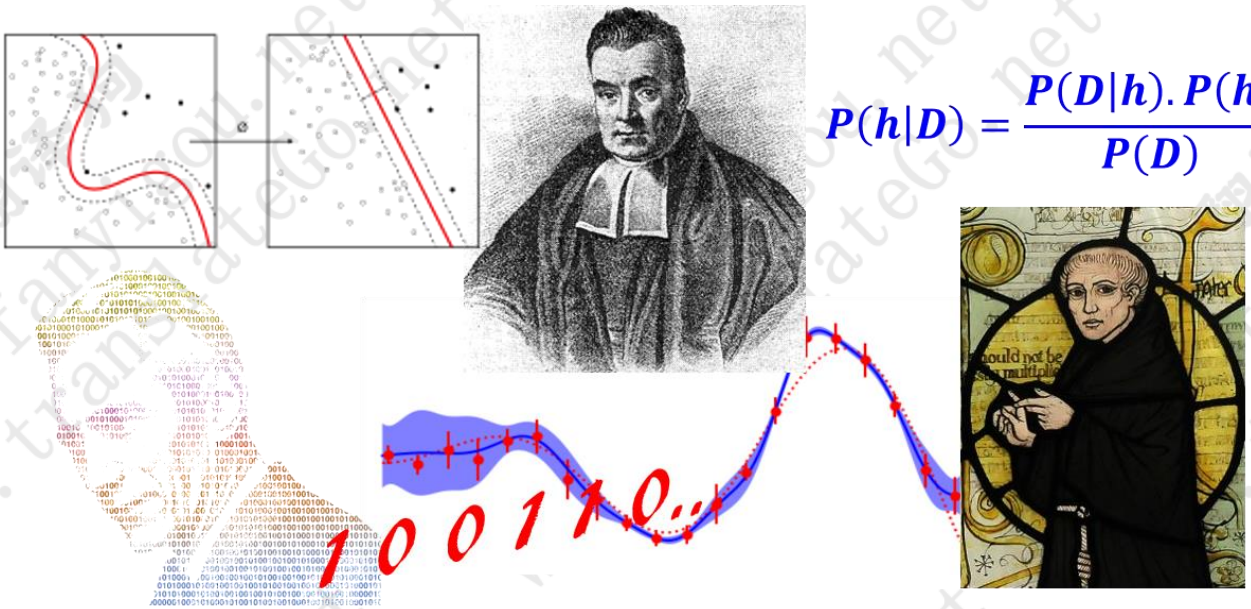
当贝叶斯，奥卡姆和香农走到一起来定义机器学习时

TIRTHAJYOTI SARKAR

一个美丽的想法，它将统计学，信息理论和哲学的概念联系在一起。



介绍

有点令人惊讶的是，在机器学习的所有高级流行语中，我们并没有听到太多关于将统计学习，信息理论和自然哲学的一些核心概念融合到一个三个词组中的短语。

并且，它不仅仅是一个用于机器学习（ML）博士和理论家的模糊和迂腐的短语。 对于任何有兴趣探索的人来说，它都具有精确且易于访问的含义，并为ML和数据科学的从业者提供了实用的支持。

我在谈论最小描述长度。 你可能在想什么是.....

让我们分层o看看它有多么有用.....

贝叶斯和他的定理

我们从（不按时间顺序）开始 牧师□omas贝叶斯顺便说一下，他从未发表过关于如何做统计推断的想法，但后来被同名定理永生化了。

这是□□世纪下半叶，并没有数学科学的分支称为“概率论”。 仅仅通过相当可观的“*机会主义*”-以书的名字命名 *Abraham de Moivre。 一篇名为“解决问题的文章”的文章*

在“机会主义”中，首先由贝叶斯制定，但编辑和 由他的朋友修改理查德普莱斯，*被读给皇家学会并在伦敦皇家学会的哲学交易中发表*□□□□。 在这篇文章中，贝叶斯以一种相当频繁的方式描述了关于联合概率的简单定理，该定理引起逆概率的计算，即贝叶斯定理。

许多战斗都在进行 从那时起，统计科学的两个交战派系 - 贝叶斯和弗雷克斯主义者之间。 但是为了本文的目的，让我们暂时忽略历史，并专注于贝叶斯推理的机制的简单解释。 对于超级 直观的主题介绍，请参阅Brandon Rohrer的这篇精彩教程。 我将专注于等式。

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

Posterior probability　Likelihood　Prior probability

□基本上告诉你在看到数据/证据（可能性）后更新你的信念（先验概率）并将更新的信念程度分配给术语后验概率。 您可以从一个信念开始，但每个数据点都会强化或削弱这种信念，并且您会一直更新您的假设。

听起来简单直观？ 大。

我在段落的最后一句中做了一个技巧。 你注意到了吗？ 我一言不发地说“假设”。 *assumptions*不是正常的英语。 *assumptions*是正式的

stu :)

在统计推断的世界中，假设是一种信念。 它是关于过程的真实性质（我们永远无法观察到）的信念，即产生随机变量（我们可以观察或测量，尽管不是没有噪声）的背后。 在统计学中，它通常被定义为概率分布。 但是在机器学习的背景下，可以考虑任何一组规则（或逻辑或过程），我们认为这些规则可以产生示例或训练数据，我们可以学习这个神秘过程的隐藏性质。

因此，让我们尝试用不同的符号重构贝叶斯定理 - 与数据科学有关的符号。 让我们用d表示数据，用h表示假设。
□是指我们应用贝叶斯的公式，试图确定数据来自哪个假设，给定数据。 我们把定理重写为，

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

现在， 一般来说， 我们有一个很大的（通常是无限的）假设空间， 即许多假设可供选择。 贝叶斯推断的本质是我们想要检查数据，以最大化一个最有可能产生观察数据的假设的概率。 我们基本上想要确定P（h | D）的argmax， 即我们想知道哪个h， 观察到的D最有可能。 为此， 我们可以安全地将该术语放在分母P（D）中， 因为它不依赖于假设。 [illegible] 方案通过相当于舌头扭曲的最大后验（MAP）名称而闻名。

现在，我们应用以下数学技巧，

- 事实上，最大化对于对数和原始函数的工作方式类似，即采用对数不会改变最大化问题。
- 产品的对数是各个对数的总和
- 数量的最大化等同于负数量的最小化

$$
\begin{aligned}
h_{MAP} &= arg\ max\ P(D|h).P(h) \\
&= arg\ max\ log_2(P(D|h).P(h)) \\
&= arg\ max\ [log_2 P(D|h) + log_2 P(h)] \\
&= arg\ min\ [-log_2 P(D|h) - log_2 P(h)]
\end{aligned}
$$

Curiouser和curiouser .....那些带有负对数的□看起来很熟悉...来自Information□edry!

进入克劳德香农。

这需要 很多卷 描述克劳德•香农的天才和陌生生活，克劳德•香农几乎单枪匹马地奠定了信息理论的基础，并将我们带入了现代高速通信和信息交流的时代。

香农的麻省理工学院硕士论文 在电子工程领域被称为□□世纪最重要的硕士论文：其中□□yearoldShannon展示了□其中uryury数学家George Boole的逻辑代数如何使用继电器的电子电路来实现。

开关。 [illegible]数字计算机设计的最基本特征 – 将“真”，“假”，“□”和“□”表示为开关或闭合开关，并使用电子逻辑门进行决策并进行算术运算 – 可以追溯到香农论文中的见解。

但这不是他最伟大的成就。

在□□□□，□Shannon去了贝尔实验室，在那里他从事战争事务，包括密码学。 他还在研究背后的原始理论 信息和通信。 在[illegible]这项工作出现在贝尔实验室研究期刊上发表的一篇广受欢迎的论文中。

香农定义了一个来源产生的信息量 – 为 例如，消息中的数量 – 通过类似于在物理学中确定热力学熵的等式的公式。 从最基本的角度来说， 香农的信息熵是编码消息所需的二进制数字的数量。 对于具有概率p的消息或事件，该消息的最有效（即紧凑）编码将需要log[illegible](p）位。

而这正是出现在贝叶斯定理中的最大后验表达式中出现的那些术语的本质！

因此，我们可以说，在贝叶斯推理的世界中，最可能的假设取决于两个引起长度感的术语 – 而不是最小长度。

$$h_{MAP} = \arg\min\,[\mathit{length}(D|h) + \mathit{length}\,(h)]$$

> 但那些长度的概念可能是什么呢？

长度（h）：奥卡姆的剃刀

奥卡姆的威廉（大约□□□□□□□□□□□□）是英国方济各会修士和 神学家和一个中立的中世纪 哲学家。他作为一个伟大的逻辑学家的流行名声在于他所谓的格言 **奥卡姆剃刀**。术语剃刀是指通过“剃掉”不必要的假设或切断两个类似的结论来区分两个假设。

归于他的确切词汇是：entia non sunt multiplicanda praeter necessitatem（实体不得超过必要性）。 用统计学的说法，这意味着我们必须努力使用能够令人满意地解释所有数据的最简单的假设。

其他名人也回应了类似的原则。

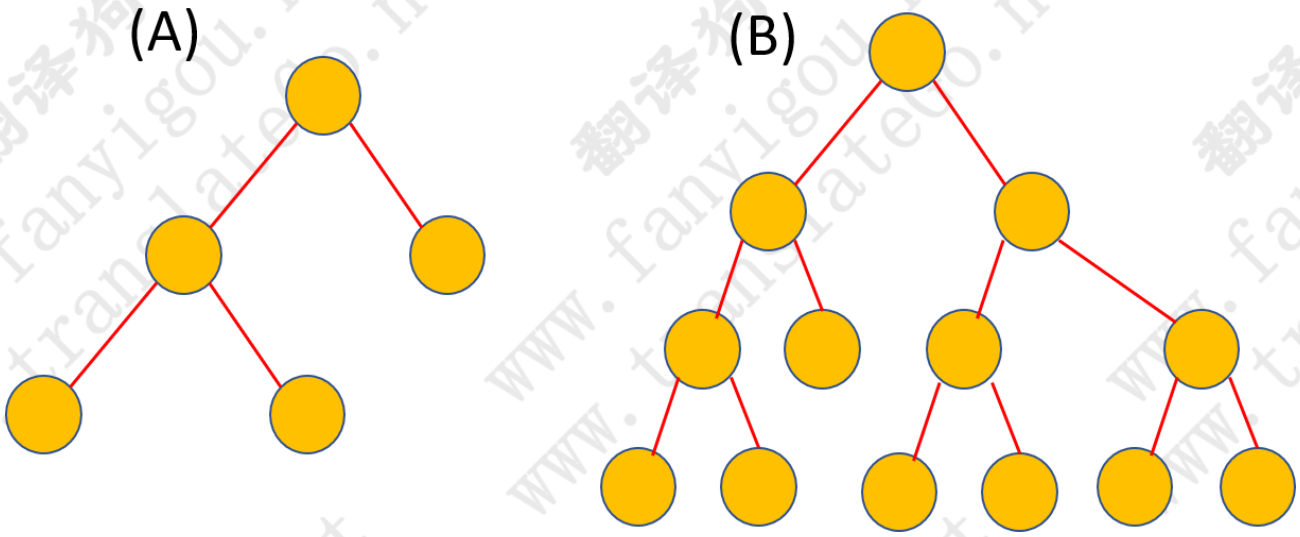
伊萨克●牛顿爵士（Sir Issac Newton）：“我们不会承认自然事物的原因，而不是真实和充分地解释它们的外表。”

Bertrand Russell：“只要有可能，用已知实体的结构代替未知实体的推论。”

> 总是喜欢较短的假设。

需要一个关于假设的长度的例子吗？

以下哪个决策树的长度较短？ A还是B？



即使没有对假设的“长度”进行精确定义，我相信你会认为左边的树（A）看起来更小或更短。 当然，你是对的。因此，较短的假设是具有较少的自由参数，或较不复杂的决策边界（对于分类问题），或这些属性的某种组合，其可以表示其简洁性。

那么’长度（D | h）’怎么样？

给定假设是数据的长度。 那是什么意思？

直觉上，它与假设的正确性或表征能力有关。 除其他事项外，它还包含一个假设，即“推断”数据的好坏程度。 如果假设真的很好地生成数据并且我们可以测量数据无错误，那么我们根本不需要数据。

□ey，第一次出现时 *原理*，他们背后没有任何严格的数学证明。 觎不是定理。 根据对自然体运动的观察，□ey很像假设。 但是他们真的很好地描述了数据。 因此，他们成为了物理定律。

这就是为什么你不需要维护和记忆所有可能加速度数的表格，作为施加在身体上的力的函数。 你只相信紧凑假设，即法则F = ma，并且相信你需要的所有数字，只要在必要时就可以从中计算出来。 它使长度（D | h）非常小。

但是如果数据偏离紧凑假设很多，那么你需要对这些偏差是什么，对它们的可能解释等进行详细描述。

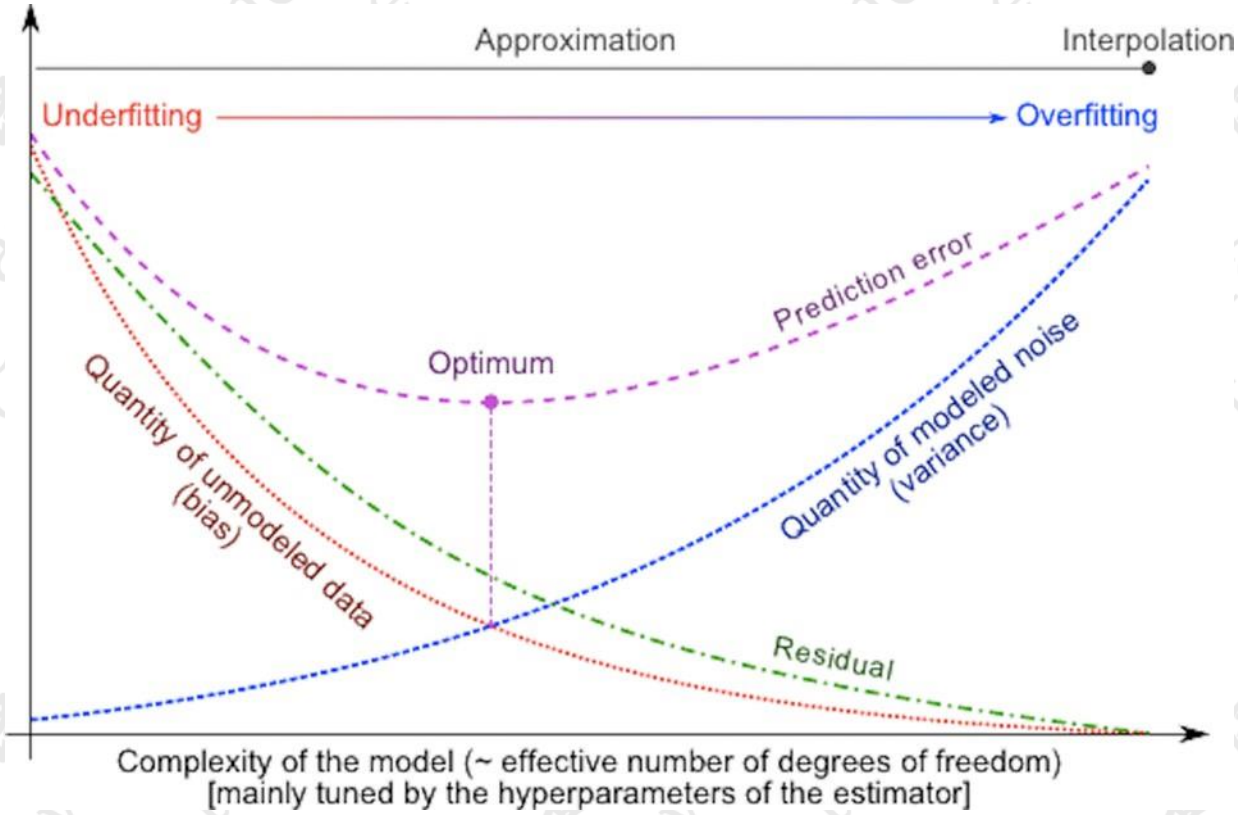
> 因此，长度（D | h）简洁地捕捉了“给定假设的数据有多好”的概念。

实质上，它是错误分类或错误率的概念。 对于一个完美的假设，在极限情况下它很短，为零。 对于一个完全没有数据的假设，它往往很长。

> 而且，这就是贸易。

如果你用一个大奥卡姆剃刀剃掉你的假设，你可能会留下一个简单的模型，一个不能完成所有数据的模型。 因此，您必须提供更多数据才能获得更好的信任。 另一方面，如果你创建一个复杂（和长）的假设，你可能能够很好地训练你的训练数据，但这实际上可能不是正确的假设，因为它违背了MAP原则，即假设小熵。

> 听起来像偏见变数贸易？ 是的，也是:）



资源：
https://www.reddit.com/r/mlclass/comments/mmlfu/a_nice_alternative_explanation_of_bias_and/

把它们放在一起

因此，贝叶斯推断告诉我们，最好的假设是最小化两个项之和的假设：假设的长度和错误率。

> 在这个深刻的句子中，它几乎捕获了所有（受监督的）机器学习。

its ra,

- 线性模型的模型复杂度 - 选择多项式的程度， 如何减少sumofsquare残差
- 选择神经网络的架构 - 如何不公开训练数据并获得良好的验证准确性，但减少分类错误。

- 支持向量机正则化和内核选择 - 软边界与硬边界之间的平衡，即与决策边界非线性交换精度。

我们真正得出的结论是什么？

我们从最小描述长度（MDL）原理的分析中得出什么结论？

> 这是否一劳永逸地证明了短假设是最好的？

没有。

MDL显示的是，如果选择假设的表示使得假设h的大小为-log恿P（h），并且如果选择异常（错误）的表示，则给定h的D的编码长度相等记录P（D | h），然后MDL原则产生MAP假设。

然而，为了表明我们有这样的表示，我们必须知道所有先验概率P（h），以及P（D | h）。 is是没有理由相信MDL假设相对于假设和错误/错误分类的任意编码应该是首选。

> 对于实际的机器学习，人类设计者有时可能更容易指定捕获关于假设的相对概率的知识的表示，而不是完全指定每个假设的概率。

二是知识表示和领域专业知识变得至关重要的地方。 它使（通常）无限大的假设空间短路，并引导我们走向一组极有可能的假设

我们可以对其进行最佳编码并努力确定其中的MAP假设集。

总结和思考后

一个很好的事实是，对概率论的基本身份进行这样一套简单的数学操作可以导致对监督机器学习的基本限制和目标进行如此深刻和简洁的描述。 对于这些问题的简明处理，读者可以参考 这博士 论文，来自卡内基梅隆大学的“为什么机器学习工作”。 考虑所有这些问题也值得考虑 连接到 NoFreeLunch定理.

如果您对此领域的深入阅读感兴趣