9

برچسب	نوع فایل ضمیمه	تعداد خطوط	موضوع	تعداد کلمات	طول عنوان
اسپم	Pdf	10	تبلیغات	45	15
اسپم	Pdf	7	خرید	30	20
غیراسپم	Txt	6	خرید	25	10
اسپم	Txt	8	فروش	60	30
اسپم	Txt	9	فروش	50	25
غیراسپم	Png	7	خرید	40	18
غیراسپم	Txt	8	خرید	20	12
اسپم	Pdf	9	تبلیغات	35	22
اسپم	Pdf	7	فروش	28	17
غیراسپم	Txt	6	خرید	22	14
اسپم	Png	12	تبلیغات	55	27
اسپم	Png	7	فروش	48	23
غیراسپم	Pdf	8	تبلیغات	37	19
غیراسپم	Png	10	خرید	30	16

- با کمک الگوریتم بیز ساده پیش‌بینی کنید که آیا ایمیل با مشخصات زیر اسپم خواهد بود یا خیر؟ توزیع ویژگی‌های پیوسته را نرمال درنظربگیرید.

$(18, 45, \text{خرید}, 15, Pdf)$

فاطمه شاکری، دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر

خرید

$P(\text{خرید} \mid \text{اسپم}) = 1/8$

$P(\text{خرید} \mid \text{غیراسپم}) = 5/6$

Pdf

$P(Pdf \mid \text{اسپم}) = 4/8 = ½$

$P(Pdf \mid \text{غیراسپم}) = 1/6$

طول عنوان

$\mu(\text{اسپم}) = (15 + 20 + 30 + 25 + 22 + 17 + 27 + 23)/8 = 22.375$

variance(اسپم) =
sqrt((54.391+5.641+58.141+6.891+0.141+28.891+21.391+0.391)/7)
=5.012

$P(\text{طول عنوان} = 18 \mid \text{اسپم}) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}}{\sigma_i\sqrt{2\pi}} = 0.05438$

μ(غیر اسپم) = (10+18+12+14+19+16)/6 = 14.833

Variance (غیراسپم) = sqrt((23.357+10.029+8.025+0.694+17.363+1.361)/5) = 3.488

P(طول عنوان =18 | غیراسپم) $\frac{e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}}{\sigma_i\sqrt{2\pi}}$ = 0.07573

تعداد کلمات

μ(اسپم) = (45+30+60+50+35+28+55+48)/8 = 43.875

variance(اسپم) = sqrt((1.266 + 192.516 + 260.016 + 37.516 + 78.766 + 252.016 + 123.766 + 17.016)/7)

= 11.728

P(تعداد کلمات = 45 | اسپم) $\frac{e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}}{\sigma_i\sqrt{2\pi}}$ = 0.03386

μ(غیر اسپم) = (25 + 40 + 20 + 22 + 37 + 30)/6 = 29

Variance (غیراسپم) = sqrt((16 + 121 + 81 + 49 + 64 + 1)/5) = 8.149

P(تعداد کلمات =45 | غیراسپم) $\frac{e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}}{\sigma_i\sqrt{2\pi}}$ = 0.00712

تعداد خطوط

μ(اسپم) = (10+7+8+9+9+7+12+7)/8 = 8.625

variance(اسپم) = sqrt((1.891 + 2.641 + 0.391 + 0.141 + 0.141 + 2.641 + 2.641 + 11.391)/7)

= 1.768

P(تعداد خطوط=15| اسپم) $\frac{e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}}{\sigma_i\sqrt{2\pi}}$ = 0.00033

μ(غیر اسپم) = (6 + 7 + 8 + 6 + 8 + 10)/6 = 7.5

Variance (غیراسپم) = sqrt((2.25 + 0.25 + 0.25 + 2.25 +0.25 + 6.25)/5) = 1.517

P(تعداد خطوط=15 | غیراسپم) $\frac{e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}}{\sigma_i\sqrt{2\pi}}$ = 0.00000129489

P(اسپم | (18, 45 , خرید , 15, Pdf)) relative to:

8/14 * 0.05438 * 0.03386 * 1/8 * 0.00033 * ½ = 2.17011 * 10^(-8)

P(غیر اسپم| (18, 45 , خرید , 15, Pdf)) relative to:

6/14 * 0.07573 * 0.00712 * 5/6 * 0.00000129489 * 1/6 =

4.15596*10^(-11)

بنابراین تحت شرایط خواسته شده پیشبینی این است که ایمیل اسپم باشد.