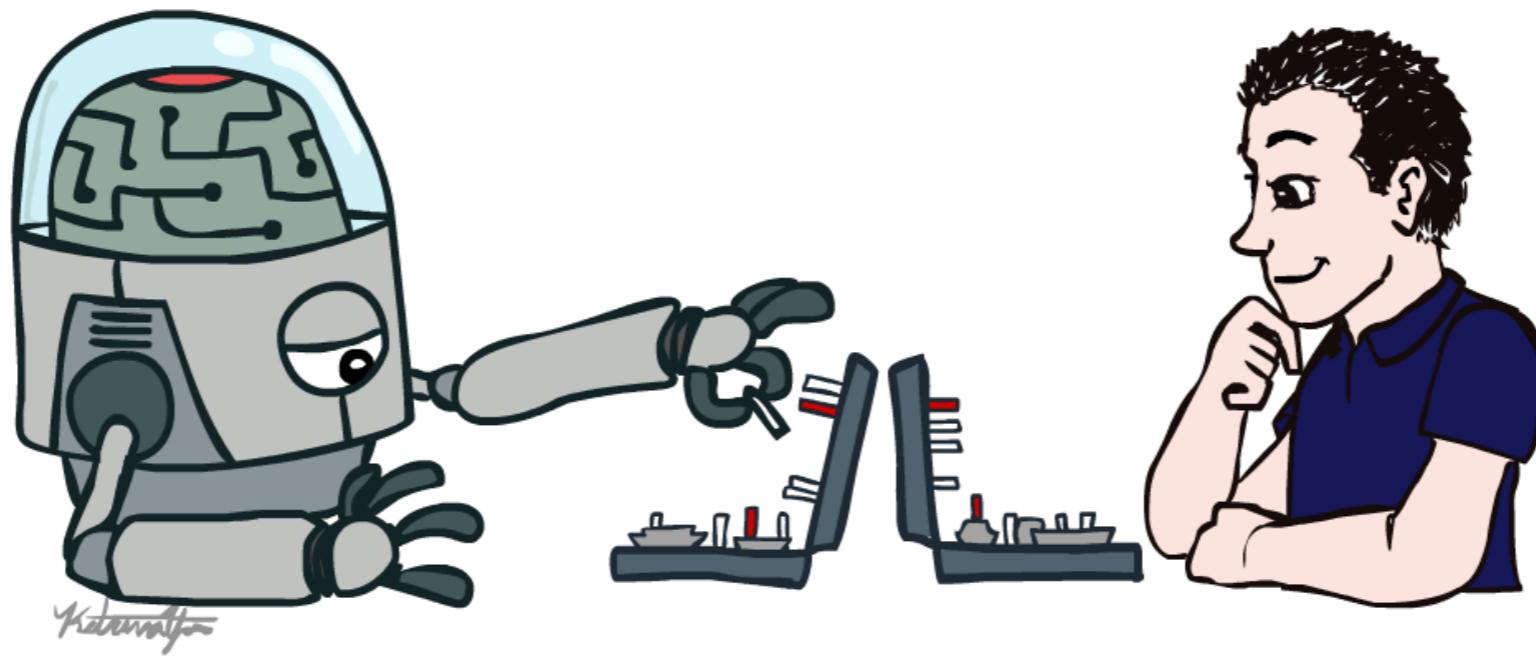
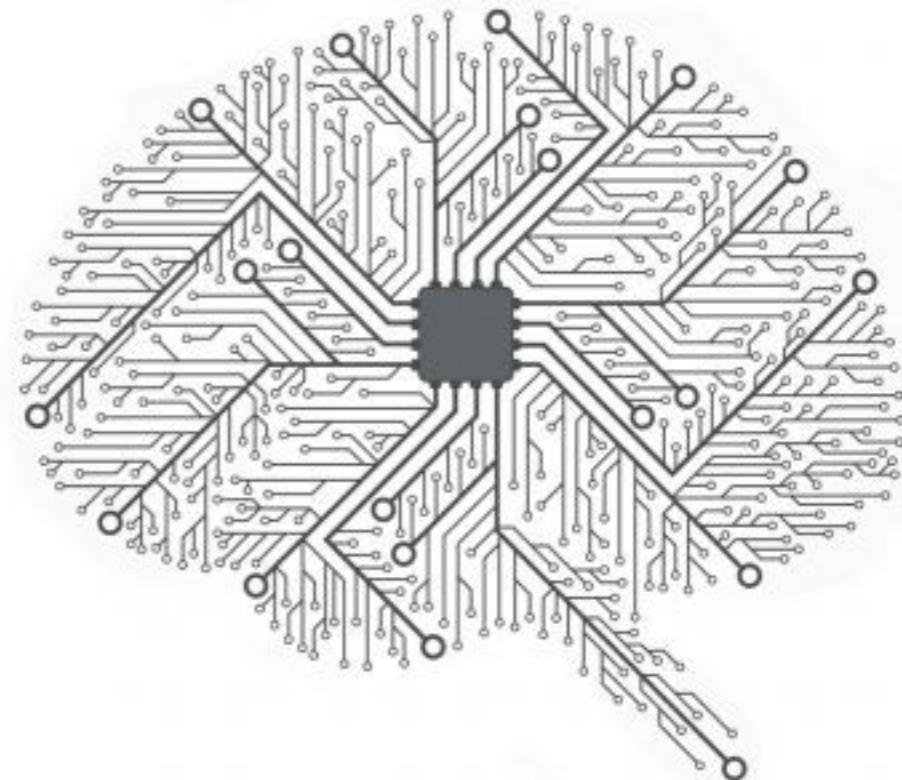


人工智能



第八章·机器学习与朴素贝叶斯

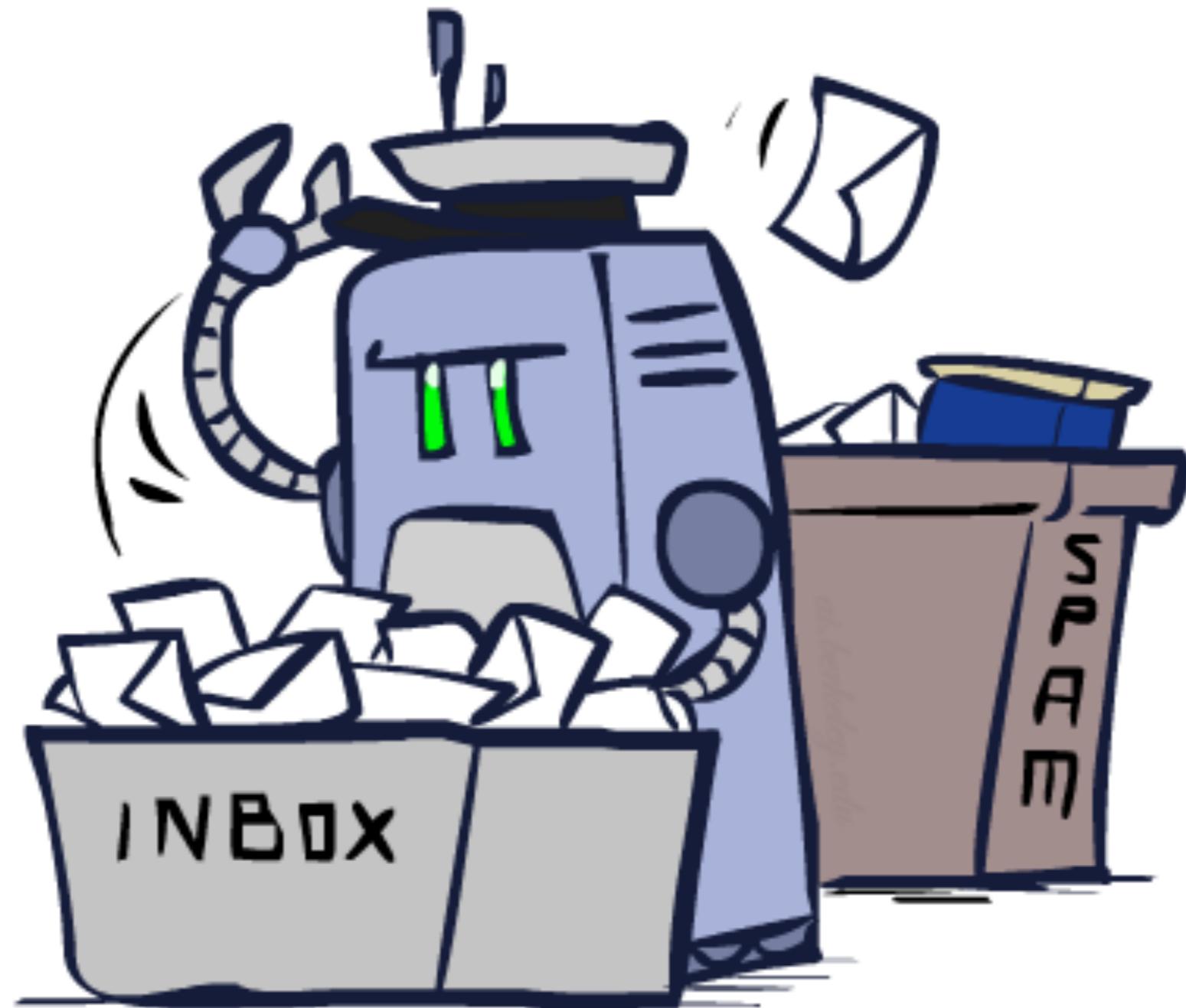
- 机器学习
- 基于模型的分类任务
- 朴素贝叶斯



机器学习

- 到目前为止：如何使用模型做出最佳决策
- 机器学习：如何从数据/经验中获取模型
 - 学习参数
 - 学习结构
 - 学习隐藏概念
- 这一章：基于模型的分类——朴素贝叶斯

分类



示例：垃圾邮件过滤器

- 输入：email
- 输出：正常邮件(ham)/垃圾邮件(spam)
- 问题描述：
 - 获取大量示例电子邮件，每个都标记为“spam”或“ham”
 - 注意：必须有人预先手动标记所有这些数据！
 - 想学习如何对新的电子邮件进行标签预测
- 特征：用于决策的一些属性
 - Words: FREE!
 - Text Patterns: \$dd, CAPS
 - Non-text: SenderInContacts, WidelyBroadcast
 - ...



Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential 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, I know 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.

示例：数字识别

- 输入: 图像 / 像素矩阵

- 输出: 数字0-9

- 问题描述:

- 获取大量的示例图像集合，每个图像都标有数字
- 注释: 必须有人手动标记所有这些数据!
- 想学习如何对新的数字图像进行标签预测

- 特征: 用于决策的一些属性

- 像素: $(6,8)=\text{ON}$
- 形状模式: NumComponents, AspectRatio, NumLoops
-
- 越来越多地引入特征，而不是筛选淘汰特征

 0

 1

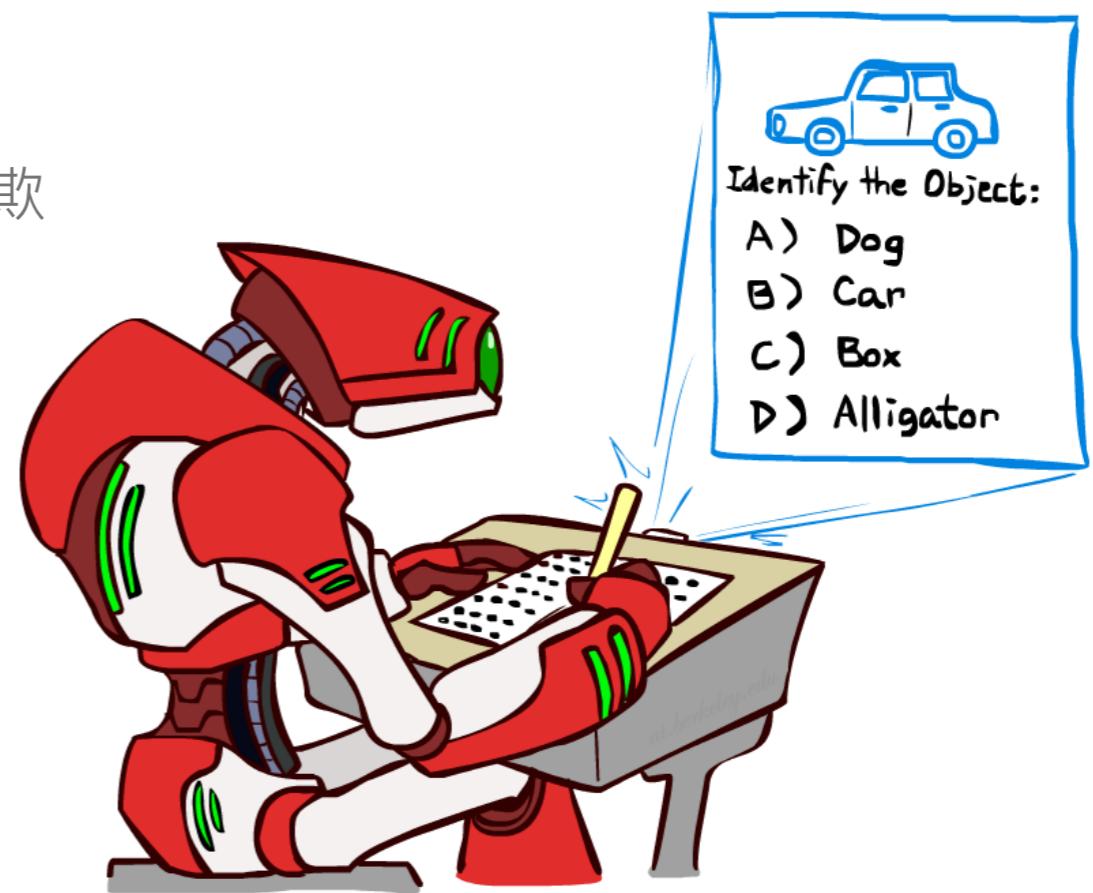
 2

 1

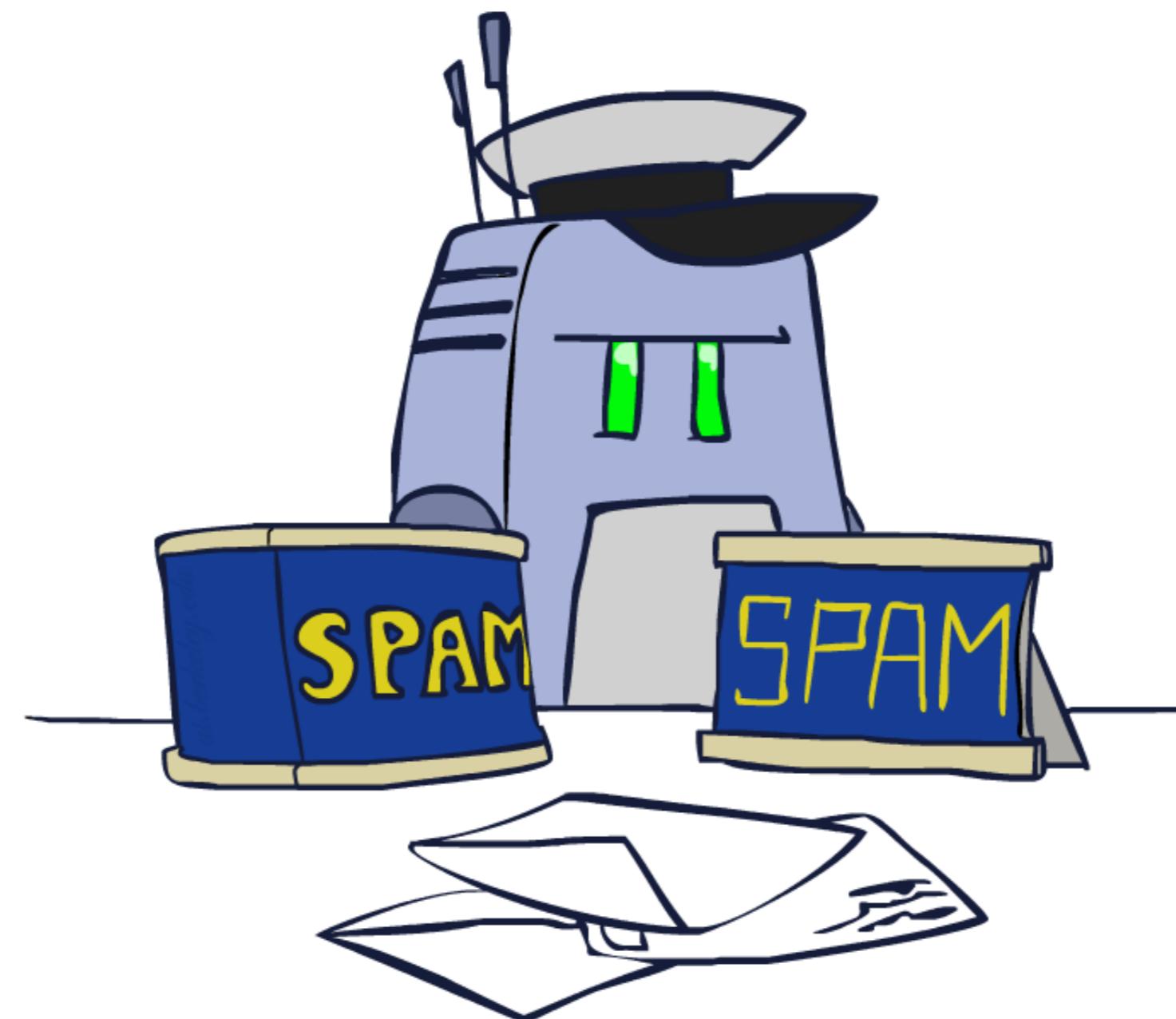
 ??

其他分类任务

- 分类任务: 给定输入 x , 预测标签 (类别) y
- 例如
 - 医疗诊断 (输入: 症状, 类别: 疾病)
 - 欺诈检测 (输入: 帐户活动, 类别: 欺诈/无欺诈)
 - 自动论文评分 (输入: 文档, 类别: 分数)
 - 客户服务
 - 电子邮件路由
 - 语言识别
 - ...更多
- 分类是一项有着重要商业应用的技术!

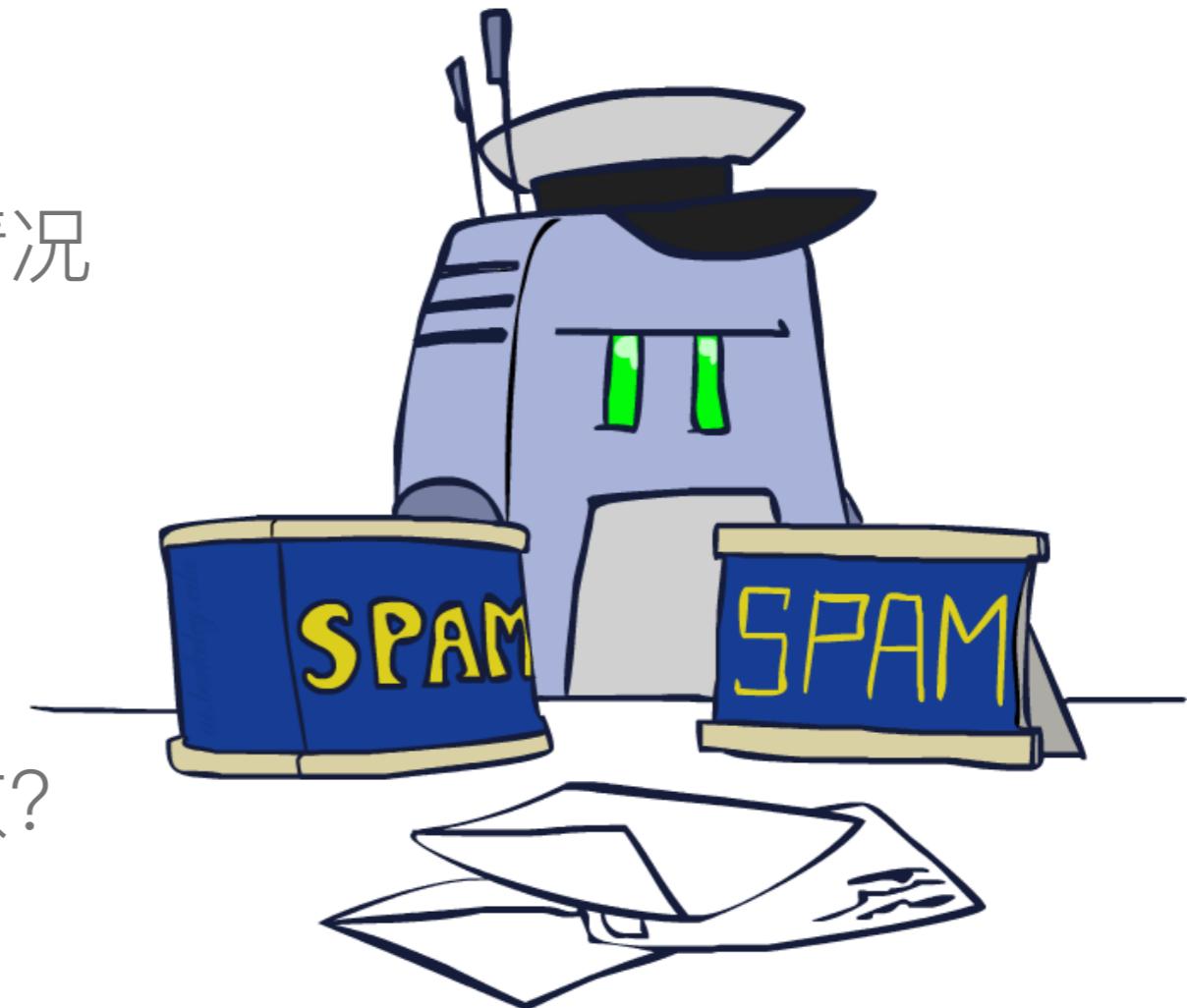


基于模型的分类任务



基于模型的分类任务

- 基于模型的方法
 - 建立一个输出标签和输入特征都是任意变量的模型
 - 实例化任何观察到的特征
 - 根据特征查询标签的分布情况
- 问题
 - 模型应该有什么样的结构?
 - 我们应该如何学习它的参数?

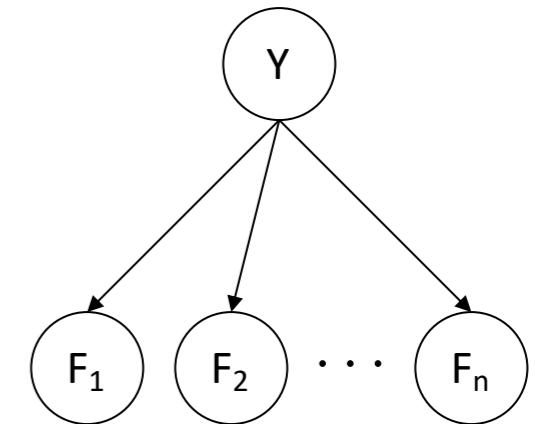


朴素贝叶斯 Naïve Bayes

- 朴素贝叶斯：假设所有特征都独立作用于标签

- 简单数字识别版本

- 每个网格位置 $<i,j>$ 作为一个特征（变量） $F_{i,j}$



- 特征取值为 on / off, 基于该位置像素的灰度是否大于 0.5

- 每个输入映射到一个特征向量, 例如

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

- 有很多特征, 每个特征都是二值的

- 朴素贝叶斯模型: $P(Y|F_{0,0} \dots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$

- 需要学习的是什么?

朴素贝叶斯

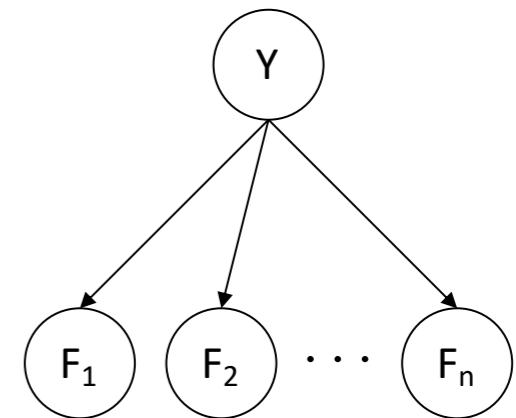
- 一般的朴素贝叶斯模型：

$|Y|$ parameters

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

$|Y| \times |F|^n$ values

$n \times |F| \times |Y|$
parameters



- 我们只需要指定每个特征如何依赖于类别
- 参数总数以 n 为单位呈线性
- 模型非常简单，但通常都是有效的

朴素贝叶斯的模型推断

- 目标：计算标签变量Y的后验分布
 - 第1步：为每个标签获取标签和证据(特征)的联合概率

$$P(Y, f_1 \dots f_n) = \begin{bmatrix} P(y_1, f_1 \dots f_n) \\ P(y_2, f_1 \dots f_n) \\ \vdots \\ P(y_k, f_1 \dots f_n) \end{bmatrix} \xrightarrow{\text{ }} \frac{\begin{bmatrix} P(y_1) \prod_i P(f_i|y_1) \\ P(y_2) \prod_i P(f_i|y_2) \\ \vdots \\ P(y_k) \prod_i P(f_i|y_k) \end{bmatrix}}{P(f_1 \dots f_n)}$$

- 第2步：求和得到证据的概率
- 第3步：将步骤1除以步骤2进行规范化

$$P(Y|f_1 \dots f_n)$$

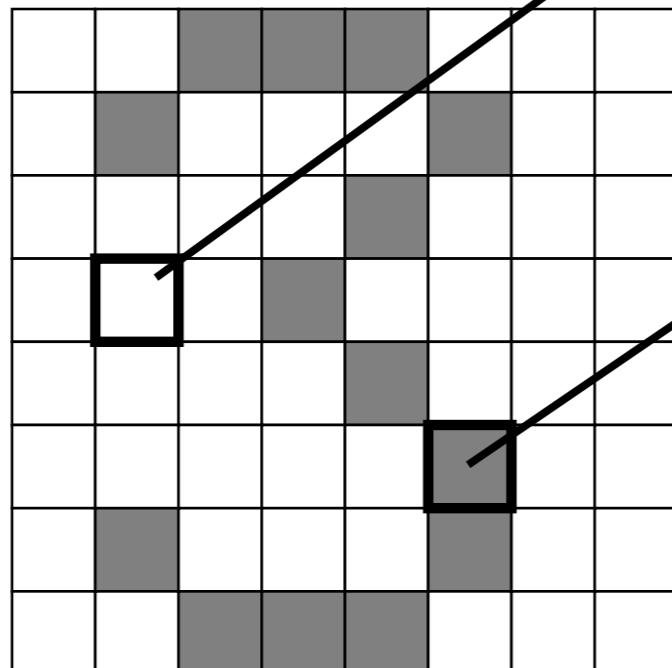
我们需要什么才能使用朴素贝叶斯?

- 推断方法
 - 需要一系列的概率: $P(Y)$ 和 $P(F_i|Y)$
 - 从而计算得到 $P(Y|F_1 \dots F_n)$
- 局部条件概率表的估计
 - $P(Y)$, 标签的先验概率
 - $P(F_i|Y)$ 对于每个特征的条件概率
 - 这些概率统称为模型的参数, 用 θ 表示, 通常来自训练数据的统计

示例：条件概率

$$P(Y)$$

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



$$P(F_{3,1} = \text{on}|Y) \quad P(F_{5,5} = \text{on}|Y)$$

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

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

文本的朴素贝叶斯

- 基于词袋(Bag-of-words)的朴素贝叶斯:
 - 特征: W_i 是在位置 i 的词
 - (和之前一样)基于特征变量预测标签 (spam vs. ham)
 - (和之前一样)假设特征是条件独立的
 - (和之前不一样)每个 W_i 分布相同
 - 模型: $P(Y, W_1 \dots W_n) = P(Y) \prod_i P(W_i|Y)$
 - 理论上, 应该分别计算每个位置上的分布 $P(F|Y)$
 - 但是在词袋模型中, 假设每个位置的分布是相同的, 于是所有的位置共享一个条件概率 $P(W|Y)$
 - 称为“词袋”, 因为模型对词序不敏感
- Word at position
 i , not i^{th} word in
the dictionary!*

示例：垃圾邮件过滤

- 模型： $P(Y, W_1 \dots W_n) = P(Y) \prod_i P(W_i|Y)$
- 参数（特征）：

$P(Y)$	$P(W \text{spam})$	$P(W \text{ham})$
ham : 0.66 spam: 0.33	the : 0.0156 to : 0.0153 and : 0.0115 of : 0.0095 you : 0.0093 a : 0.0086 with: 0.0080 from: 0.0075 ...	the : 0.0210 to : 0.0133 of : 0.0119 2002: 0.0110 with: 0.0108 from: 0.0107 and : 0.0105 a : 0.0100 ...

- 如何计算得到这些表格的？

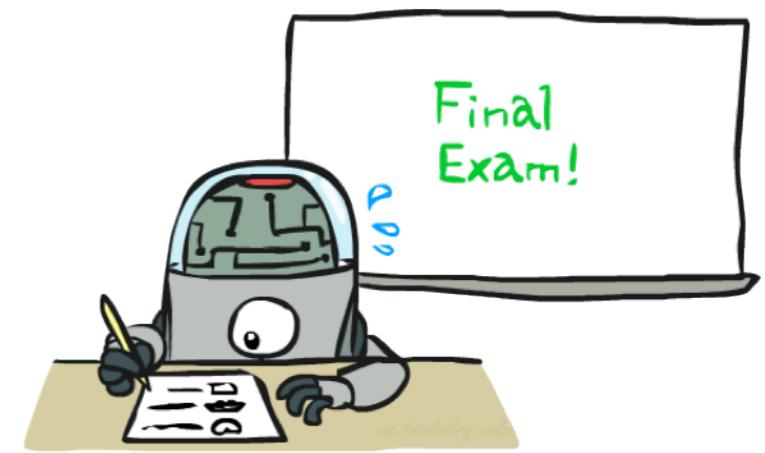
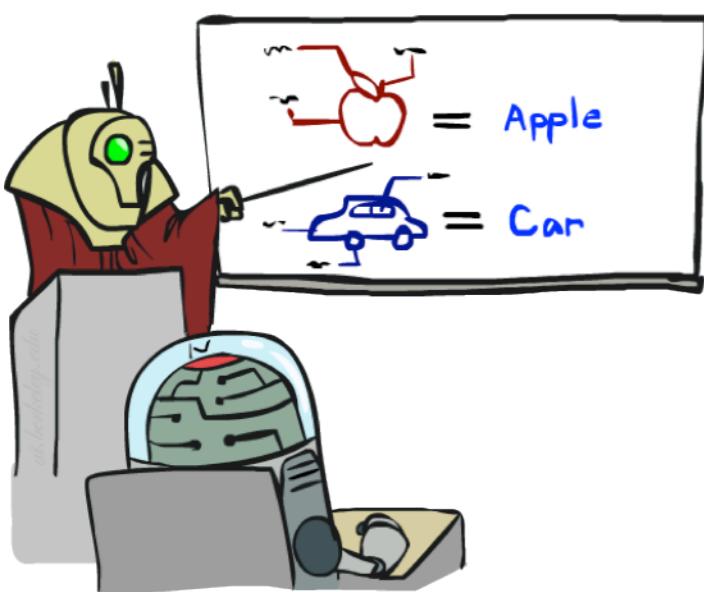
示例：垃圾邮件

Word	P(w spam)	P(w ham)	Tot Spam ^{log_e}	Tot Ham ^{log_e}
(prior)	0.33333	0.66666	-1.1	-0.4
Gary	0.00002	0.00021	-11.8	-8.9
would	0.00069	0.00084	-19.1	-16.0
you	0.00881	0.00304	-23.8	-21.8
like	0.00086	0.00083	-30.9	-28.9
to	0.01517	0.01339	-35.1	-33.2
lose	0.00008	0.00002	-44.5	-44.0
weight	0.00016	0.00002	-53.3	-55.0
while	0.00027	0.00027	-61.5	-63.2
you	0.00881	0.00304	-66.2	-69.0
sleep	0.00006	0.00001	-76.0	-80.5

$$P(\text{spam}|w) = \text{Exp}[-76.0]/(\text{Exp}[-76.0]+\text{Exp}[-80.5])= 0.989$$

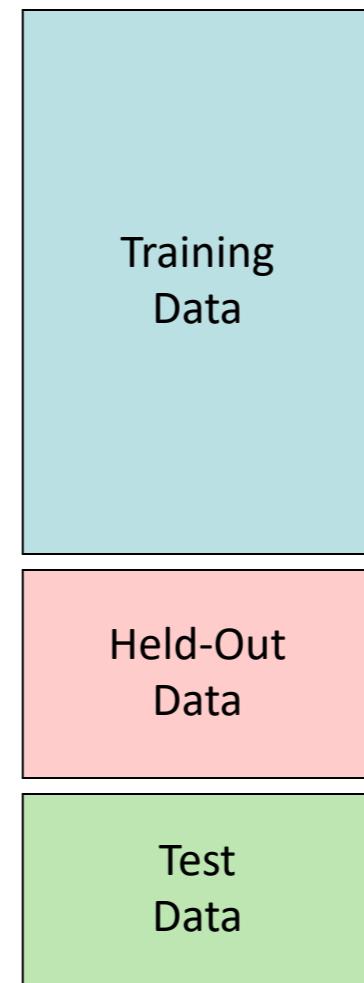
$$P(\text{ham}|w) = \text{Exp}[-80.5]/(\text{Exp}[-76.0]+\text{Exp}[-80.5])= 0.011$$

训练和测试



重要概念

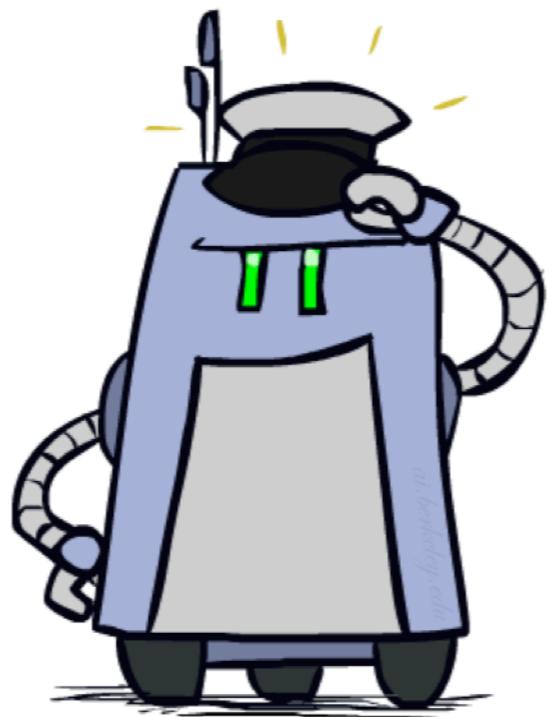
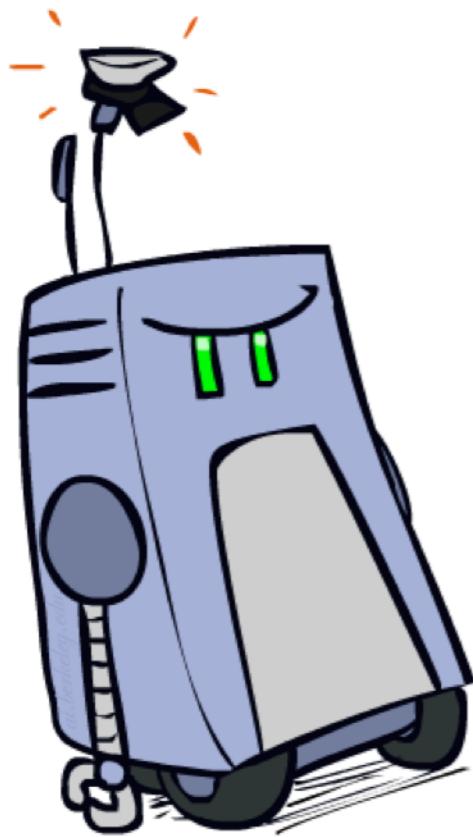
- 数据: 带标注的实例 (例如 电子邮件标注为 spam/ham)
 - 训练集(Train set)
 - 留出集(Held out set)
 - 测试集(Test set)
 - 实验过程
 - 在训练集上学习参数 (例如模型概率)
 - 在留出集上调整超参数
 - 在测试集上测试准确率
 - 非常重要: 永远不要“偷看”测试集!
 - 评估 (可能有许多指标, 例如Accuracy, Precision, Recall, F1)
 - Accuracy: 正确预测的实例比例
 - 过拟合与泛化
 - 过拟合: 训练数据拟合得很好, 但在测试集上表现糟糕
 - 欠拟合: 训练数据拟合得很不好



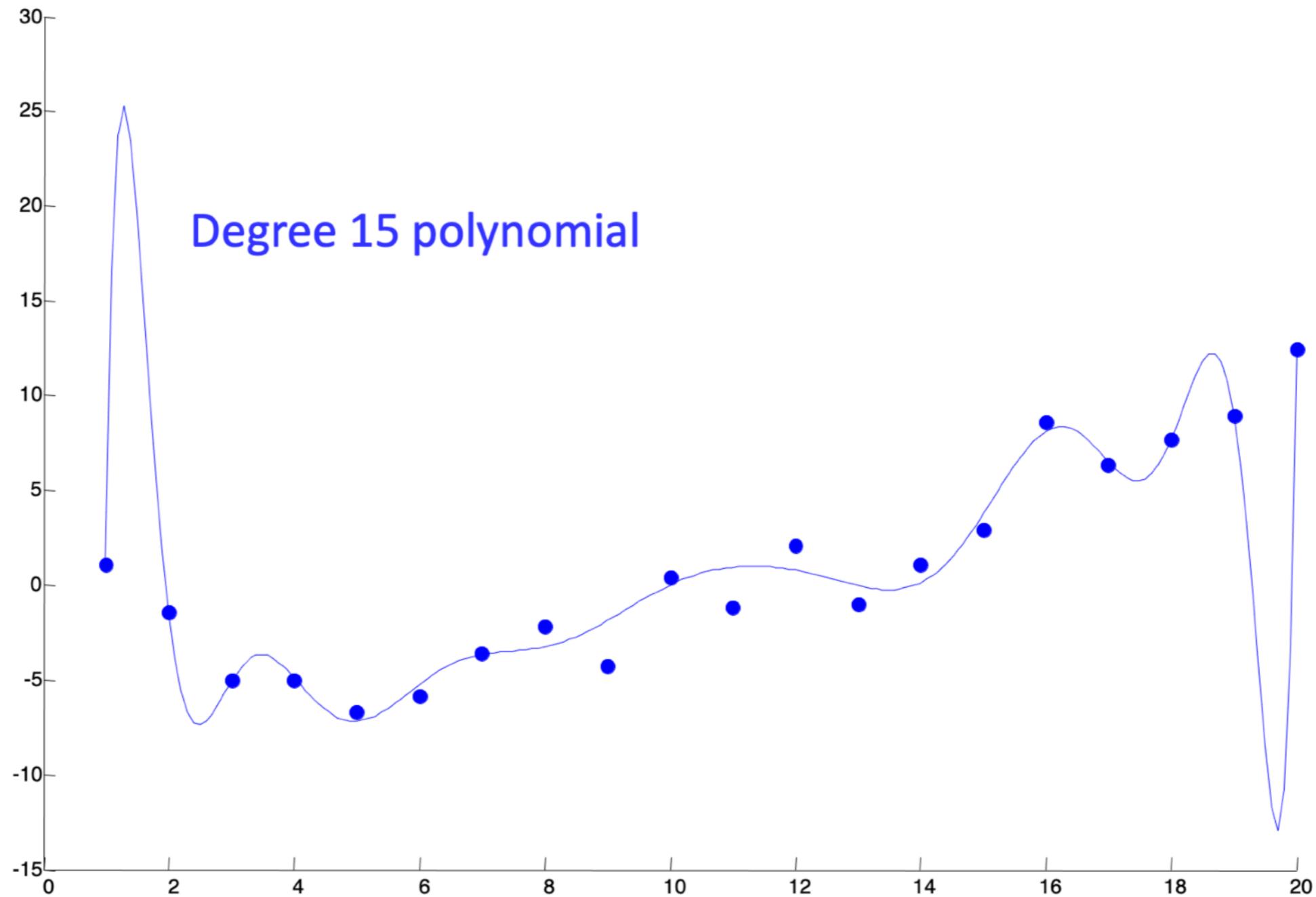
经验风险最小化

- 经验风险最小化
 - 我们希望模型（分类器）在真正的数据分布上做得最好
 - 由于不知道真正的分布，所以从我们的训练集中挑选最好的模型
 - 在训练集上寻找“最佳”模型是一个优化问题
- 主要问题：过度适应训练数据（Overfitting）
 - 训练数据越多越好（抽样方差越小，训练越像测试）
 - 限制假设的复杂性（正则化和/或小假设空间）会更好

过拟合与泛化



过拟合



示例：过拟合

$P(\text{features}, C = 2)$

$P(C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.8$

$P(\text{on}|C = 2) = 0.1$

$P(\text{off}|C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.01$

$P(\text{features}, C = 3)$

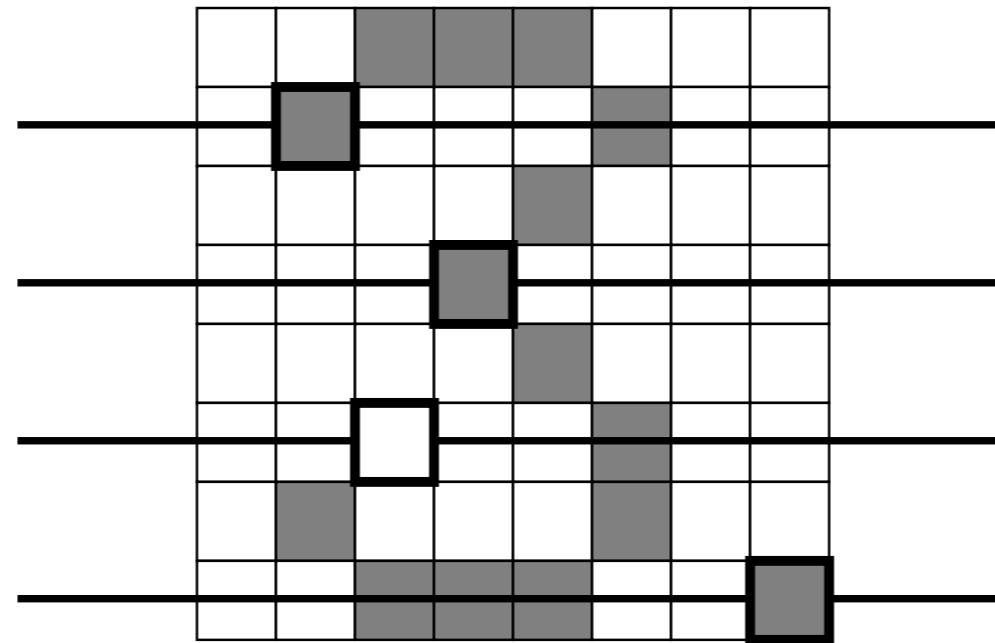
$P(C = 3) = 0.1$

$P(\text{on}|C = 3) = 0.8$

$P(\text{on}|C = 3) = 0.9$

$P(\text{off}|C = 3) = 0.7$

$P(\text{on}|C = 3) = 0.0$



$$P(\mathbf{Y}, \mathcal{F}_1 \dots \mathcal{F}_n) = P(\mathbf{Y}) \prod_i P(\mathcal{F}_i | \mathbf{Y})$$

2 wins!!

示例：过拟合

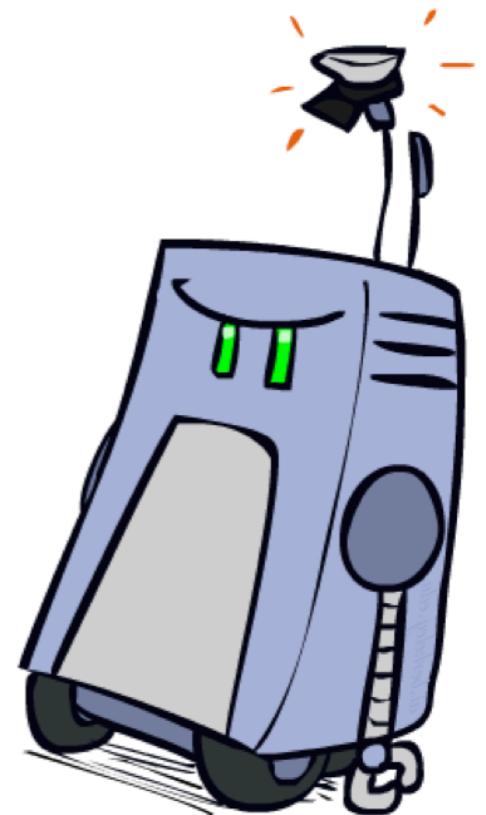
- 由相对概率决定后验概率

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

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

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

```
screens      : inf  
minute       : inf  
guaranteed   : inf  
$205.00      : inf  
delivery     : inf  
signature    : inf  
...  
...
```

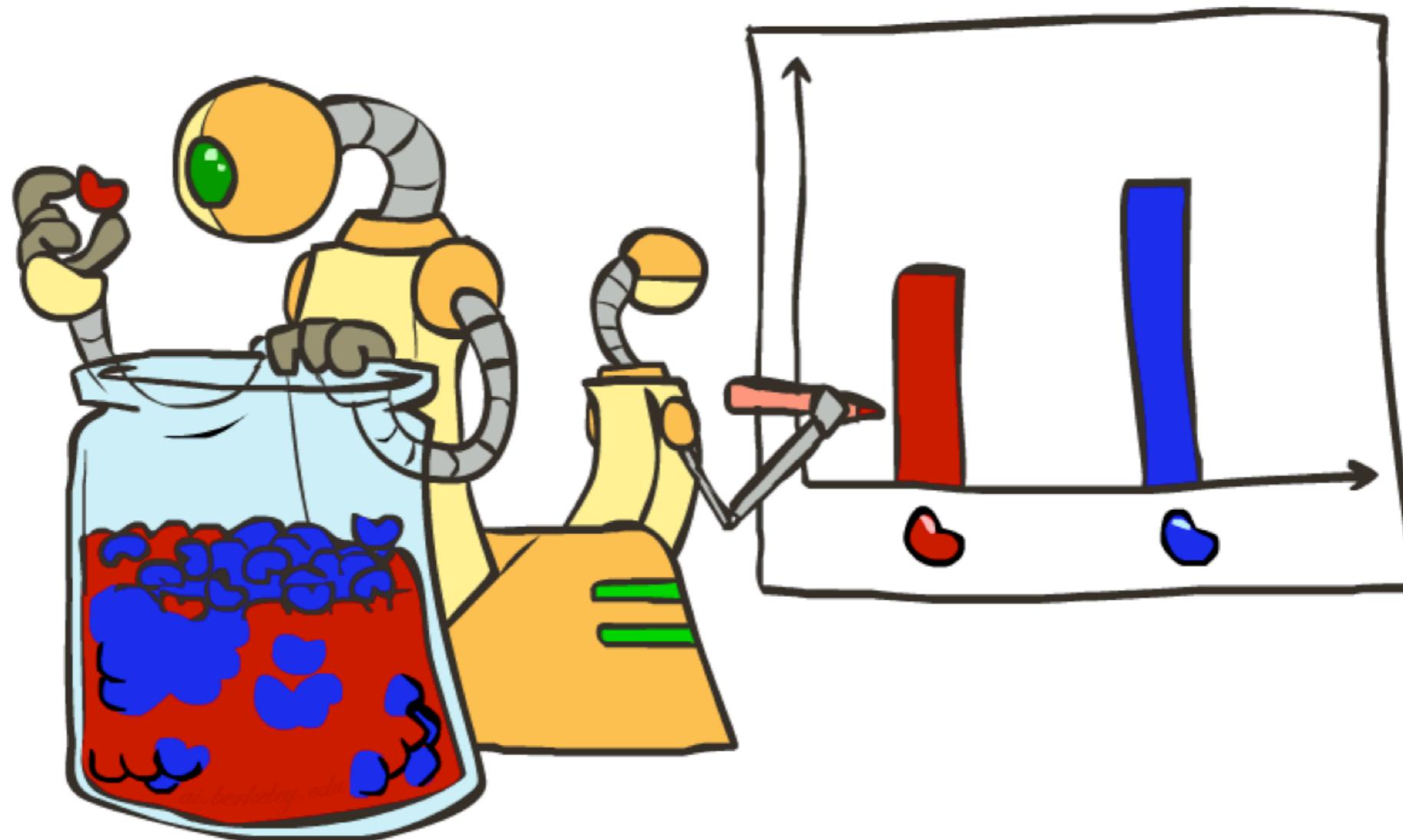


这里出什么问题了？

过拟合与泛化

- 过度拟合训练数据
 - 在训练集中没有像素(15,15)为on的 3 的样本，不意味着测试集中也没有
 - 不太可能每次出现“minute”都是100%的垃圾邮件
 - 同样不太可能每次出现“seriously”都是100%的正常邮件
 - 那些训练集里根本没有出现的单词呢？一般来说，我们不能让没见过的事件发生的概率为零
- 作为一个极端情况，如果使用整个电子邮件作为唯一的特征（例如文档ID）会让模型很容易完美地匹配训练数据，而这个模型一点泛化能力都没有
 - 用词袋(bag-of-words)使得模型具备了一定的泛化能力，但是还不够
- 为了更好地泛化：平滑(smoothing) or 正规化(regularization)

参数估计

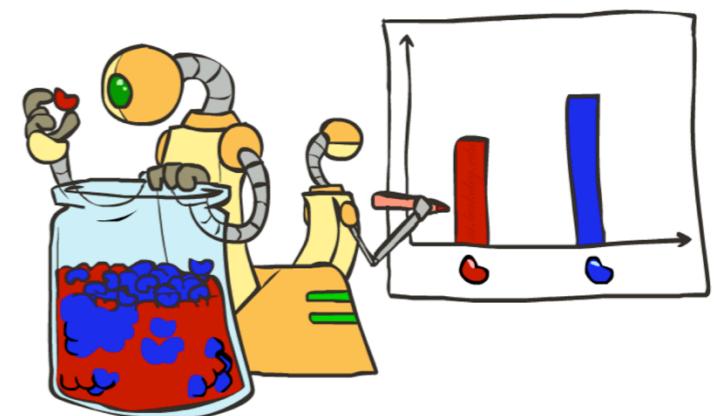


参数估计

- 估计一个随机变量的分布
- 使用训练数据 (机器学习)

- 例如：对于每个结果 x , 查看该值的经验比率：

$$P_{\text{ML}}(x) = \frac{\text{count}(x)}{\text{total samples}}$$



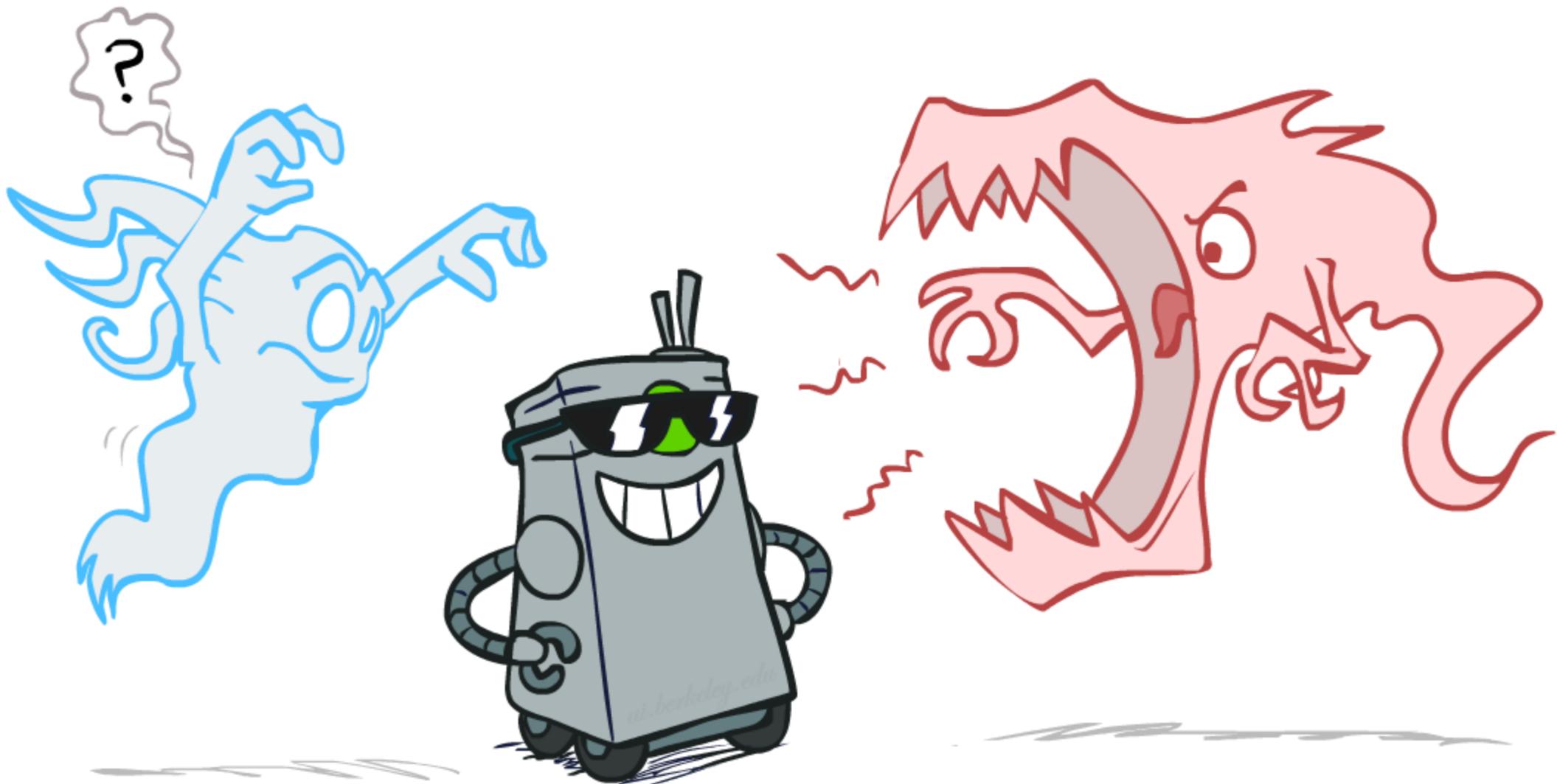
r r b

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

- 这是使数据可能性最大化的估计值

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

平滑 smoothing



最大似然?

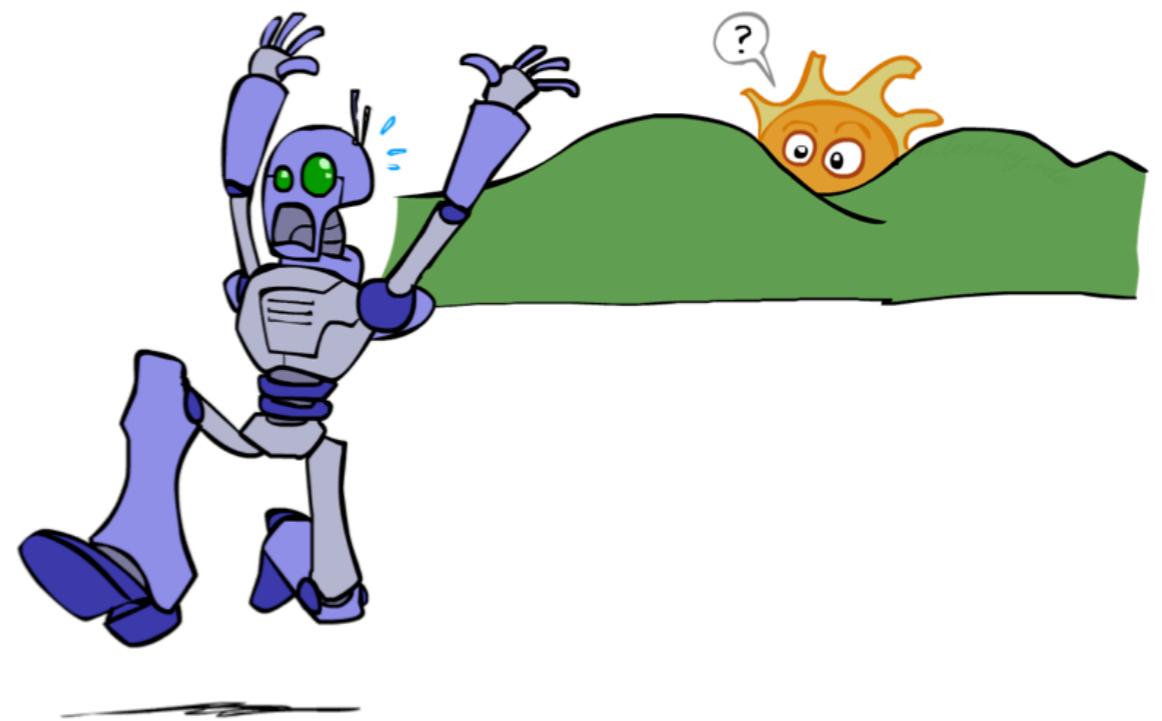
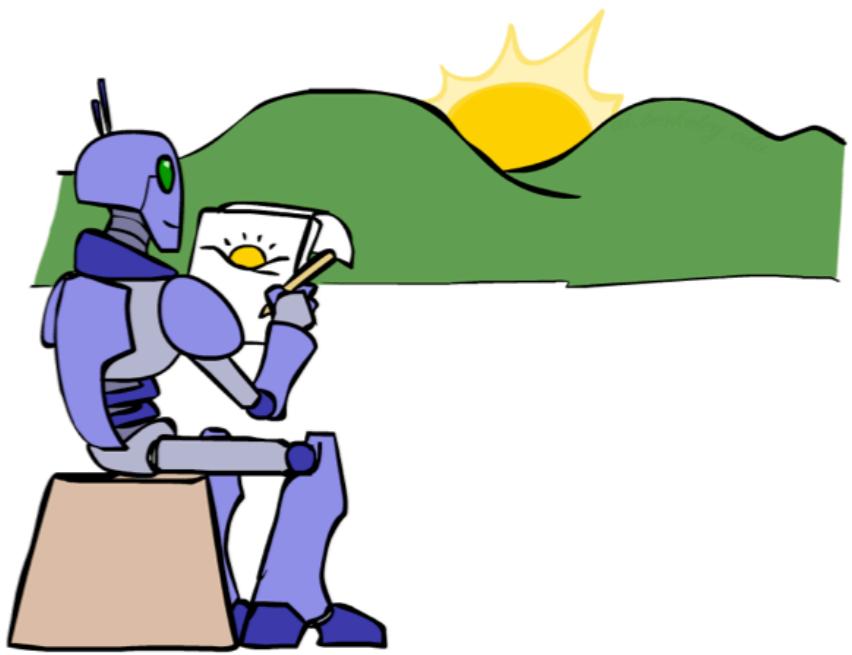
- 最大似然作为相对概率

$$\begin{aligned}\theta_{ML} &= \arg \max_{\theta} P(\mathbf{X}|\theta) \\ &= \arg \max_{\theta} \prod_i P_{\theta}(X_i)\end{aligned}\quad \Rightarrow \quad P_{ML}(x) = \frac{\text{count}(x)}{\text{total samples}}$$

- 另一个选择是考虑给定数据的最有可能的参数值

$$\begin{aligned}\theta_{MAP} &= \arg \max_{\theta} P(\theta|\mathbf{X}) \\ &= \arg \max_{\theta} P(\mathbf{X}|\theta)P(\theta)/P(\mathbf{X}) \\ &= \arg \max_{\theta} P(\mathbf{X}|\theta)P(\theta)\end{aligned}\quad \Rightarrow \quad \text{????}$$

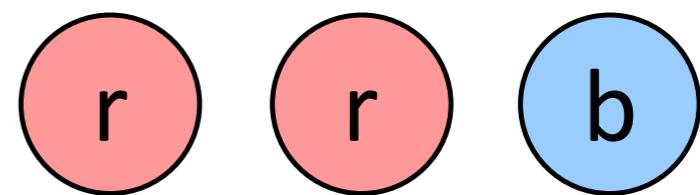
没见过的样本



拉普拉斯平滑

- 拉普拉斯估计：

- 假装你比实际看到的每一个结果都多看一次



$$P_{LAP}(x) = \frac{c(x) + 1}{\sum_x [c(x) + 1]}$$

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

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

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

拉普拉斯平滑

- 拉普拉斯估计 (扩展):

- 假装你比实际看到的每一个结果都多看k次

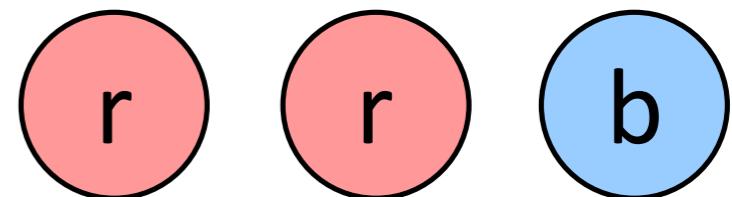
$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

- k 是先验的强度

- 带条件的拉普拉斯估计:

- 独立平滑每个条件:

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



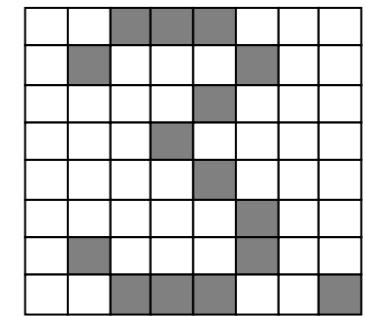
$$P_{LAP,0}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

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

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

估计：线性插值

- 在现实任务中，拉普拉斯在 $|X|$ 或者 $|Y|$ 非常大时表现不佳
- 另一个选择：线性插值
 - 同样从数据中得到经验 $P(X)$
 - 确保 $P(X|Y)$ 的估计值与经验 $P(X)$ 没有太大差异



$$P_{LIN}(x|y) = \alpha \hat{P}(x|y) + (1.0 - \alpha) \hat{P}(x)$$

现实应用: 平滑化估计参数

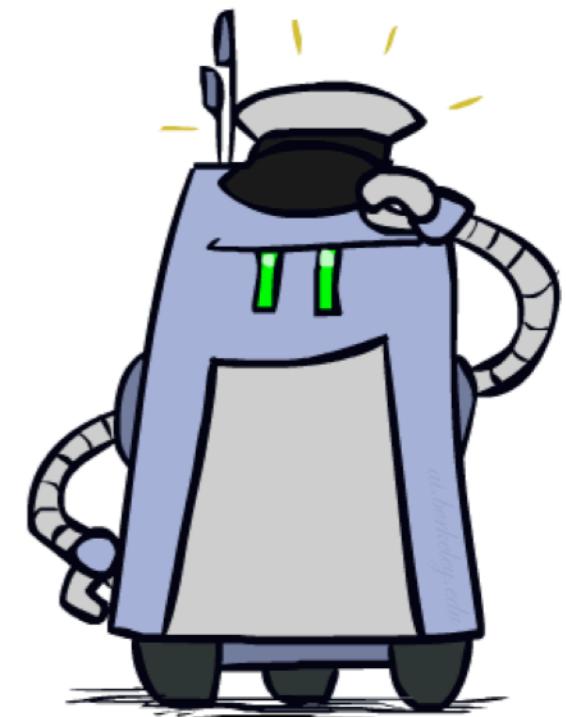
- 对于实际的分类问题, 平滑是至关重要的
- 新条件概率比:

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

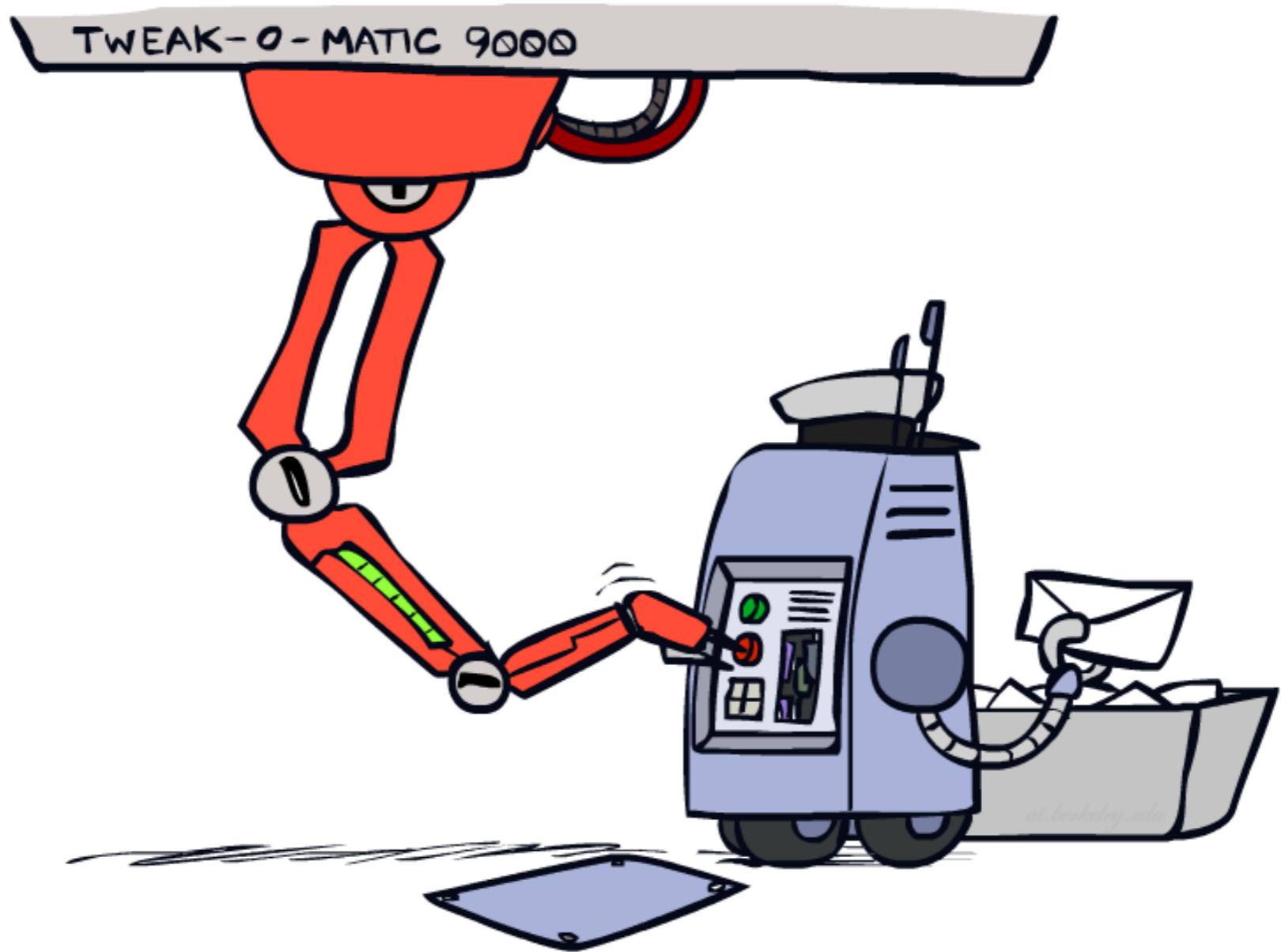
helvetica	:	11.4
seems	:	10.8
group	:	10.2
ago	:	8.4
areas	:	8.3
...		

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

verdana	:	28.8
Credit	:	28.4
ORDER	:	27.2
	:	26.9
money	:	26.5
...		

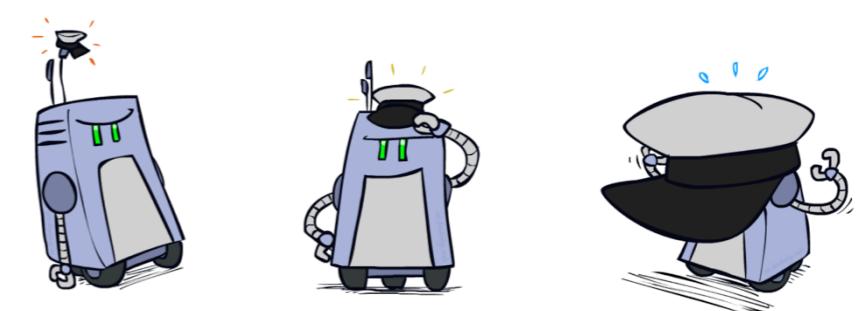
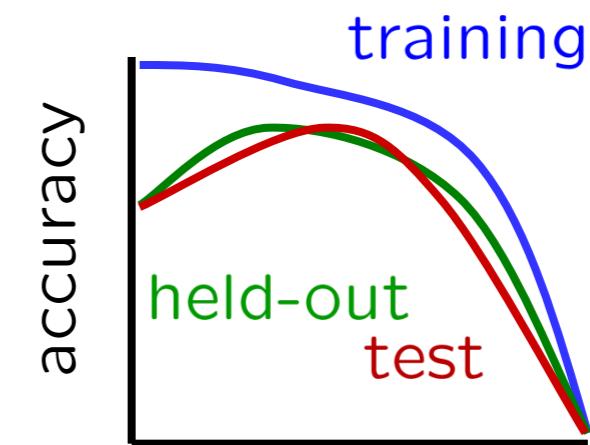


调优Tuning

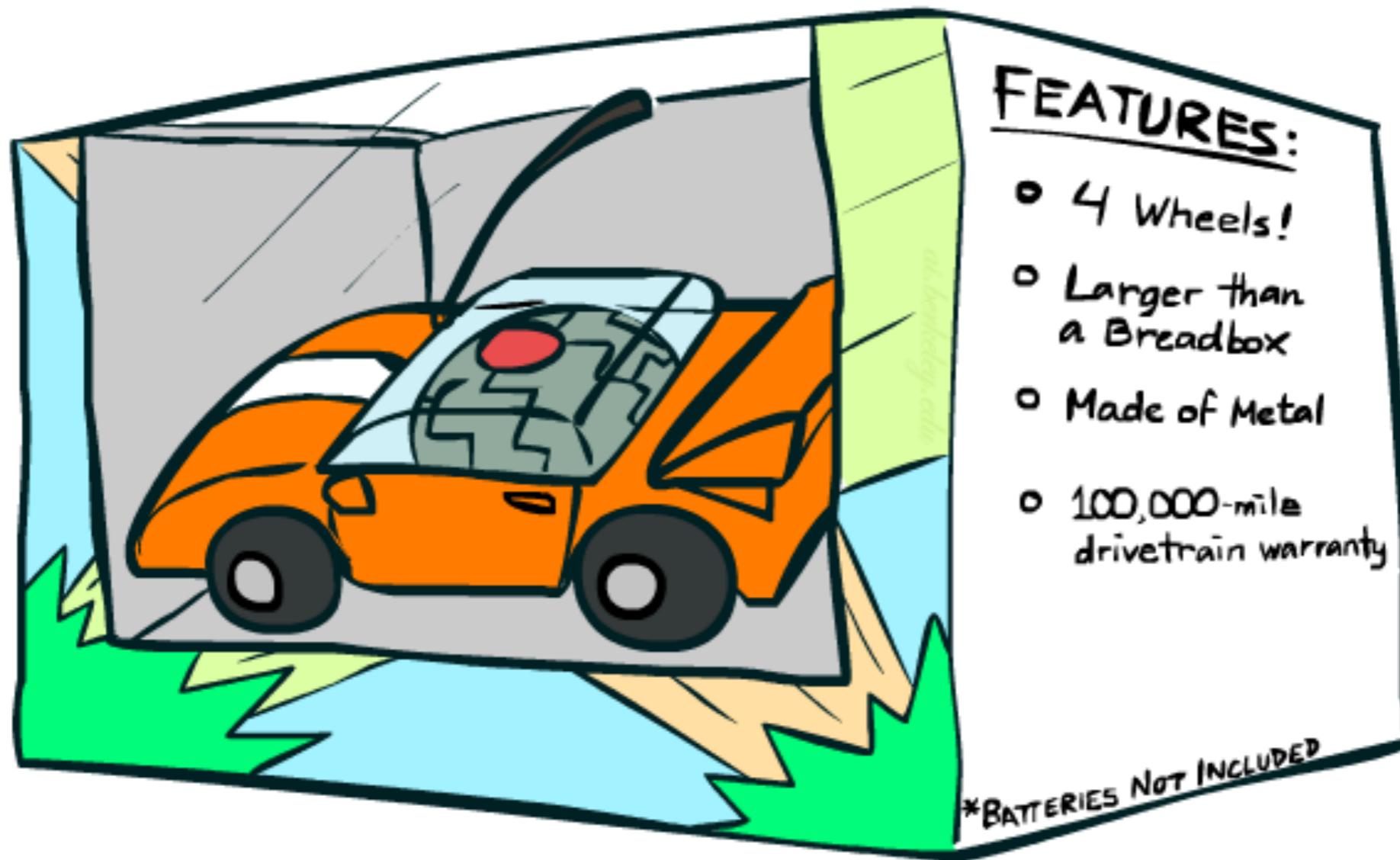


在留出数据集上调优

- 现在有了两组未知数：
 - 参数：概率 $P(X|Y), P(Y)$
 - 超参数：比如平滑的方法和参数 k, a
- 如何进行学习？
 - 从训练数据中学习参数
 - 在不同的数据上调整超参数
- 对于超参数的每个值，在留出集上测试以选择最佳值，并对测试数据进行最终测试



特征



错误，以及该怎么做

- 错误示例

Dear GlobalSCAPE Customer,

GlobalSCAPE has partnered with ScanSoft to offer you the latest version of OmniPage Pro, for just \$99.99* - the regular list price is \$499! The most common question we've received about this offer is - Is this genuine? We would like to assure you that this offer is authorized by ScanSoft, is genuine and valid. You can get the . . .

. . . To receive your \$30 Amazon.com promotional certificate, click through to

<http://www.amazon.com/apparel>

and see the prominent link for the \$30 offer. All details are there. We hope you enjoyed receiving this message. However, if you'd rather not receive future e-mails announcing new store launches, please click . . .

如何处理错误？

- 需要更多的特征
 - 以前给收信人发过邮件吗？
 - 有其他收到同样的邮件吗？
 - 是否一直在发送邮件？
 - 邮件内容全大写？
 - 邮件中有URL嘛？
- 可以将这些信息作为新特征添加到朴素贝叶斯模型中

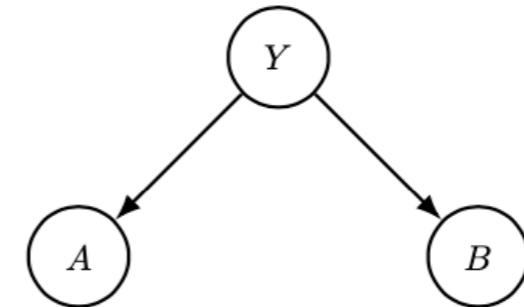


总结

- 贝叶斯法则使得我们可以利用类似因果关系的条件概率，对未知变量进行诊断式查询（推理计算）
- 朴素贝叶斯模型假设所有特征变量之间都是条件于分类标签相互独立的
- 我们可以构建分类器，使用训练数据计算朴素贝叶斯模型的参数
- 对模型参数的平滑化估计在现实应用中很重要

作业

- 我们将用朴素贝叶斯模型，基于两个特征变量 A 和 B ，来判断标签变量 Y 。所有这三个变量都是二元变量，值域都是集合 $\{0, 1\}$ 。以下给出了 10 个训练样本，这些样本点将用来估计模型参数。



A	B	Y
1	1	1
1	0	1
1	0	0
1	1	0
0	1	0
1	1	1
0	1	1
1	0	0
1	1	0
1	1	0

- 1. 请计算该模型的概率分布表的最大似然法估计值是多少？
- 2. 给定一个新的测试样本 ($A = 1, B = 1$)，请计算该模型对这个样本的预测标签是什么？
- 3. 请应用 Laplace 平滑方法重新计算概率分布 $P(A|Y)$ 的值，假定 Laplace 平滑参数 $k = 2$ 。

A	Y	$P(A Y)$
0	0	
1	0	
0	1	
1	1	

B	Y	$P(B Y)$
0	0	
1	0	
0	1	
1	1	

Y	$P(Y)$
0	
1	