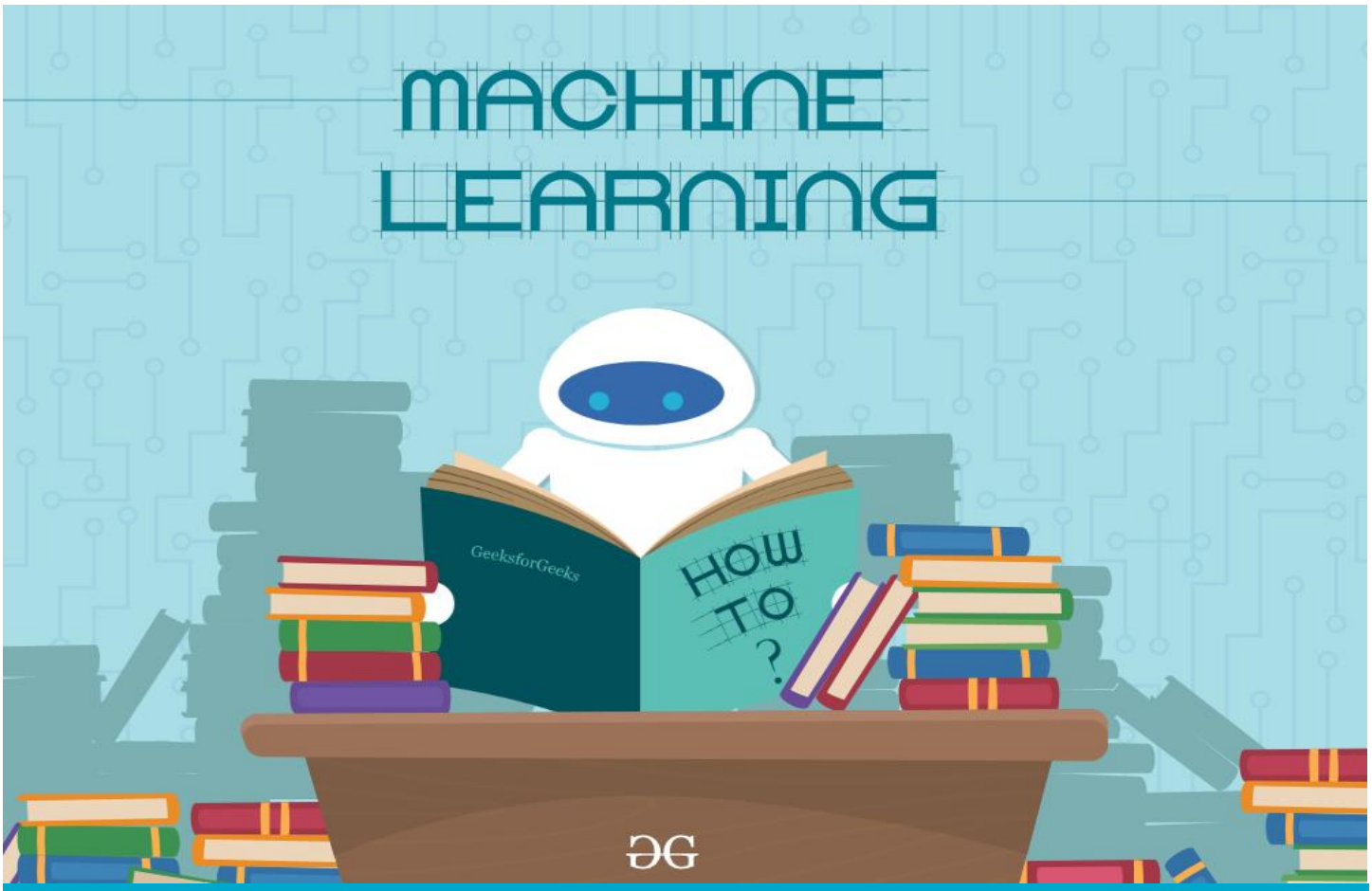


دوره جامع یادگیری ماشین

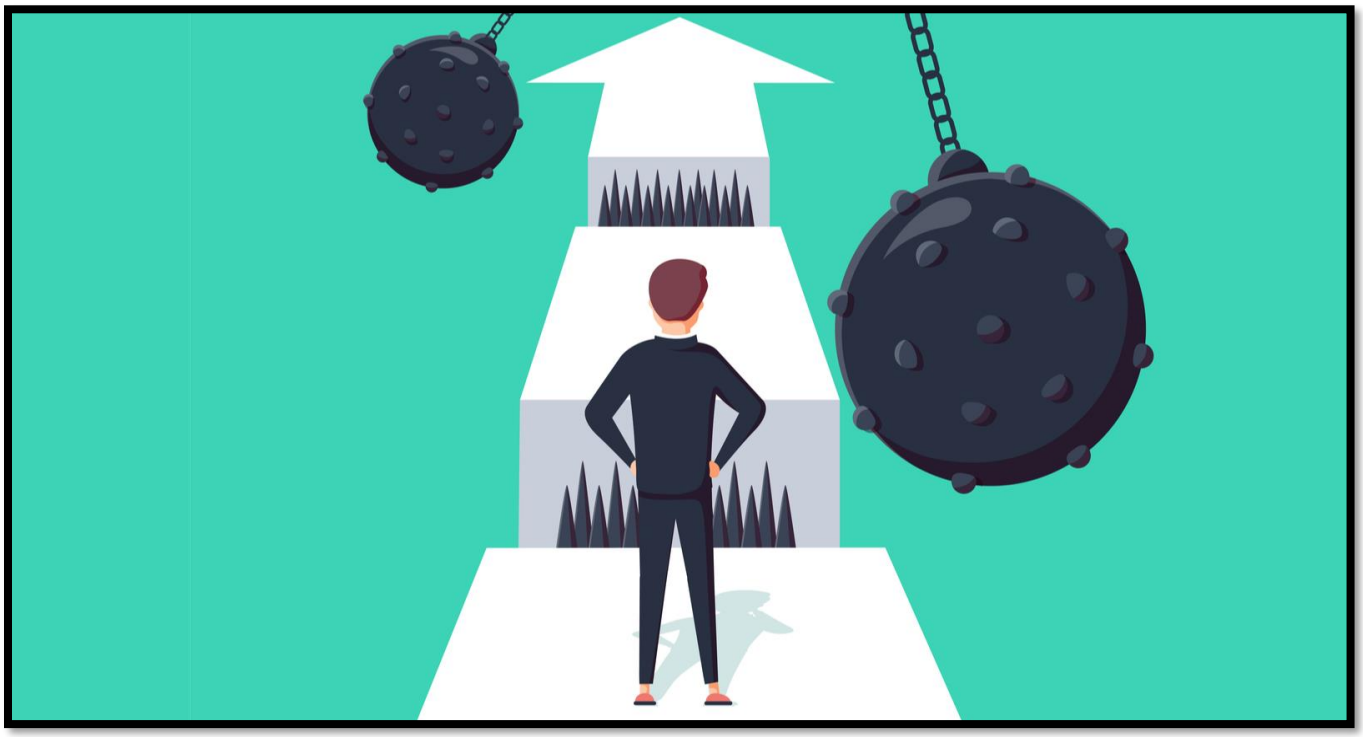
قسمت نهم، اُور فیتینگ و آندر فیتینگ!

چالش های یادگیری ماشین ۲

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

چالش های اصلی در یادگیری ماشین (2)

فهمیدیم که دو چالش مهمی که با آن سر و کار خواهیم داشت، الگوریتم بد و همچنین داده های به درد نخور خواهد بود. جلسه گذشته به بررسی چالش های مرتبط با دیتا پرداختیم و در ادامه، چالش های مربوط به الگوریتم ها را مورد بررسی قرار می دهیم.



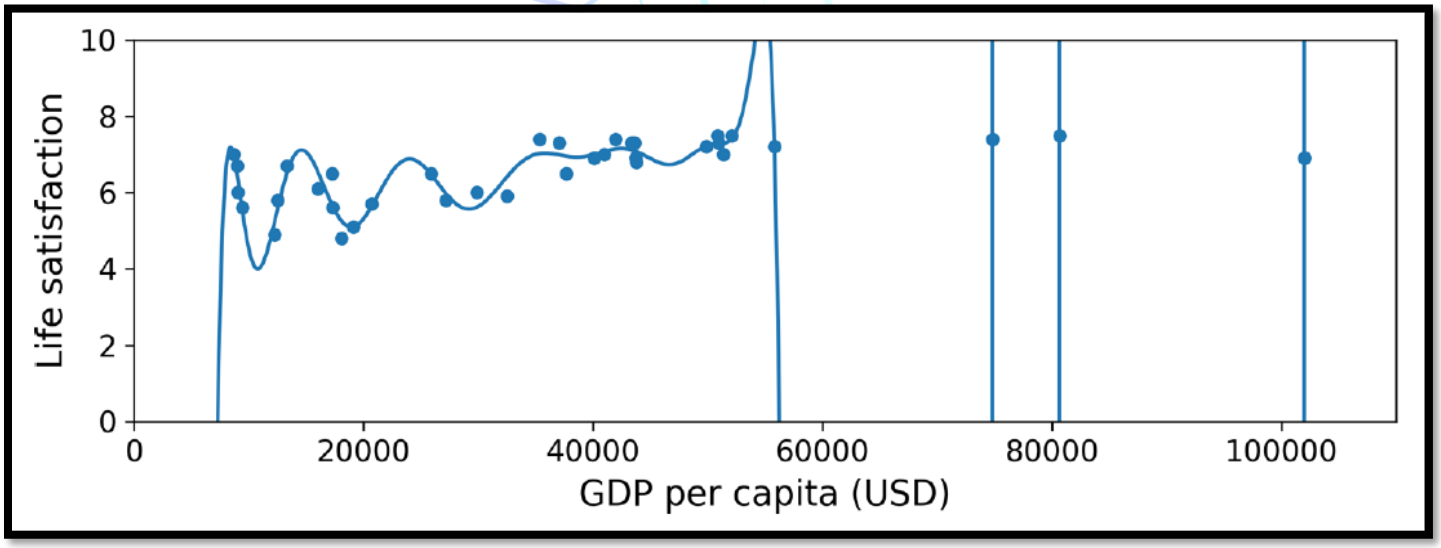
Overfitting the Training Data

فرض کنین به یه شهری سفر می کنیم و زمانی که سوار تاکسی می شیم، راننده قسمتی از پول مون رو بهمون بر نمی گردونه؛ ممکنه این اتفاق این نتیجه گیری رو در پی داشته باشه که همه راننده های این شهر بدجنس هستن! خداییش این نتیجه گیری درسته؟ هم بله و هم نه!!

در واقع ما یک سری دیتا دریافت می کنیم.(هر سری که سوار یک تاکسی در اون شهر شدیم) نتیجه گیریمون بر اساس اون دیتا کاملا صحیحه، اما زمانی که می خواینم نتیجه گیریمون رو به کل راننده های شهر تعمیم بدیم (Generalization)، اینجاست که دچار اشتباه می شیم.

متاسفانه ماشین ها از این اشتباهات مستثنی نیستند. به این خطا (Overfitting) می گیم، یعنی زمانی که ماشین نسبت به داده های آموزش (training data) درست کار می کنه اما نسبت به داده های جدید (when we try to generalize the model)، نتیجه اصلا به اون چیزی که باید، شبیه نیست.

واسه نمونه خروجی زیر که نشون دهنده ارتباط بین GDP و نرخ رضایت رو در نظر بگیرین. مدل کاملا بر روی training data، منطبق هست. پس به نظر عالی داره کار می کنه اما واقعا می تونیم به این ماشین برای پیشبینی نمونه جدید اعتماد کنیم؟

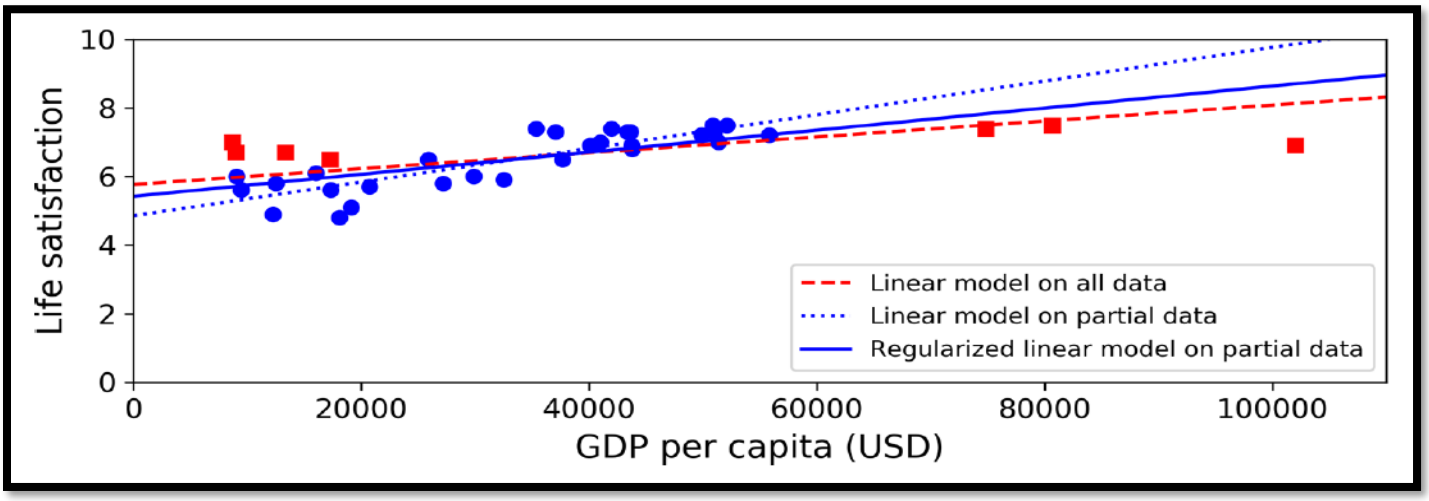


به طور کلی overfitting زمانی اتفاق می افته که مدلمون بیش از حد به نویز های داده ها وابسته باشه. راه حل های زیر می تونن برای حل این مشکل، کارگشا باشن:

- **ساده تر کردن** مدل، مثلا کاهش تعداد پارامتر ها (features) و یا **محدود کردن** اثر اون ها
- جمع آوری دیتا های بیشتر
- کاهش نویز داخل داده ها، به عنوان مثال حذف مقادیر outliers

محدود کردن مدل برای ساده تر کردن اون (simplify) و کاهش ریسکِ overfit شدنش با عنوان **Regularization** شناخته میشه. روال کلی به این صورته که به جای حذف پارامتر های مختلف (پاک کردن صورت مسئله!)، میایم و اثر پارامتر های مختلف رو کمتر می کنیم. اینطوری انگار هم هستن و هم اونقدری قدرت ندارن که مدل رو به خودشون overfit کنن. با یه تیر، دو نشون می زنیم!😎👌

در تصویر زیر سه مدل با سه روکرد متفاوت یعنی حذف برخی داده ها(نقطه چین)، در نظر گرفتن همه داده ها (خط چین) و حالت سوم یعنی حفظ همه داده ها و کاهش اثر پارامتر ها (خط ممتد) رو مشاهده می کنین و کاملا مشخصه که regularization می تونه به خوبی شانس overfit شدن رو کاهش بده.



دونستن این نکته هم خالی از لطف نیست که میزان اثر کاهش فیچر ها در مدل ها با پارامتری تحت عنوان **hyper-parameter** تنظیم میشه. اگر این مقدار خیلی زیاد باشه مدل میره به سمت تبدیل شدن به یک خط صاف (با شیب صفر). اگر این طوری باشه قطعا مدلتون overfit نخواهد شد اما به نظرتون می تونه خوب کار کنه؟ فردی رو در نظر بگیرین که هیچ چیزی از دیگران یاد نمیگیره و دوست داره خودش همه چی رو امتحان کنه! خب قطعا چنین آدمی سرشار هست از کلی اشتباه رنگارنگ **پس high – hyper – parameter نباشیم.** 🥲

Underfitting the Training Data

حدستون کاملا درسته. این پدیده دقیقا برعکس overfitting هست. زمانی اتفاق می افته که مدلمون بیش حد ساده باشه طوری که حتی نتونه پیشبینی خوبی برای دیتا های آموزش (trainig data) باشه. در اصل اینقدر ماشینمون خنگه که هیچی بلد نیس! خنگ نیست البته، اون طور که باید یاد نگرفته. مشکل از اون افرادیه که بهش آموزش دادن(بلانسبت شما البته 😃).

برای حل این خطا میتونیم رویکرد های زیر رو در نظر بگیریم:

- تابع پیچیده تری انتخاب کنیم مثلا به جای یه مدل خطی ساده، یه مدل چند جمله ای رو امتحان کنیم.
- فیچر های بهتری رو برای مدل امتحان کنیم. مثلا برای پیشبینی قیمت خونه، علاوه بر متراژ، تعداد اتاق ها رو هم به مدل بخورونیم!
- کاهش محدودیت ها، مانند کم تر کردن hyper – parameter برای افزایش اثر پارامتر ها.

با هم مهم ترین چالش ها از دید الگوریتم ها رو بررسی کردیم. در جلسه بعدی یک مروری بر روی آنچه که تا الان یادگرفتیم خواهیم کرد و همچنین میریم برای **قدم نهایی یعنی ارزیابی عملکرد مدل!**

به امید دیدار 🤠