# 和平之城中的鸟类识别(案例研究)

## 一、问题陈述

这个例子来源于实际项目，但是为了保护机密性，我们会对细节进行保护。



  现在你是和平之城的著名研究员，和平之城的人有一个共同的特点：他们害怕鸟类。为了保护他们，你必须设计一个算法，以检测**飞越和平之城的任何鸟类**，同时警告人们有鸟类飞过。市议会为你提供了10,000,000张图片的数据集，这些都是从城市的安全摄像头拍摄到的。它们被命名为:

y = 0: 图片中没有鸟类；

y = 1: 图片中有鸟类。

你的目标是设计一个算法，能够对和平之城安全摄像头拍摄的新图像进行分类。

有很多决定要做：

评估指标是什么？

你如何将你的数据分割为训练/开发/测试集?

**成功的指标**

市议会告诉你，他们想要一个算法：

1.拥有较高的准确度；

2.快速运行，只需要很短的时间来分类一个新的图像；

3.可以适应小内存的设备，这样它就可以运行在一个小的处理器上，它将用于城市的安全摄像头上。

**请注意**: **有三个评估指标使您很难在两种不同的算法之间进行快速选择，并且会降低您的团队迭代的速度，是真的吗？**

* 【★】正确
* 【 】错误

## 二、问题2

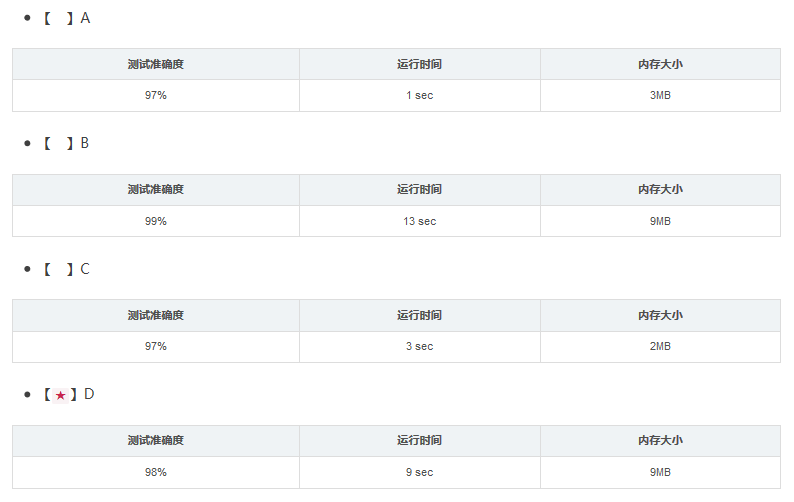
经过进一步讨论，市议会缩小了它的标准：

“我们需要一种算法，可以让我们尽可能精确的知道一只鸟正飞过和平之城。”

“我们希望经过训练的模型对新图像进行分类不会超过10秒。”

“我们的模型要适应10MB的内存的设备。”

**如果你有以下三个模型，你会选择哪一个？**



## 三、问题3

**根据城市的要求，您认为以下哪一项是正确的？**

【★】准确度是一个优化指标；运行时间和内存大小是令人满意的指标。

【 】准确度是一个令人满意的指标；运行时间和内存大小是一个优化指标。

【 】准确性、运行时间和内存大小都是优化指标，因为您希望在所有这三方面都做得很好。

【 】准确性、运行时间和内存大小都是令人满意的指标，因为您必须在三项方面做得足够好才能使系统可以被接受。

## 四、问题4

## **结构化你的数据**

**在实现你的算法之前，你需要将你的数据分割成训练/开发/测试集，你认为哪一个是最好的选择？**



## 五、问题5

  在设置了训练/开发/测试集之后，市议会再次给你了1,000,000张图片，称为“公民数据”。显然，和平之城的公民非常害怕鸟类，他们自愿为天空拍照并贴上标签，从而为这些额外的1,000,000张图像贡献力量。这些图像与市议会最初给您的图像分布不同，但您认为它可以帮助您的算法。

**你不应该将公民数据添加到训练集中，因为这会导致训练/开发/测试集分布变得不同，从而损害开发集和测试集性能，是真的吗？**

【 】True

【★】False

## 六、问题6

**市议会的一名成员对机器学习知之甚少，他认为应该将1,000,000个公民的数据图像添加到测试集中，你反对的原因是：**

【★】这会导致开发集和测试集分布变得不同。这是一个很糟糕的主意，因为这会达不到你想要的效果。

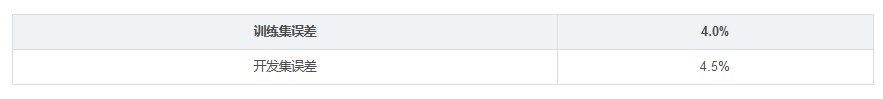
【 】公民的数据图像与其他数据没有一致的x- >y映射(类似于纽约/底特律的住房价格例子)。

【 】一个更大的测试集将减慢迭代速度，因为测试集上评估模型会有计算开销。

【★】测试集不再反映您最关心的数据(安全摄像头)的分布。（博主注：训练集是摄像头拍的，用他人拍的数据去测试摄像头拍的，势必会导致准确度下降，要添加也应该添加到整个数据集中，保证同一分布。）

## 七、问题7

你训练了一个系统，其误差度如下（误差度 = 100％-准确度）：



**这表明，提高性能的一个很好的途径是训练一个更大的网络，以降低4%的训练误差。你同意吗？**

【 】是的，因为有4%的训练误差表明你有很高的偏差。

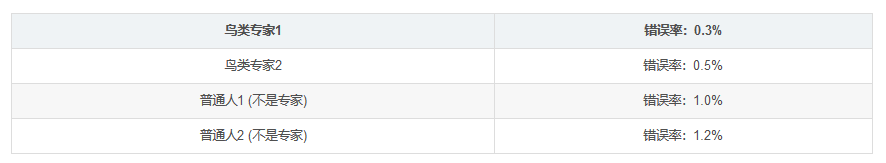
【 】是的，因为这表明你的模型的偏差高于方差。

【 】不同意，因为方差高于偏差。

【★】不同意，因为没有足够的信息，这什么也说明不了。（博主注：想一下贝叶斯最优误差，我们至少还要一个人们对图片的识别误差值，请看下面的题。）

## 八、问题8

你训练了一个系统，其误差度如下（误差度 = 100％-准确度）：



**如果您的目标是将“人类表现”作为贝叶斯错误的基准线（或估计），那么您如何定义“人类表现”？**

【 】0.0% (因为不可能做得比这更好)

【★】0.3% (专家1的错误率)

【 】0.4% (0.3 到 0.5 之间)

【 】0.75% (以上所有四个数字的平均值)

## 九、问题9

**您同意以下哪项陈述？**

【★】学习算法的性能可以优于人类表现，但它永远不会优于贝叶斯错误的基准线。

【 】学习算法的性能不可能优于人类表现，但它可以优于贝叶斯错误的基准线。

【 】学习算法的性能不可能优于人类表现，也不可能优于贝叶斯错误的基准线。

【 】学习算法的性能可以优于人类表现，也可以优于贝叶斯错误的基准线。

## 十、问题10

  你发现一组鸟类学家辩论和讨论图像得到一个更好的0.1%的性能，所以你将其定义为“人类表现”。在对算法进行深入研究之后，最终得出以下结论：



**根据你的资料，以下四个选项中哪两个尝试起来是最有希望的？（两个选项。）**

【 】尝试增加正则化。

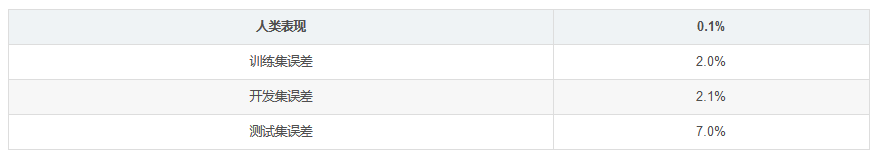
【 】获得更大的训练集以减少差异。

【★】尝试减少正则化。

【★】训练一个更大的模型，试图在训练集上做得更好。

## 十一、问题11

你在测试集上评估你的模型，并找到以下内容：



**这意味着什么？（两个最佳选项。）**

【 】你没有拟合开发集。

【★】你应该尝试获得更大的开发集。

【 】你应该得到一个更大的测试集。

【★】你对开发集过拟合了。

## 十二、问题12

在一年后，你完成了这个项目，你终于实现了：



**你能得出什么结论？（检查所有选项。）**

【★】现在很难衡量可避免偏差，因此今后的进展将会放缓。

【 】统计异常(统计噪声的结果)，因为它不可能超过人类表现。

【 】只有0.09％的进步空间，你应该很快就能够将剩余的差距缩小到0％

【★】如果测试集足够大，使得这0.05%的误差估计是准确的，这意味着贝叶斯误差是小于等于0.05的。

## 十三、问题13

  事实证明，和平之城也雇佣了你的竞争对手来设计一个系统。您的系统和竞争对手都被提供了相同的运行时间和内存大小的系统，您的系统有更高的准确性。然而，当你和你的竞争对手的系统进行测试时，和平之城实际上更喜欢竞争对手的系统，因为即使你的整体准确率更高，你也会有更多的假阴性结果(当鸟在空中时没有发出警报)。**你该怎么办？**

【 】查看开发过程中开发的所有模型，找出错误率最低的模型。

【 】要求你的团队在开发过程中同时考虑准确性和假阴性率。

【★】重新思考此任务的指标，并要求您的团队调整到新指标。

【 】选择假阴性率作为新指标，并使用这个新指标来进一步发展。

## 十四、问题14

  你轻易击败了你的竞争对手，你的系统现在被部署在和平之城中，并且保护公民免受鸟类攻击！ 但在过去几个月中，一种新的鸟类已经慢慢迁移到该地区，因此你的系统的性能会逐渐下降，因为您的系统正在测试一种新类型的数据。（博主注：以系统未训练过的鸟类图片来测试系统的性能）

**你只有1000张新鸟类的图像，在未来的3个月里，城市希望你能更新为更好的系统。你应该先做哪一个？**

【★】使用所拥有的数据来定义新的评估指标（使用新的开发/测试集），同时考虑到新物种，并以此来推动团队的进一步发展。

【 】把1000张图片放进训练集，以便让系统更好地对这些鸟类进行训练。

【 】尝试数据增强/数据合成，以获得更多的新鸟的图像。

【 】将1,000幅图像添加到您的数据集中，并重新组合成一个新的训练/开发/测试集

## 15. 问题15

  市议会认为在城市里养更多的猫会有助于吓跑鸟类，他们对你在鸟类探测器上的工作感到非常满意，他们也雇佣你来设计一个猫探测器。（哇~猫探测器是非常有用的，不是吗？）由于有多年的猫探测器的工作经验，你有一个巨大的数据集，你有100,000,000猫的图像，训练这个数据需要大约两个星期。**你同意哪些说法？（检查所有选项。）**

【★】需要两周的时间来训练将会限制你迭代的速度。

【★】购买速度更快的计算机可以加速团队的迭代速度，从而提高团队的生产力。

【★】如果100,000,000个样本就足以建立一个足够好的猫探测器，你最好用100,000,00个样本训练，从而使您可以快速运行实验的速度提高约10倍，即使每个模型表现差一点因为它的训练数据较少。

【 】建立了一个效果比较好的鸟类检测器后，您应该能够采用相同的模型和超参数，并将其应用于猫数据集，因此无需迭代。