش کردیم به مدل اموزش بدیم. (مربع های قرمز). مجبور شدیم بریم و دوباره کل داده های قبلی به همراه داده های فراموش شده رو به ماشین آموزش بدیم. مدل به دست امده (خط ممتد) کاملا متفاوت از مدل اولمون هست. از مدل جدید میشه فهمید که در کشور های خیلی ثروتمند، میزان ارتباط GDP با احساس رضایت رابطه ی کمتری داره. پس اینجاست که متوجه میشیم **پول زیاد داشتن لزوما خوشبختی نمی یاره! اما تا یه حدیش قطعا می یاره!!** 😎

یه مثال خیلی معروف دیگه ای هم که وجود داره: در یک انتخاباتی در فلان سال! با بررسی داده ها به این نتیجه می رسن که فلان فرد برنده میشه!! اما بعدا متوجه شدن که نظرسنجی هاشون از یک قشر خاصی بوده که عموما جز افراد پولدار جامعه شون محسوب می شدن. خلاصه این که در آخر فلان فرد برنده نمی شه! 🤦‍♀️

در مجموع خیلی باید مراقب این نکته باشیم که داده هامون نشون دهنده ی **کل جامعه** باشن و نه صرفا یک قشر به خصوصی.

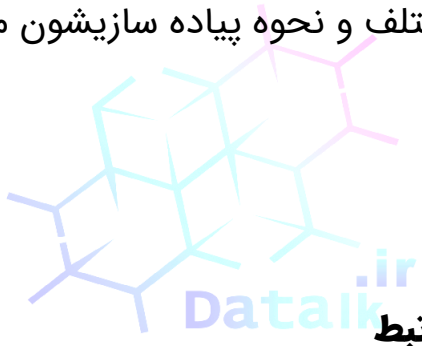
کیفیت پایین داده ها

حتما همتون شنیدین که یه دیتا ساینتیست واقعی معمولا بیشترین وقتش رو صرف آماده سازی داده ها می کنه تا مسائل دیگه. همونطور که تصویر زیر به خوبی گویای این مسئله هست. 😃



حالت های زیر، مواردی هستن که در ارتباط با کیفیت پایین داده ها ممکنه باهاشون روبه رو بشیم:

- داخل دیتا هاتون ممکنه یکسری از اونا مقادیر غیر متعارفی داشته باشد (outliers)؛ که معمولا باید بگردیم و بعد از پیدا کردنشون اونا رو حذف کنیم.
- بعضی از داده ها ممکنه خالی (missing value) باشن. مثلا کاربرمون فراموش کرده سنش رو وارد کنه. در ارتباط با این مقادیر خالی در دیتا ست می تونیم رویکرد های متفاوتی رو اتخاذ کنین. از حذف کردنشون گرفته، تا جایگزاریشون با مقادیر دیگه (مثلا میانگین سن کل افراد). پیشنهاد می کنم حتما یه نیم نگاهی به **دوره پایتون در یادگیری ماشین** مون بندازین که اونجا راجع به این رویکرد های مختلف و نحوه پیاده سازیشون مفصل صحبت کردیم.



فیچر ها یا ابعاد یا (X) های غیرمرتبط

همیشه این رو در نظر بگیرین که ماشین فقط زمانی می تونه خوب یادبگیره که الگو های خوبی داشته باشه. فیچر های شما (یادتون هست فیچر چی بود دیگه؟ می خوام قیمت خونه رو پیشبینی کنم، فیچر هاش میتونه تعداد اتاق ها، متراژ، سن هر خونه و ... باشه)باید هم کافی باشن و هم بیش از حد با هم بی ارتباط نباشن. پس یکی از وظایف اصلی شما انتخاب و یا ایجاد فیچر های مناسب برای ماشین هاست که به این پروسه مهم، **(feature engineering)** گفته میشه که میتونه شامل فرایند های زیر باشه:

- انتخاب فیچر(feature selection): انتخاب مناسب ترین فیچر ها از بین فیچر های موجود

- استخراج فیچر(feature extraction): ترکیب و یا ادغام فیچر های موجود که اگه یادتون باشه الگوریتم کاهش ابعاد برای این کار می تونه مفید باشه.
- ایجاد فیچرهای جدید: با جمع آوری داده های مورد نیا فیچر های جدید و تاثیر گذار را ایجاد می کنیم.

خب تا اینجا به بررسی چالش های ماشین لرنینگ که در ارتباط با دیتا ها وجود داره، پرداختیم. به امید خدا در قسمت بعدی **چالش های الگوریتمیک** رو در کنار هم واکاوی می کنیم.

پایدار و سلامت باشین 😊