- دلیل اینکه بعضی ها احتمال بیشتر از یک در میآورند

ممکن است بپرسید مشکل از کجا بوده که اگر مخرج کسر را حساب کنیم، احتمال ممکن است بیشتر از ۱ شود.

توضیح:

ما در این مساله از فرض naive استفاده کردیم. در نتیجه بعضی از احتمال ها را با روش عادی نباید حساب کنیم.

یکی از اونا، این کسره قرمزه هست:

شما در واقع باید مقایسه کنید. احتمال Spam بودن و احتمال NotSpam را:

$$P(Spam|a,b,c,d) = \frac{1}{p(a,b,c,d)} * \big(P(Spam) * P(a,b,c,d|Spam)\big)$$

$$P(Not\ Spam|a,b,c,d) = \frac{1}{p(a,b,c,d)} * \big(P(Not\ Spam) * P(a,b,c,d|Not\ Spam)\big)$$

در هنگام مقایسه، حساب کردن این مقدار قرمز لازم نیست. بقیه احتمال ها را نیز با روش های گفته شده حساب کنید.

واسه این لازم نیست چون یکسانه دو طرفه

یعنی کافیه که این مقایسه رو انجام بدید:

$$\frac{1}{p(a,b,c,d)} * \big(P(Spam) * P(a,b,c,d|Spam)\big) \geq^{?} \frac{1}{p(a,b,c,d)} * \big(P(Not\ Spam) * P(a,b,c,d|Not\ Spam)\big)$$

که میشه:

$$\big(P(Spam) * P(a,b,c,d|Spam)\big) \geq^{?} \big(P(Not\ Spam) * P(a,b,c,d|Not\ Spam)\big)$$

حالا اگه اصرار دارید که این احتمال را حساب کنید،

یک راه این است که:

$$P(Spam|a,b,c,d) + P(Not\ Spam|a,b,c,d) = 1$$

راه دیکر هم این است که:

$$p(a,b,c,d) = p(a,b,c,d,spam) + p(a,b,c,d,not\ spam) =$$

$$p(spam) * p(a,b,c,d|Spam) + p(not\ spam) * p(a,b,c,d|not\ Spam)$$

ولی

این راه غلطه که بگیم:

$$p(a,b,c,d) = \frac{\#(a,b,c,d)}{\#(mails)}$$

حالا یه مثال که توش استفاده از این رابطه آبیه مجاز باشه

در حالتی که فرض naive فرض واقعا درستی باشد مثلا حالت زیر:

Data	a	b	Class
1	0	0	Spam
2	0	1	Spam
3	1	0	Spam
4	1	1	Spam
5	0	0	Spam
6	0	1	Spam
7	1	0	Spam
8	1	1	Spam
9	0	1	Not Spam
10	1	1	Not Spam
11	0	1	Not Spam
12	1	1	Not Spam

حال احتمال

$$p(a,b) = \frac{\#(a,b,c,d)}{\#(mails)} = \frac{4}{12}$$

و همچنین

$$p(a,b) = p(spam) * p(a,b|Spam) + p(not\ spam) * p(a,b|not\ Spam) = \frac{4}{12}$$

چون این دو رابطه واقعا بر قرار است:

$$p(a,b|Spam) = p(a|Spam) * p(b|Spam) = \frac{1}{4}$$

$$p(a,b|not\ Spam) = p(a|not\ Spam) * p(b|not\ Spam) = \frac{1}{2}$$

پس وقتی این رابطه تو داده ها برقرار نباشه، که در تو تمرین برقرار نیست از فرمول آبی استفاده نکنید!