

پی نوشتی درباره استقلال توی مساله spam

سلام!

یه نکته ای که شاید توی صورت سوال خیلی تاکید نشده آنقدر که باید و شاید
اینه که در حالت عادی این free, money و ... مستقل از هم نیستند.
چون اگه همه چی از هم مستقل باشه که دیگه مشاهده اش کمکی نمیکنه
یعنی اینطوری نیست که بگیم

$$P(\text{Spam}, \text{free}, \text{money}) = P(\text{spam})P(\text{free})P(\text{money})$$

اینطوری که نیست جدا! چون خوب اگه اینطوری بود دونستن اینکه توی متن free اومده به هیچ دردی نمیخورد
بلکه اینطوریه:

$$P(\text{Spam}, \text{free}, \text{money}) = P(\text{Spam})P(\text{free}, \text{money}|\text{Spam})$$

که حالا فرض naïve bayes که فرض درستی هم نیست استفاده میشه

$$\begin{aligned} P(\text{Spam})P(\text{free}, \text{money}|\text{Spam}) \\ = P(\text{Spam})P(\text{free}|\text{Spam})P(\text{money}|\text{Spam}) \end{aligned}$$

یادتون باشه که الان بیاین توی جدول ها حساب کنیم این رابطه احتمالا برقرار نیست:

$$P(\text{free}, \text{money}|\text{Spam}) = P(\text{free}|\text{Spam})P(\text{money}|\text{Spam})$$

ولی اشکال نداره! اگه **واقعا** این رابطه استقلال برقرار بود و اگه **میلیارد** ها ایمیل میومد و احتمال رو اونجا حساب
میکردین این رابطه بالا برقرار میشد.

ولی خوب الان تعداد ایمیل ها کمه و شاید اصلا فرض ما هم غلط باشه!

برای اینکه یکم بیشتر قانع بشید که فرضمون اونقدر ها هم مسخره و لوس نیست، این مثال رو میزنم:
فرض کنید یه اتاق داریم که توش یه بخاری هست.

دو تا دما سنج هم داریم تو دو تا گوشه اون اتاق

متغیر $x_1=1$ یعنی گرم. $x_1=0$ یعنی سرد. $x_2=1$ گرم و $x_2=0$ یعنی سرد.

$Y=1$ روشن بوده بخاری و $Y=0$ بخاری خاموش بوده.

دمای دما سنج ها و روشن بودن بخاری رو توی یک سال اندازه گرفتیم و احتمال ها رو با این اطلاعات حساب کردیم.

حالا بدیهی هست که مستقل نیستن x_1 و x_2 :

$$P(x_1 = 1, x_2 = 0) \neq P(x_1 = 1)P(x_2 = 0)$$

اما اگه بدونیم بخاری روشن بوده، تقریباً هردو سر اتاق گرمایشون شبیه هم خواهد بود. یعنی رابطه زیر برقراره:

$$P(x_1, x_2|Y) = P(x_1|Y)P(x_2|Y)$$

Naïve bayes هم تقریباً به همین چیزیه. میگه اگه بدونی بخاری روشنه (Spam هست) اون موقع تقریباً وضعیت دمای دو نقطه مشخصه و داشتن اطلاع درمورد دمای یه نقطه اونقدر اطلاع جدید در مورد نقطه دیگه نمیده.