بیز ساده

سید ناصر رضوی n.razavi@tabrizu.ac.ir

۱۳۹۵

شبکه‌های بیز ساده

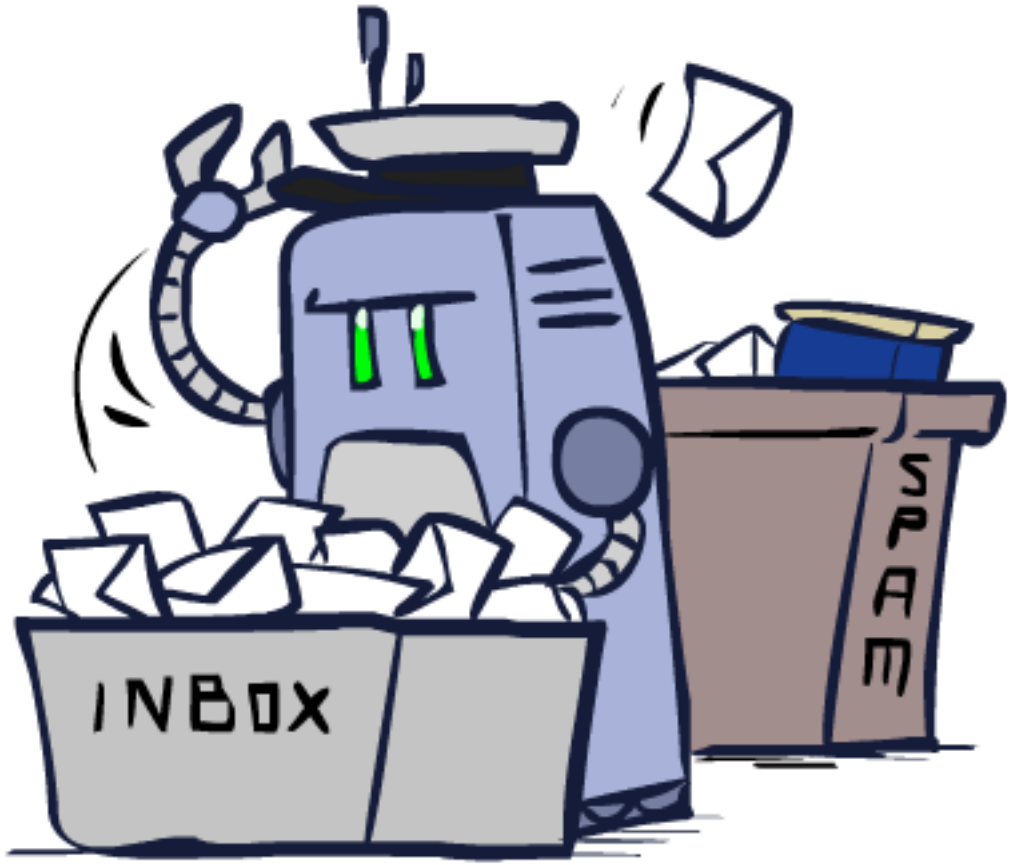
یادگیری ماشین

- تا کنون. چگونگی استفاده از یک مدل برای گرفتن تصمیم‌های بهینه.

- یادگیری ماشین. چگونگی ایجاد یک مدل از روی داده‌ها و تجارب.
  - یادگیری پارامترها [مانند احتمالات]
  - یادگیری ساختار [مانند گراف شبکه بیز]
  - یادگیری مفاهیم مخفی [مانند خوشه‌بندی]

- امروز. کلاس‌بندی مبتنی بر مدل با شبکه‌های ساده‌ی بیز!

کلاس‌بندی



مثال: تشخیص هرزنامه

- ورودی. یک ایمیل
- خروجی. هرزنامه؟ بله / خیر

- راه‌اندازی.
  - جمع‌آوری یک مجموعه بزرگ از ایمیل‌ها که کلاس هر کدام مشخص شده است.
  - هدف: پیش‌بینی کلاس ایمیل‌های جدید.

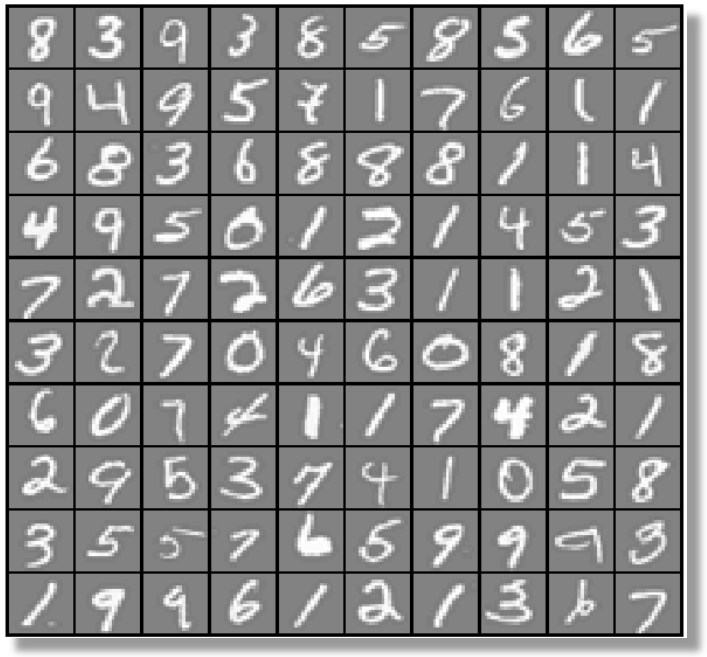
- ویژگی‌ها. صفات استفاده شده برای تصمیم‌گیری درباره کلاس ایمیل.
  - واژه‌ها: مانند رایگان، پول و ...!
  - الگوهای متنی: حروف بزرگ، دلار و بعد دو رقم.
  - خصوصیات غیرمتنی: وجود فرستنده در لیست تماس.

Dear Sir.
First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.
99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

مثال: تشخیص ارقام دست‌نویس



- ورودی. تصویر مربوط به یک رقم.
- خروجی. یک رقم از ۰ تا ۹.

- راه‌اندازی.
  - جمع‌آوری یک مجموعه بزرگ از تصاویر که کلاس هر کدام مشخص شده است.
  - هدف: پیش‌بینی کلاس مربوط به تصاویر جدید.

- ویژگی‌ها. صفات استفاده شده برای تصمیم‌گیری درباره کلاس رقم.
  - پیکسل‌ها: پیکسل (۸, ۶) برابر با ۱ است.
  - الگوهای شکلی: تعداد مولفه‌ها، تعداد حلقه‌ها، نسبت طول به عرض.

چند مثال دیگر از کلاس‌بندی

- کلاس‌بندی. با داشتن ورودی x، برچسب (کلاس) y را پیش‌بینی کن.

- مثال‌ها.
  - تشخیص پزشکی.
    - ورودی: نشانه‌ها
    - خروجی: نوع بیماری
  - نمره‌دهی خودکار به یک انشا.
    - ورودی: یک سند متنی
    - خروجی: یک نمره
  - تشخیص کلاه‌برداری.
    - ورودی: یک تراکنش مالی
    - خروجی: کلاه‌برداری یا خیر
  - ...

کلاس‌بندی یک فناوری تجاری بسیار مهم است!

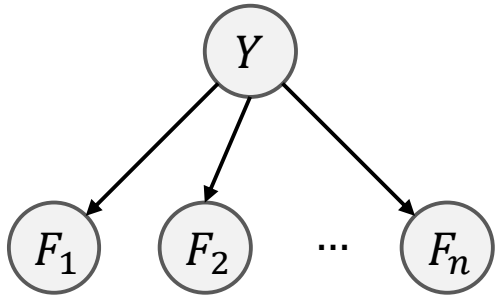
کلاس‌بندی مبتنی بر مدل

کلاس‌بندی مبتنی بر مدل

- کلاس‌بندی مبتنی بر مدل.
  - ایجاد یک مدل (مثلاً شبکه‌ی بیز) به طوری که در آن هم ویژگی‌ها و هم برچسب کلاس‌ها هر کدام یک متغیر تصادفی هستند.
  - مقداردهی به ویژگی‌های مشاهده شده.
  - انجام پرس و جو برای برچسب با توجه به ویژگی‌ها.
- چالش‌ها.
  - ساختار شبکه‌ی بیز چگونه باید باشد؟
  - پارامترهای شبکه‌ی بیز چگونه باید یاد گرفته شوند؟

بیز ساده برای تشخیص ارقام

- بیز ساده. فرض می‌کند تمام ویژگی‌ها اثرات مستقل برچسب کلاس هستند.



- نسخه ساده تشخیص ارقام.
  - یک ویژگی F_{ij} به ازای هر پیکسل $\langle i,j \rangle$.
  - مقدار هر ویژگی برابر با ۰ یا ۱ است.
  - هر تصویر ورودی به یک بردار ویژگی نگاشت می‌شود:

$$\rightarrow \langle F_{0,0}=0, F_{0,1}=0, F_{0,2}=1, F_{0,3}=1, F_{0,4}=0, \dots, F_{15,15}=0 \rangle$$

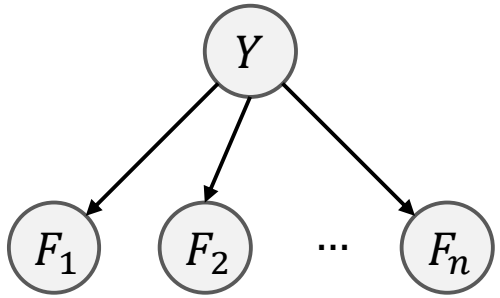
- مدل بیز ساده.

پارامترهایی که باید یاد گرفته شوند

$$P(Y|F_{0,0}, \dots, F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

مدل عمومی بیز ساده

- مدل عمومی بیز ساده.



$$P(Y, F_1, \dots, F_n) = P(Y) \prod_i P(F_i|Y)$$

اندازه نمایی — اندازه خطی

- تنها باید مشخص کنیم هر ویژگی چگونه به کلاس وابسته است.
- تعداد کل پارامترها برحسب تعداد ویژگی‌ها خطی است.
- اگرچه این مدل یک مدل بسیار ساده شده است، اما اغلب به خوبی کار می‌کند.

استنتاج در بیز ساده

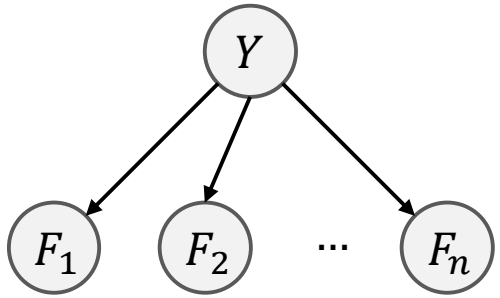
- هدف. محاسبه احتمال پسین برای متغیر کلاس Y.
  - گام ۱: ایجاد توزیع توأم برچسب و شواهد به ازای هر برچسب.

$$P(Y, F_1, \dots, F_n) = \begin{bmatrix} P(y_1, f_1, \dots, f_n) \\ P(y_2, f_1, \dots, f_n) \\ \vdots \\ P(y_k, f_1, \dots, f_n) \end{bmatrix} = \begin{bmatrix} P(y_1)\Pi_i P(f_i|Y) \\ P(y_2)\Pi_i P(f_i|Y) \\ \vdots \\ P(y_k)\Pi_i P(f_i|Y) \end{bmatrix} \Big) + \quad P(f_1, f_2, \dots, f_n)$$

  - گام ۲: محاسبه مجموع برای به دست آوردن احتمال شواهد. $P(f_1, f_2, \dots, f_n)$
  - گام ۳: نرمال‌سازی با تقسیم نتیجه گام ۱ بر نتیجه گام ۲. $P(Y|f_1, \dots, f_n)$

مدل عمومی بیز ساده

- س. برای استفاده از بیز ساده به چیزهایی نیاز داریم؟
  - روش استنتاج. [در صفحه‌ی قبل]
    - با مجموعه‌ای از احتمالات شروع کن: یعنی احتمال $P(Y)$ و جداول $P(F_i|Y)$.
    - از روش استنتاج استاندارد برای محاسبه استفاده $P(Y|F_1, \dots, F_n)$ کن.

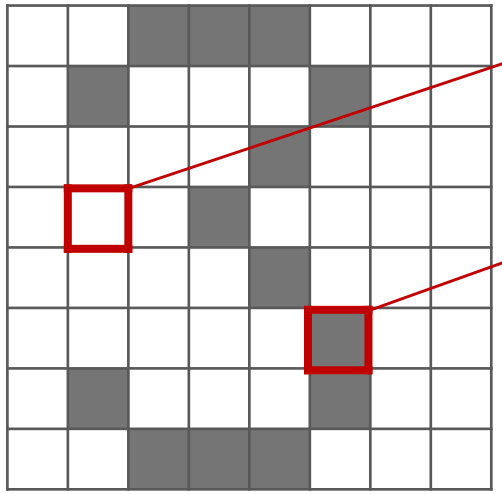


  - تخمین جداول احتمال شرطی محلی.
    - احتمال پیشین به ازای هر کلاس، $P(Y)$
    - احتمال شرطی به ازای هر ویژگی، $P(F_i|Y)$
    - احتمالات فوق در مجموع پارامترهای مدل نامیده می‌شوند و به وسیله‌ی θ نشان داده می‌شوند.
    - احتمالات فوق از روی یک مجموعه‌ی آموزشی محاسبه می‌شوند.

مثال: احتمال‌های شرطی

- محاسبه‌ی احتمال‌های شرطی.

$P(Y)$	
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$P(F_{3,1} = 1\|Y)$	
1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(F_{5,5} = 1\|Y)$	
1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

کاربرد بیز ساده در داده‌های متنی

- بیز ساده به صورت یک کیسه از کلمات.
  - ویژگی‌ها: W_i کلمه‌ی واقع در مکان i است.
  - مانند قبل: پیش‌بینی متغیر کلاس با توجه به ویژگی‌ها (هرزنامه یا خیر).
  - مانند قبل: با دانستن برچسب کلاس متغیرها به صورت شرطی مستقل هستند.
  - یک فرض جدید: هر کلمه‌ی W_i به صورت یکسان توزیع شده است.

- مدل تولید کننده. ترتیب کلمات اهمیت ندارد!

$$P(Y, W_1, \dots, W_n) = P(Y) \prod_i P(W_i|Y)$$

کلمه در مکان i ام، نه i امین کلمه در دیکشنری

کاربرد بیز ساده در داده‌های متنی

- کیسه کلمات.
  - معمولاً هر متغیر دارای یک احتمال شرطی $P(F|Y)$ مخصوص به خود است.
  - اما در مدل کیسه کلمات:
    - هر مکان به صورت یکسان توزیع شده است.
    - یعنی همه مکان‌ها از یک توزیع شرطی به صورت $P(W|Y)$ پیروی می‌کنند.
  - «کیسه کلمات» به این معنی است که مدل به ترتیب کلمات حساس نیست!

مثال: تشخیص هرزنامه

- مدل.

$$P(Y, W_1, \dots, W_n) = P(Y) \prod_i P(W_i|Y)$$

- پارامترها.

$P(Y)$

ham:	0.66
spam:	0.33

$P(W|spam)$

the:	0.0156
to:	0.0153
and:	0.0115
of:	0.0095
you:	0.0093
a:	0.0086
with:	0.0080
from:	0.0075
…	

$P(W|ham)$

the:	0.0210
to:	0.0133
of:	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and:	0.0105
a:	0.0100
…	

مثال

Word	P(w|spam)	P(w|ham)	Tot spam	Tot ham

$$P(spam|w) = \frac{e^{-76}}{e^{-76} + e^{-80.5}}$$

P(spam | w) = 98.9

آموزش و آزمایش

مفاهیم مهم

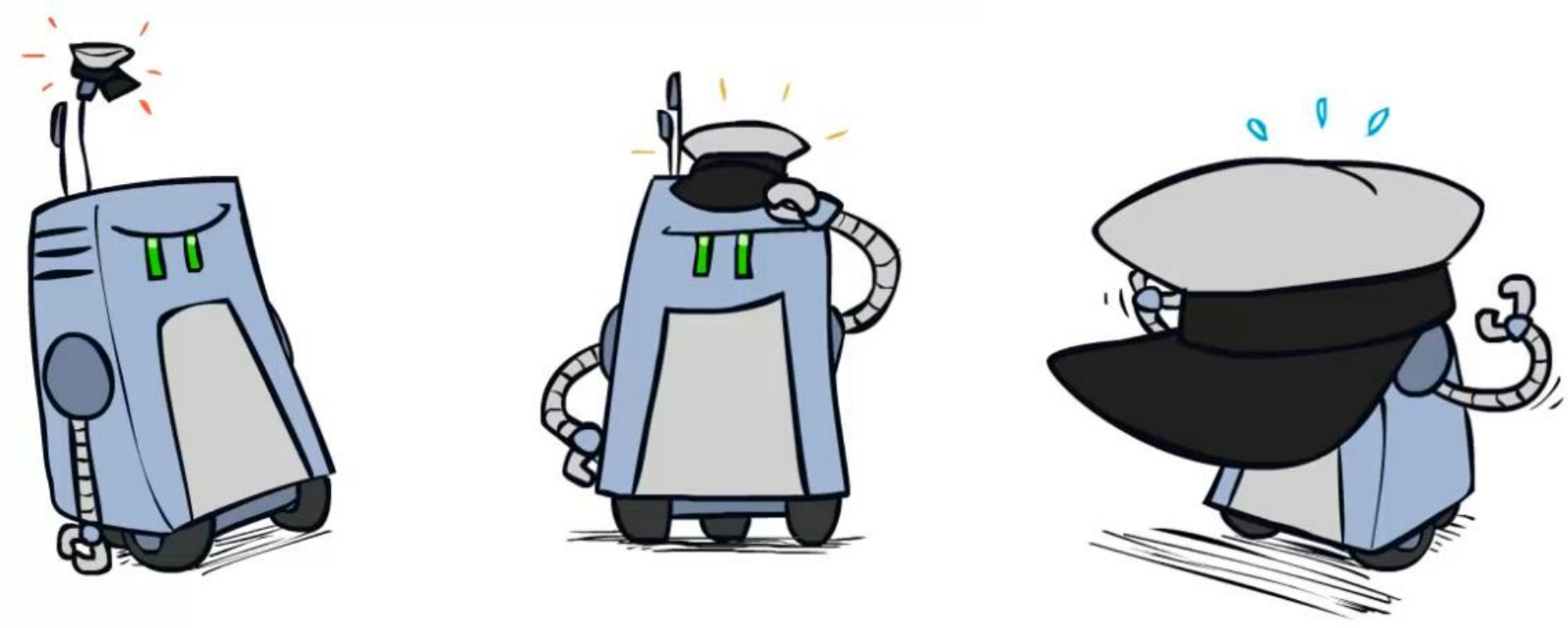
- داده‌ها. نمونه‌های برچسب‌گذاری شده.
  - مجموعه‌ی آموزشی
  - مجموعه‌ی اعتبارسنجی
  - مجموعه‌ی آزمایشی
- ویژگی‌ها. زوج‌های صفت-مقدار که مشخص کننده‌ی ورودی x هستند.
- چرخه‌ی یادگیری.
  - یادگیری پارامترها با استفاده از مجموعه‌ی آموزشی.
  - تنظیم ابرپارامترها با استفاده از مجموعه‌ی اعتبارسنجی.
  - محاسبه‌ی دقت بر روی مجموعه‌ی آزمایشی.
- ارزیابی. دقت کلاس‌بندی!
  - درصدی از نمونه‌ها که به درستی پیش‌بینی شده‌اند.

مجموعه‌ی آموزشی

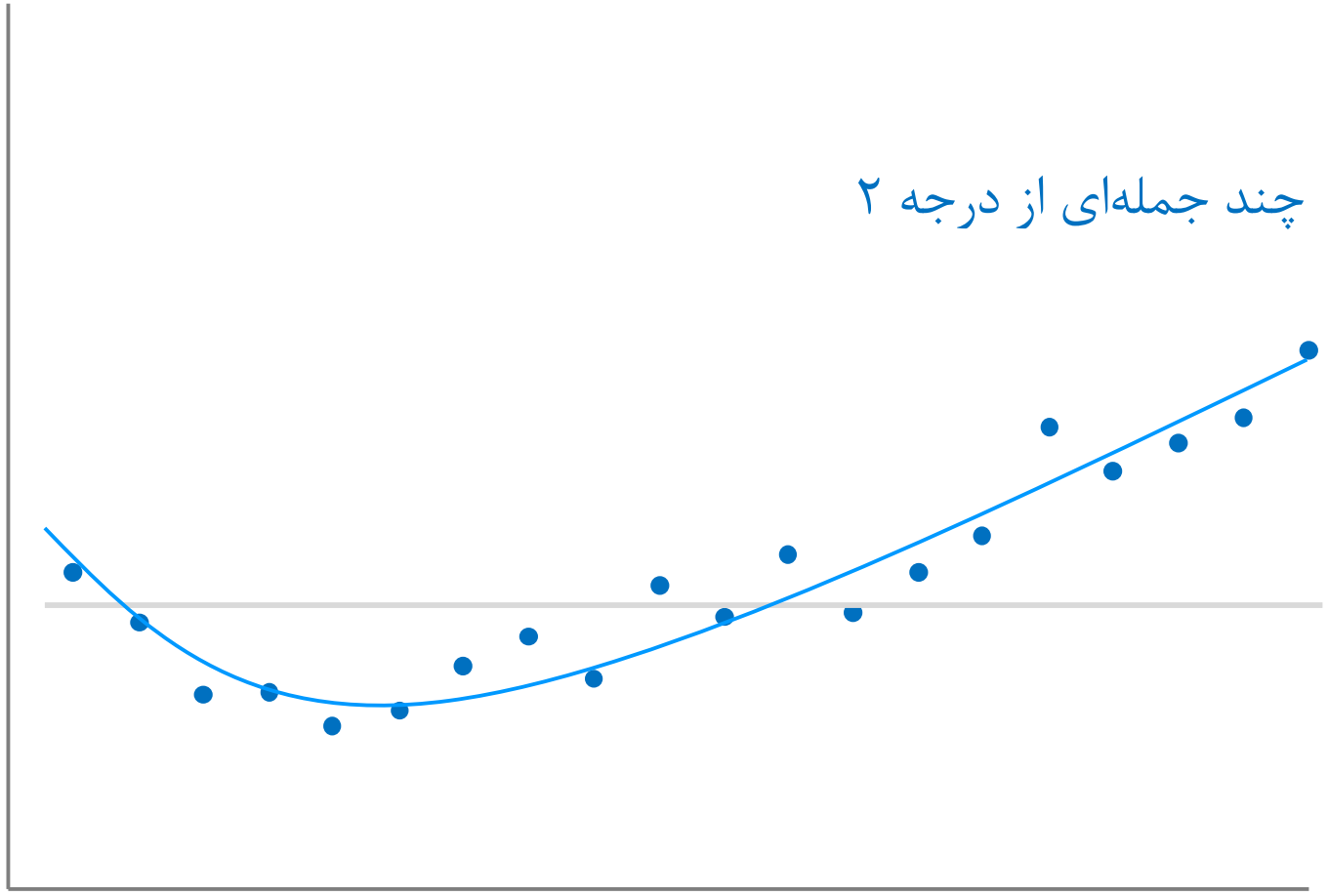
مجموعه‌ی اعتبارسنجی

مجموعه‌ی آزمایشی

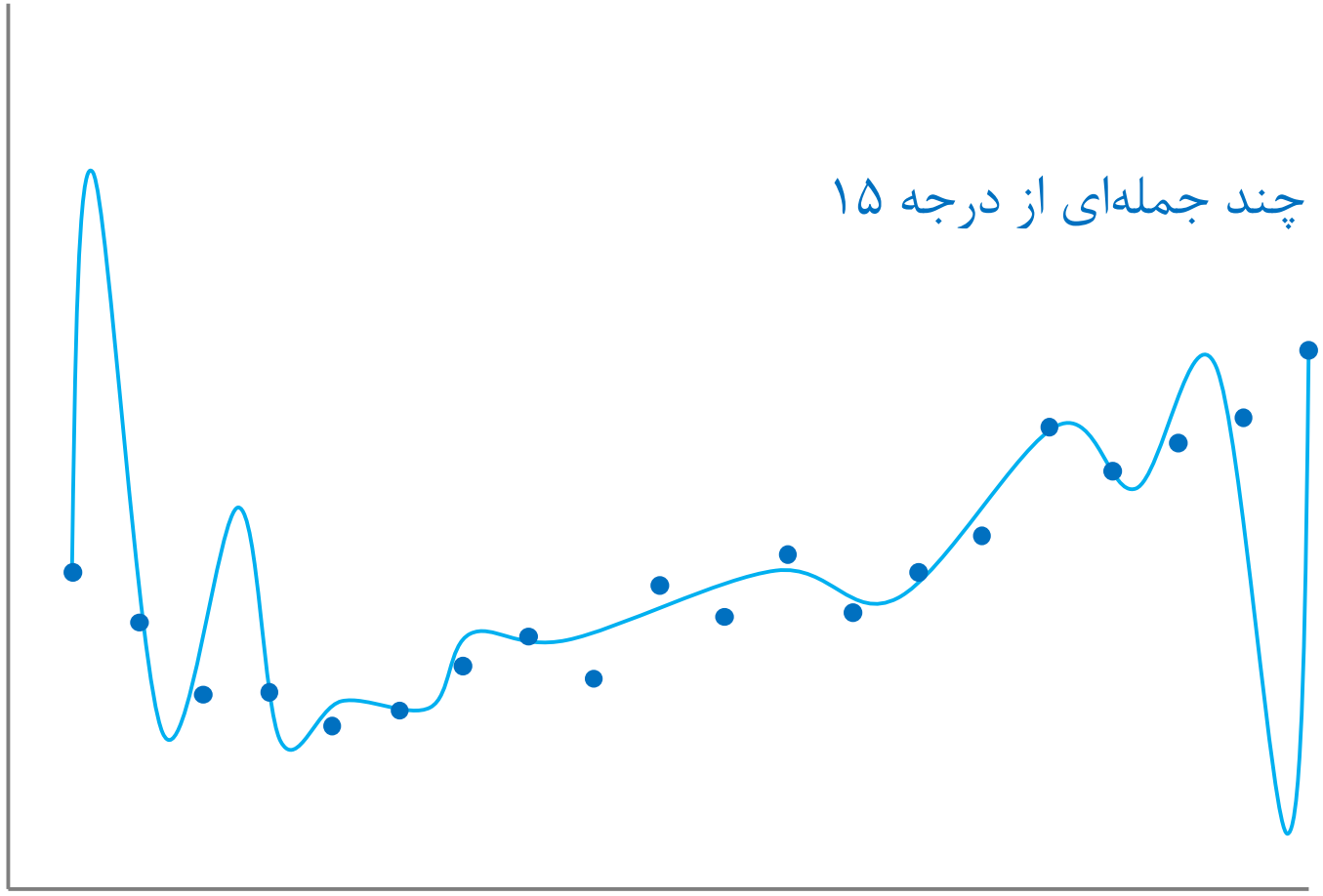
قابلیت تعمیم و بیش‌برازندگی

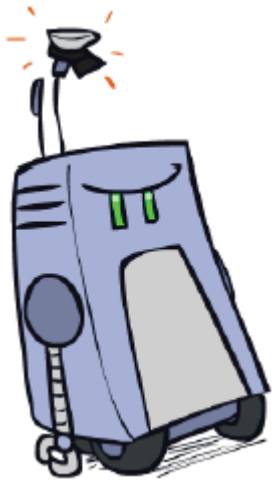


قابلیت تعمیم و بیش‌برازندگی



قابلیت تعمیم و بیش‌برازندگی





مثال: بیش‌برازش

$P(features, C = 2)$

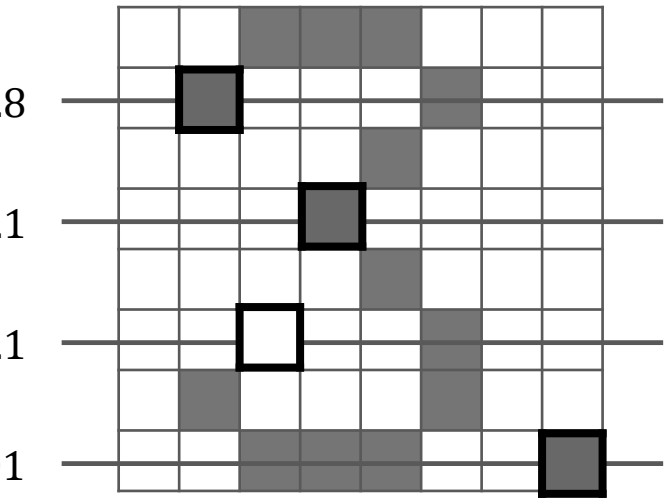
$P(C = 2) = 0.1$

$P(on|C = 2) = 0.8$

$P(on|C = 2) = 0.1$

$P(off|C = 2) = 0.1$

$P(on|C = 2) = 0.01$



$P(features, C = 3)$

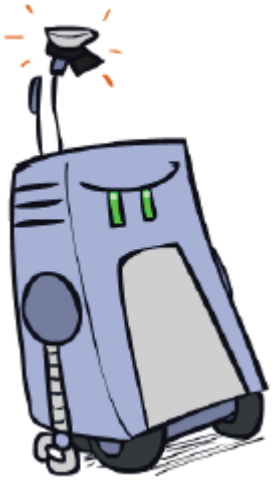
$P(C = 3) = 0.1$

$P(on|C = 3) = 0.8$

$P(on|C = 3) = 0.9$

$P(off|C = 3) = 0.7$

$P(on|C = 3) = 0.0$



کلاس ۲ برنده می‌شود!

مثال: بیش‌برازش

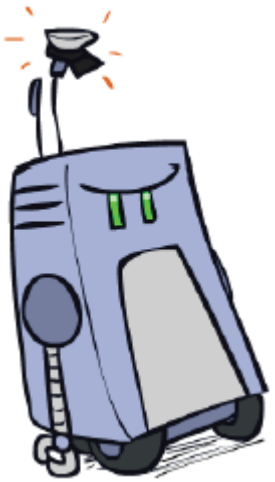
- احتمال‌های پسین به وسیله احتمال‌های نسبی تعیین می‌شوند.

$$\frac{P(W|spam)}{P(W|ham)}$$

screens:	inf
minute:	inf
guaranty:	inf
$205.00:	inf
delivery:	inf
signature:	inf
...	

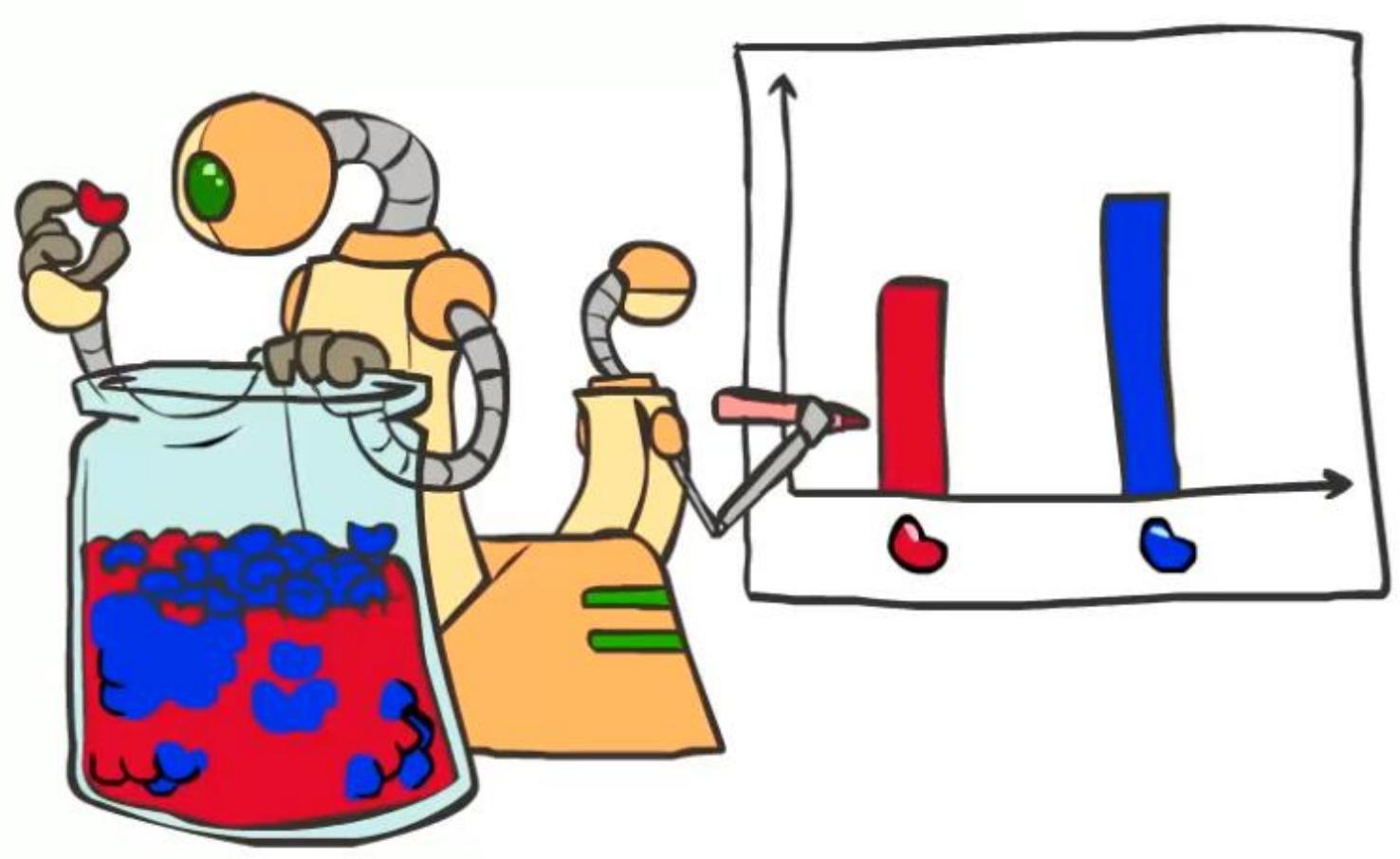
$$\frac{P(W|ham)}{P(W|spam)}$$

south-west:	inf
nation:	inf
morally:	inf
nicely:	inf
extent:	inf
seriously:	inf
...	

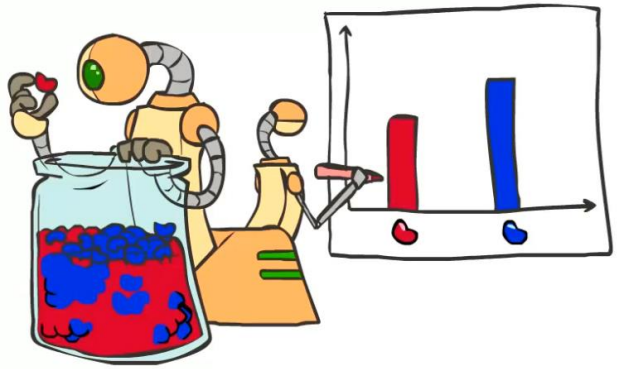


چه چیزی در اینجا اشتباه است؟

تخمین پارامتر



تخمین پارامتر



- تخمین پارامتر. تخمین توزیع یک متغیر تصادفی.
  - فراخوانی: پرسیدن از یک انسان (چرا سخت است؟)
  - تجربی: با استفاده از داده‌های آموزشی (یادگیری!)

- تخمین بیشترین درست‌نمایی.

$$P_{ML}(x) = \frac{count(x)}{\#\ of\ samples}$$

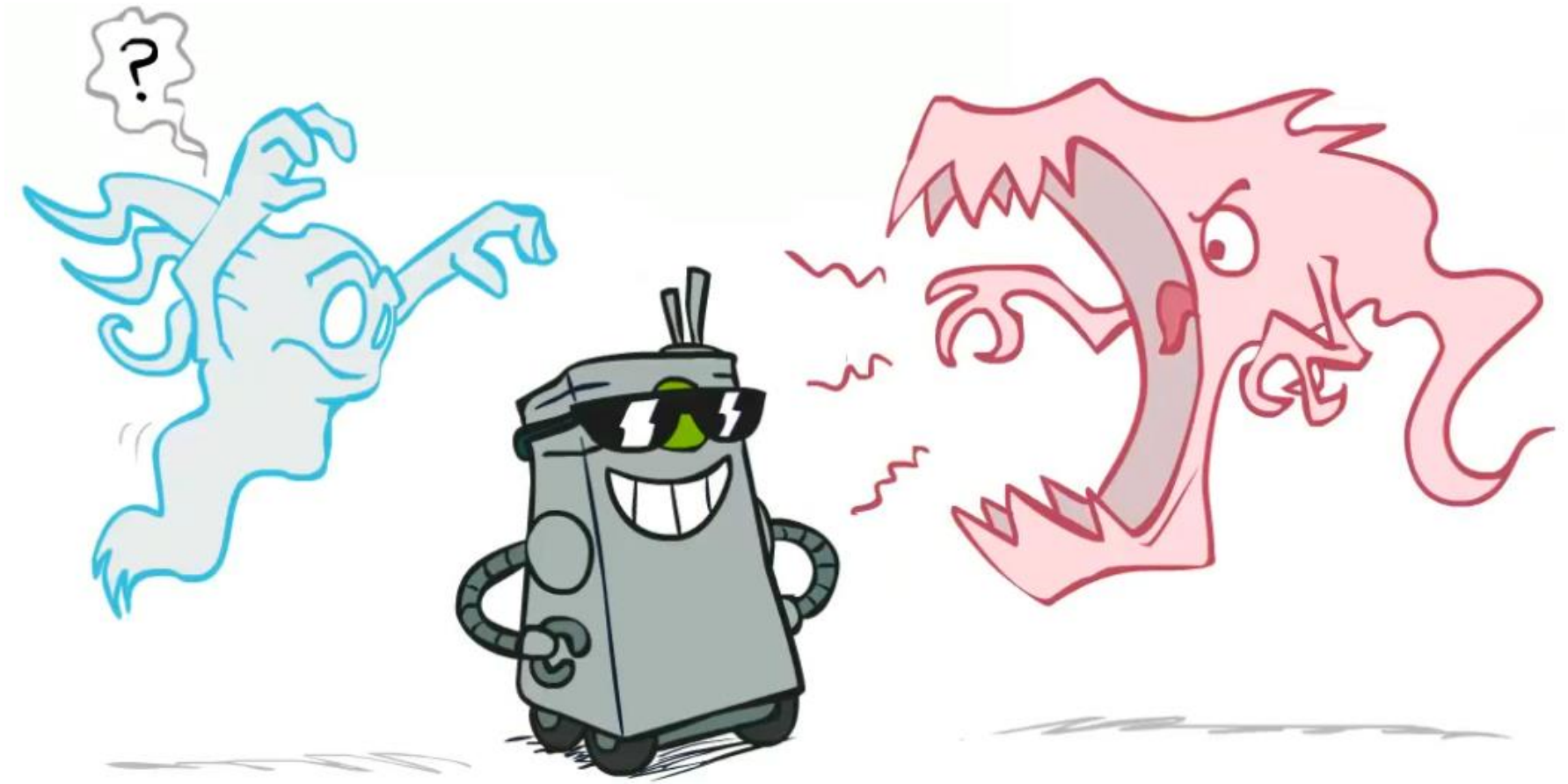
r r b $\qquad P_{ML}(r) = 2/3$

- این تخمین درست‌نمایی داده‌ها را به حداکثر می‌رساند.

$$L(x, \theta) = \prod_i P_\theta(x_i)$$

$$\max_r [r^2(1-r)] \Rightarrow r = 2/3$$

همواره‌سازی



بیشینه‌سازی درست‌نمایی؟

- فراوانی‌های نسبی همان تخمین بیشترین درست‌نمایی هستند.

$$\theta_{ML} = \arg\max_{\theta} P(X|\theta)$$
$$= \arg\max_{\theta} \Pi_i P_\theta(X_i)$$

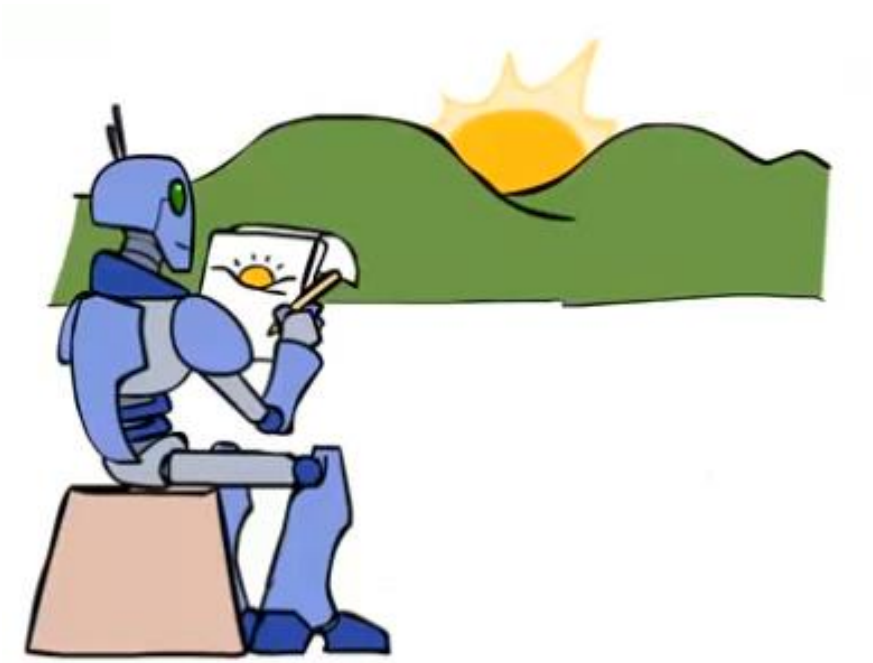
$$\Rightarrow \quad P_{ML}(x) = \frac{count(x)}{\#\ of\ samples}$$

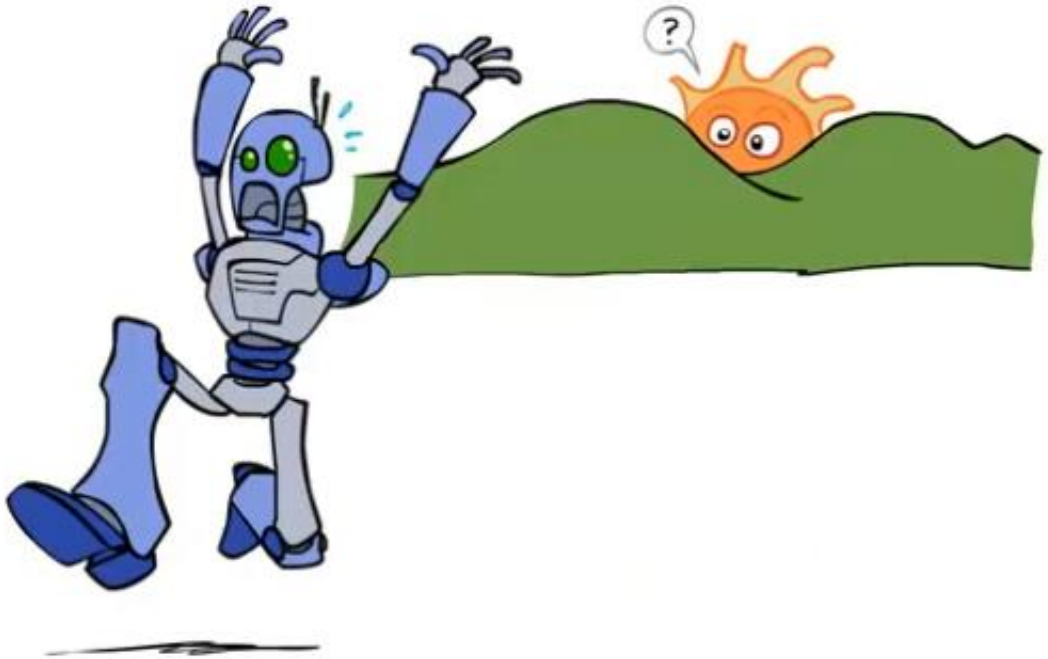
- یک گزینه‌ی دیگر. در نظر گرفتن محتمل‌ترین مقدار برای پارامترها با توجه به داده‌ها.

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|X)$$
$$= \arg\max_{\theta} P(X|\theta)P(\theta)/P(X)$$
$$= \arg\max_{\theta} P(X|\theta)P(\theta)$$

امکان استفاده از دانش در مورد احتمال‌های پیشین!!!

رویدادهای پیش‌بینی نشده



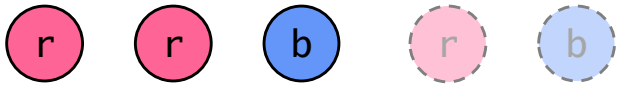


هموارسازی لاپلاس

- تخمین لاپلاس.
  - وانمود کن که هر پیامد را یک بار بیشتر از مقدار واقعی دیده‌ای.

$$P_{LAP}(x) = \frac{count(x)+1}{\Sigma_x[count(x)+1]}$$

$$= \frac{count(x)+1}{N+|X|}$$



$$P_{ML}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

نزدیک‌تر به توزیع یکنواخت ⟵ $P_{LAP}(X) = \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle$

هموارسازی لاپلاس

- تخمین لاپلاس (تعمیم یافته).
  - وانمود کن که هر پیامد را k بار بیشتر از مقدار واقعی دیده‌ای.

$$P_{LAP,k}(x) = \frac{count(x) + k}{N + k|X|}$$

شرت احتمال پیشین

$$P_{LAP,0}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

$$P_{LAP,1}(X) = \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle$$

$$P_{LAP,100}(X) = \left\langle \frac{102}{203}, \frac{101}{203} \right\rangle$$

هموارسازی

- برای مسائل کلاس‌بندی در دنیای واقعی، هموارسازی یک امر حیاتی است.

$$\frac{P(W|spam)}{P(W|ham)}$$

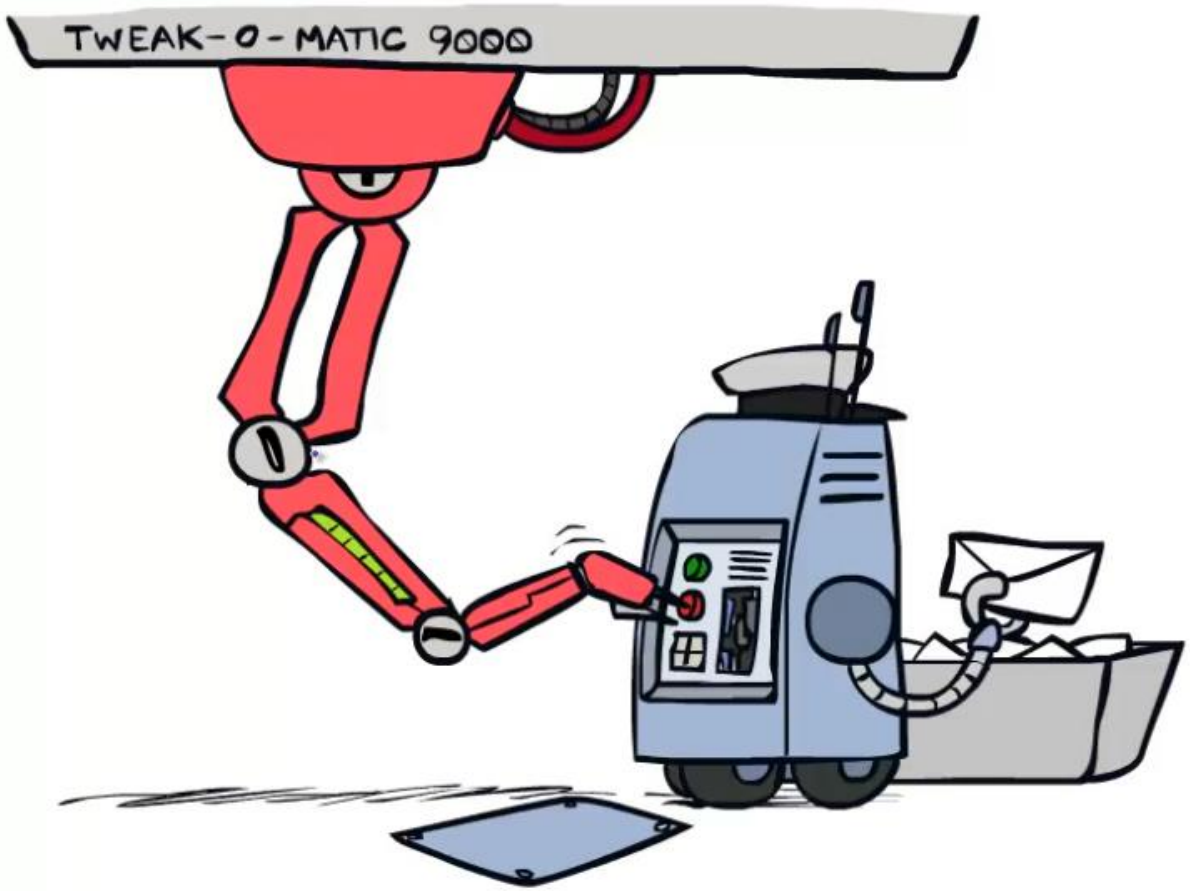
verdana:	28.8
Credit:	28.4
ORDER:	27.2
<FONT>:	26.9
money:	26.5
...	

$$\frac{P(W|ham)}{P(W|spam)}$$

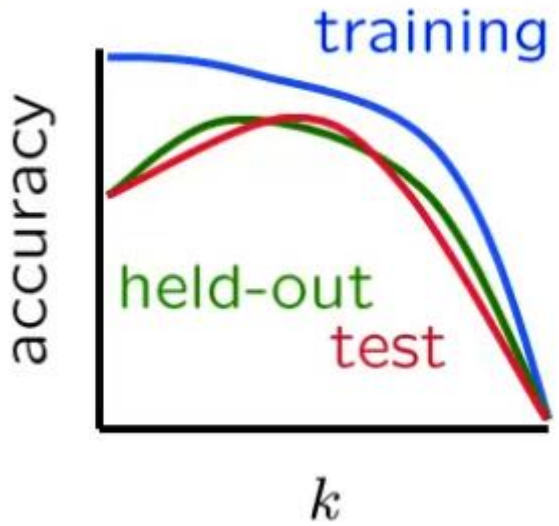
Helvetica:	11.4
seems:	10.8
group:	10.2
ago:	8.4
areas:	8.3
...	

آیا نسبت به قبل منطقی‌تر است؟

تنظیم ابرپارامترها



تنظیم بر روی داده‌های اعتبارسنجی



- مقادیری که باید یاد گرفته شوند.
  - پارامترها: مقادیر $P(Y)$ و $P(X|Y)$.
  - ابرپارامترها: مقدار یا نوع هموارسازی مانند k و α.
- یادگیری پارامترها و ابرپارامترها.
  - یادگیری پارامترها از روی مجموعه‌ی آموزشی.
  - تنظیم مقدار ابرپارامترها بر روی داده‌های متفاوت.
    - چرا؟
  - به ازای هر مقدار ابرپارامترها، آموزش و سنجش بر روی مجموعه‌ی اعتبارسنجی.
  - انتخاب بهترین مقدار برای ابرپارامترها و انجام سنجش نهایی بر روی داده‌های آزمایشی.