

## 003 | 数据科学家基础能力之概率统计

2017-10-10 洪亮劫

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 10:19 大小 4.73M



学习人工智能的工程师，甚至是在人工智能相关领域从业的数据科学家，往往都不重视概率统计知识的学习和培养。有人认为概率统计知识已经过时了，现在是拥抱复杂的机器学习模型的时候了。实际上，概率统计知识和数据科学家的日常工作，以及一个人工智能项目的正常运作都密切相关，**概率统计知识正在人工智能中发挥着越来越重要的作用。**

和机器学习一样，概率统计各个领域的知识以及研究成果浩如烟海。今天我就和你聊一聊，如何从这么繁多的信息中，掌握能够立即应用到实际问题中的概率统计知识，以及如何快速入手一些核心知识，并能触类旁通学习到更多的内容。

### 使用概率的语言

概率统计中的“概率”，对于学习和掌握人工智能的诸多方面都有着举足轻重的作用。这里面最重要的，恐怕要数概率论中各种分布的定义。初学者往往会觉得这部分内容过于枯燥乏味，实际上，**概率论中的各种分布就像是一门语言的基本单词，掌握了这些基本的“建模语言”单词，才能在机器学习的各个领域游刃有余。**

值得注意的是，目前火热的深度学习模型，以及在之前一段时间占领机器学习统治地位的概率图模型（Probabilistic Graphical Models），都依赖于概率分布作为这些框架的基本建模语言。因此，能够真正掌握这些分布就显得尤为重要。

对于分布的掌握其实可以很容易。只要对少量几个分布有一定的认识后，就能够很容易地扩展开来。首先，**当你遇到一个实际场景的时候，你要问自己的第一个问题是，这个场景是针对离散结果建模还是针对连续数值建模？**这是一个最重要的分支决策，让你选择正确的建模工具。

当面对离散结果的时候，最需要掌握的分布其实就是三个：

1. 伯努利分布
2. 多项分布
3. 泊松分布

这三种分布是其他离散分布的重要基础。对于这三种分布，记忆其实也相对容易。比如，任何时候，如果你的场景是一个二元问题（例如用户是否点击，是否购买），伯努利分布都是最直接的选择。当你遇到的场景需要有多于两种选择的时候，那自然就用多项分布。另外，文本建模常常可以看做这样一个问题，那就是在特定语境下，如何从上千甚至上万的可能性中选择出最恰当的字词，因此多项分布也广泛应用在文本建模领域。泊松分布则常常用在可对数的整数进行建模，比如一些物品的总个数。

**了解应用场景和他们所对应的分布之间的联系，是掌握这些“语言”的重要环节。**当你面临的问题是连续数值的时候，绝大多数情况下，你需要掌握和理解正态分布，有时候称为高斯分布。正态分布的重要性是再怎么强调都不为过的。任何你可以想到的场景，几乎都可以用正态分布来建模。由于中心极限定理的存在，在大规模数据的情况下，很多其他分布都可以用正态分布来近似或者模拟。因此，**如果说学习概率知识中你只需要掌握一种分布的话，那无疑就是正态分布。**

在理解概率分布的过程中，还需要逐渐建立起关于“随机数”和“参数”的概念。衡量一个分布是离散还是连续，指的是它产生的“随机数”是离散还是连续，和这个分布的“参数”没有关系。比如伯努利分布是一个离散分布，但是伯努利分布的参数则是一个介于 0 和 1 之间的实数。理解这一点常常是初学者的障碍。另外，建立起参数的概念以后，所有的分布就有了模型（也就是分布本身）和参数的估计过程两个方面。这对理解机器学习中模型和算法的分离有很直接的帮助。

当理解了这些概率最基础的语言以后，下面需要做的就是，了解贝叶斯统计中，怎么针对概率分布定义先验概率，又怎么推导后验概率。

了解贝叶斯统计不是说一定要做比较困难的贝叶斯估计，而是说，怎么利用先验概率去对复杂的现实情况进行建模。比如说，针对用户是否购买某一件商品而言，这个问题可以用一个伯努利分布来建模。假如我们又想描述男性和女性可能先天上就对这个商品有不同的偏好，这个时候，我们就可以在伯努利分布的参数上做文章。也就是说，我们可以认为男性和女性拥有不同的参数，然而这两个参数都来自一个共同的先验概率分布（也可以认为是全部人群的购买偏好）。那么这个时候，我们就建立起了一个具有先验的模型来描述数据。这个思维过程是需要初学者去琢磨和掌握的。

## 假设检验

如果说概率基础是一般学习人工智能技术工程师和数据科学家的薄弱环节，假设检验往往就是被彻底遗忘的角落。我接触过的很多统计背景毕业的研究生甚至博士生，都不能对假设检验完全理解吃透。实际上，**假设检验是现实数据分析和数据产品得以演化的核心步骤。**

对于一款数据产品，特别是已经上线的产品来说，能够持续地做线上 A/B 测试，通过 A/B 测试检测重要的产品指标，从而指导产品迭代，已经成为产品成败的关键因素。这里面，通过 A/B 测试衡量产品指标，或多或少就是做某种形式的假设检验。

你期望提高产品性能，那么如何理解假设检验，选取合适的工具，理解 P 值等一系列细节就至关重要，这些细节决定了你辛辛苦苦使用的复杂人工智能模型算法是否有实际作用。

首先，我们要**熟悉假设检验的基本设定**。比如，我们往往把现在的系统情况（比方说用户的点击率、购买率等）当做零假设，或者通常叫做  $H_0$ 。然后把我们的改进的系统情况或者算法产生的结果，叫做备择假设，或者叫做  $H_1$ 。

接下来，一个重要的步骤就是**检验目前的实验环境**，看是否满足一些标准检验的假设环境，比如 T 检验、Z 检验等。这一步往往会困扰初学者甚至是有经验的数据科学家。一个非常粗略的窍门则是，因为中心极限定理的存在，Z 检验通常是一个可以缺省使用的检验，也就是说，在绝大多数情况下，如果我们拥有大量数据可供使用，一般会选择 Z 检验。当然，对于初学者而言，最常规的也是最需要的就是掌握 T 检验和 Z 检验，然后会灵活使用。

在选择了需要的检验以后，就要**计算相应的统计量**。然后根据相应的统计量以及我们选好的检验，就可以得到一系列的数值，比如 P 值。然后利用 P 值以及我们预先设定的一个范围值，比如经常设置的 0.95（或者说 95%），我们往往就可以确定，H1 是否在统计意义上和 H0 不同。如果 H1 代表着新算法、新模型，也就意味着新结果比老系统、老算法有可能要好。

需要你注意的是，这里说的是“有可能”，而不是“一定”、“确定”。从本质上来说，假设检验并不是金科玉律。假设检验本身就是一个统计推断的过程。我们在假设检验的流程中计算的，其实是统计量在 H0 假设下的分布中出现的可能性。可能性低，只能说，我们观测到的现象或者数值并不支持我们的 H0，但这个流程并没有去验证这些现象或者数值是不是更加支持 H1。

另外，即便“可能性”低，也并不代表绝对不出现。这也是初学者常常过度相信假设检验所带来的问题。**比较正确的对待假设检验的态度，就是把这个流程提供的结果当做工具，与更加复杂的决策过程结合起来，从而对目前的系统、目前的产品有一个综合的分析。**

值得注意的是，和假设检验有关联的一个概念“**置信区间**”往往也很容易被忽视。尽管初看没有太大作用，置信区间其实被广泛应用在推荐系统的“利用和探索”（Exploitation & Exploration）策略中。因此，明白置信区间的概念很有益处，对实际的计算有很大帮助。

## 因果推论

最后我想提一下**因果推论**（Causal Inference）。因果推论不是一般的统计教科书或者工程类学生接触到的统计教科书里的基本内容。然而最近几年，这个领域在机器学习界获得了越来越多的关注。对于学习机器学习前沿知识的朋友来说，了解因果推论十分必要。

同时，对于工程产品而言，并不是所有情况都能通过 A/B 测试来对一个希望测试的内容、模型、产品设计进行测试，并在一定时间内找到合理的结果。在很多情况下是不能进行测试的。因此，**如何在不能进行测试的情况下，还能通过数据研究得出期望的结果，这就是因果推论的核心价值**。基于此，越来越多的互联网公司开始关注这个技术。

对于多数人工智能工程师而言，因果推论所需要的场景其实无时无刻不陪伴着我们。一个常见的情况是，我们需要用数据来训练新的模型或者算法。这里面的数据采集自目前线上的系统，比如一个新闻推荐系统。然而，现在的线上系统是有一定偏差的，例如比较偏好推荐娱乐新闻。那么，这个偏差就会被记录到数据里，我们收集的数据就侧重于娱乐新闻。那么，**要想在一个有偏差的数据中，依然能够对模型和算法进行无偏差的训练和评测，就可以运用因果推论为机器学习带来的一系列工具。**

## 小结

今天我为你讲了掌握概率统计基础知识的一些核心思路。一起来回顾下要点：第一，学习概率分布的语言对于理解、甚至是创造新的机器学习模型和算法都有着重要作用。第二，假设检验是常常被人工智能工程师和数据科学家遗忘的知识。然而，它对我们做产品开发却至关重要。第三，因果推论是一个新兴的统计和机器学习结合的领域，希望你能有所了解。

最后，给你留一个思考题，我们之前说到假设检验约等于我们计算统计量在  $H_0$  里发生的可能性，那么，为什么我们不直接计算在  $H_1$  里发生的可能性呢？

欢迎你给我留言，和我一起讨论。




# AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管  
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

上一篇 002 | 聊聊2017年KDD大会的时间检验奖

下一篇 004 | 数据科学家基础能力之机器学习

## 精选留言 (14)

写留言



沉 置顶

2017-10-17

6

感觉找到了一个很好的指导手册，对假设检验和置信区间的意义有了更深的理解



吴文敏

2017-10-19

39

可否在每期文章的最后推荐一些相应的参考资料或书籍？

展开



rayeaster

2017-10-17

8

统计方面的好书求推荐

展开



JIA

2017-10-16

7

概率知识确实很重要，但是好多内容都忘记了，请问洪老师，怎样可以最快地补起来，有什么资料推荐吗？



五岳寻仙

2018-09-21

4

H0往往是现有状况，分布和参数都已知，容易计算出概率。

比如，要检验某种药物是否有降压作用。H0：药物无降压作用；H1：药物有降压作用。

计算H0概率就很简单，因为我们知道正常人血压值和波动程度，就能很容易地计算出出现某种情况的概率。

展开



阿珂

2018-09-21

👍 1

之所以在 $H_0$ 中计算是由于 $H_0$ 假设有着足够的情景下的数据样本进行计算。

展开 ▾



李志鹏

2018-07-28

👍 1

能不能推荐些比较的好的统计学书，比如因果推论

展开 ▾



yaolixu

2018-06-18

👍 1

假设计算 $H_1$ 发生的可能性, 然后不容易估计 $H_0$ 出错的概率值, 进而计算置信区间了。



开心果

2018-03-09

👍 1

全称命题可以证伪，而不能证实。

展开 ▾



夏

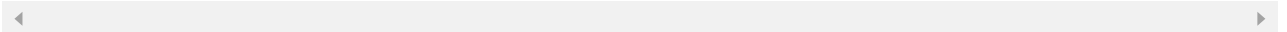
2017-12-12

👍 1

有没有因果推论相关的科普文章推荐？

展开 ▾

作者回复: 可以看看David Sontag在ICML 2016上的Causal Inference for Observational Studies的讲座。



徐涛

2019-03-25

👍

不懂老师的问题。 $H_0$ 和 $H_1$ 都要算啊，两者数据有进行比较才能接受或拒绝假设啊。





halo128



因为假设检验对构造统计量上基于 $H_0$ 进行构造的，所以检验判断也就针对 $H_0$ 的接受与否。当然 $H_0$ 的接收与拒绝可能性，又与 $H_0$ 本身的真假有关，涉及到统计里面犯第一类错误、第二类错误的概率。

如果想要得到 $H_0$ 和 $H_1$ 各自的检验可能性，可以考虑使用贝叶斯的相关假设检验方法。

展开 ▾



小千

2018-12-02



还有一个最基础的问题，就是概率的有效性，测度论告诉我们，随机事件是可以用概率描述的，但是还有不确定事件是不能用概率描述的（或者说用概率近似描述会出很大的问题），因为不确定事件是勒贝格不可积。这就是为什么很多模型在这预测问题上失效的原因之一，因为很多预测问题是不确定的。



黄淮

2017-11-21



思考题的答案是？为何不直接计算  $H_1$  可能性？

展开 ▾