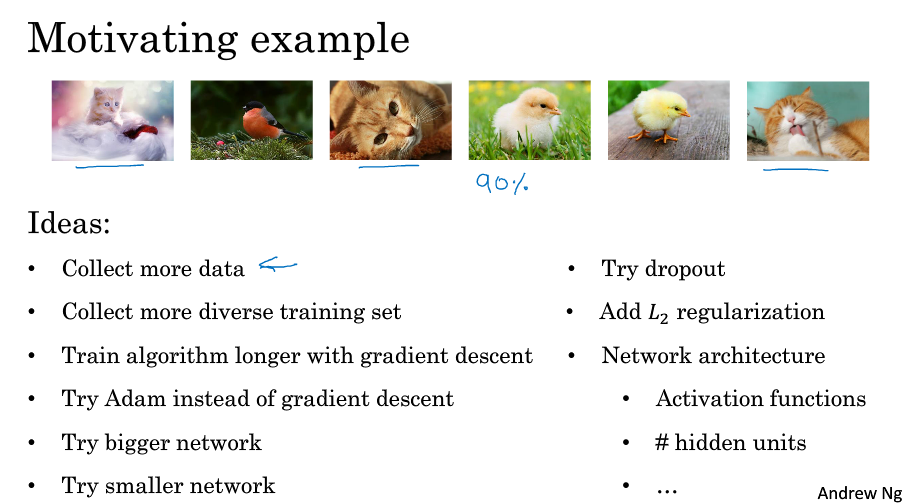
**第三节课**

**第一周：**



正交化正交性是一种系统设计属性，它确保修改一条指令或算法的一个组件不会给系统的其他组件带来或传播副作用。它使得独立验证算法变得更加容易，减少了测试和开发时间当设计监督学习系统时，

这4个假设需要是真实和正交的。

1. 良好的成本管理培训如果它不适合，使用更大的神经网络或切换到更好的优化算法可能会有所帮助。

2. 在成本功能上适合开发如果不适合，正则化或使用更大的训练集可能会有所帮助。

3.成本函数的拟合测试如果不太合适，使用更大的开发集可能会有所帮助

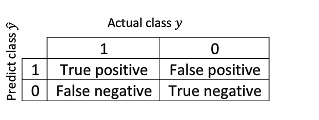
4. 在现实世界中表现良好如果它执行得不好，则开发测试集没有正确设置，或者成本函数没有正确设置。没有计算正确的值

**单数评价指标**

为了选择一个分类器，一个定义良好的开发集和一个评估度量将加速迭代过程。

例如:Cat vs非Cat

y=1, Cat图像检测



Precison（精度）

我们预测y=1的图像中，有多少是有猫的?

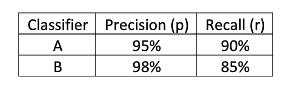


Recall（召回率）

在所有有猫的图片中，我们正确识别有猫的图片占了多少比例?



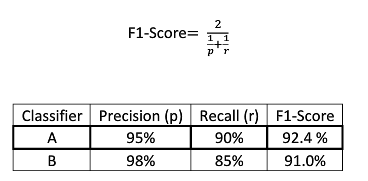
让我们比较两个分类器A和B，用来评估是否有cat图像:



在这种情况下，评估指标是精确度和召回率。

对于分类器A，有95%的几率图像中有一只猫，并且有90%的几率它正确地检测到了一只猫。而对于分类器B来说，图像中有98%的概率是猫，而它正确检测到猫的概率是85%

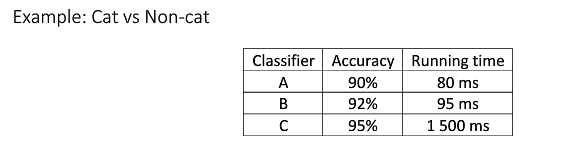
使用精确度/召回率作为评估指标的问题是您不确定哪一个是。更好的是，在这种情况下，两者都有很好的et召回精度。F1-score是一个调和平均值，它将精确度和召回率结合起来



分类器A是更好的选择。f1评分不是唯一可以使用的评价指标，例如，平均值也可以作为使用哪个分类器的指标

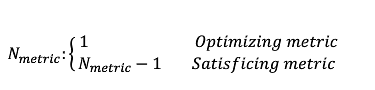
**满足和优化指标**

有不同的指标来评价分类器的性能，它们被称为评价矩阵。它们可以分为满足矩阵和优化矩阵。需要注意的是，这些评估矩阵必须在训练集、开发集或测试集上进行评估。



在这种情况下，精度和运行时间是评价矩阵。准确性是优化的指标，因为vou希望分类器能够尽可能准确地正确检测cat图像。在本例中设置为小于100ms的运行时间为满足度规，即满足度规必须满足期望集。

一般规则：



满足度量

优化度量

**训练/开发/测试集划分**

建立培训、开发和测试集对生产力有巨大的影响。从相同的分布中选择开发集和测试集是很重要的，必须从所有开发集和测试集中随机选取数据。

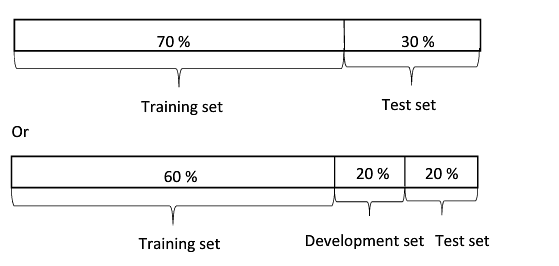
**指导方针**

选择一个开发集和测试集，以反映您希望在未来获得的数据，并考虑做好这些数据的重要性。

**开发和测试集的大小**

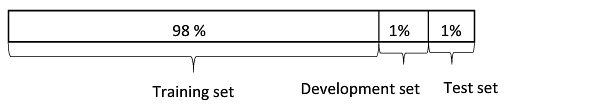
**旧的数据分割方法**

我们有更小的数据集，因此我们必须使用更大比例的数据来开发和测试想法和模型。



现代——大数据

现在，因为有大量的数据可用，我们不需要做太多的妥协，可以使用更大的一部分来训练模型



**指导方针**

设置测试集的大小，以对系统的总体性能给予高度的信心。

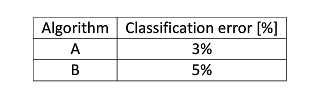
测试集可以帮助评估最终分类器的性能，最终分类器的性能可以小于整个分类器的30%数据集。

开发集必须足够大以评估不同的想法。

**1.7何时更改开发/测试集和度量标准**

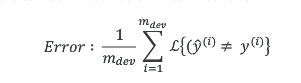
例子:猫vs非猫

猫的分类器试图找到大量的猫的图像显示给爱猫的用户。使用的评价指标是分类错误。



算法A似乎比算法B好，因为只有3%的错误，但是由于某种原因，算法A让很多色情图片通过。算法B有5%的错误，分类的图像较少，但没有色情图像。从公司的角度来看，以及从用户接受的角度来看。算法B实际上是一个更好的算法。评价指标未能正确排序算法之间的顺序偏好。应该更改评估指标或开发集或测试集。

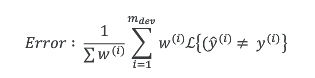
误分类误差度量可以写成如下函数:



这个函数计算错误分类的例子的数量。这个评估指标的问题在于，它对色情图片和非色情图片一视同仁。改变这个评价指标的方法是添加权重项w(i)。



函数变为：



指导方针

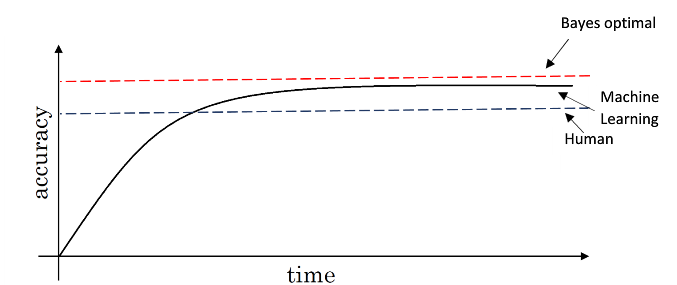
1. 正确定义有助于更好地排列顺序分类器的评估度量

2. 优化评价指标

**1.8为什么是人类的表现?**

今天，机器学习算法可以与人类水平的性能竞争，因为它们在很多应用中更高效、更可行。此外，设计和构建机器学习系统的工作流程比以前更加高效，而且，人类所做的一些任务接近“完美”，这就是机器学习试图模仿人类水平的性能的原因。

下图显示了人类和机器学习在一段时间内的表现。



当机器学习超过人类水平的表现时，它进展缓慢。其中一个原因是，人类水平的性能可以接近贝叶斯最优误差，尤其是自然感知问题。

贝叶斯最优误差定义为可能的最佳误差。换句话说，这意味着从x到y的任何函数映射都不能超过一定的精度。此外，当机器学习的性能不如人类时，您可以使用不同的工具来改进它。一旦其性能超过人类水平，它们就很难使用了。

这些工具是:

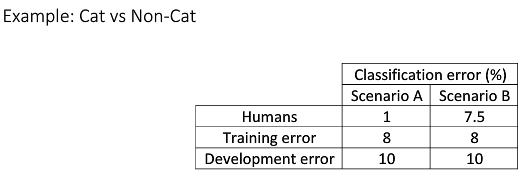
从人类那里获取标记的数据从人工错误分析中

获得洞察力:为什么一个人会做对这件事?

更好的偏见/方差分析

1.9可避免的偏差

通过了解人类水平的性能是什么，就有可能判断一个训练集的性能是否良好。



在本例中，由于人类善于识别图像，所以将人为级错误作为贝叶斯误差的代理。如果你想提高训练集的性能，但你不能做得比贝叶斯误差更好，否则训练集是过拟合的。通过了解贝叶斯误差，更容易关注偏差规避策略或方差规避策略是否会提高模型的性能

情形A

训练集的性能与人为水平误差之间有7%的差距。这意味着该算法与训练集的拟合不太好，因为目标约为1%。为了解决这个问题，我们使用了偏置减少技术，如训练一个更大的神经网络或运行更长的训练集

情形B

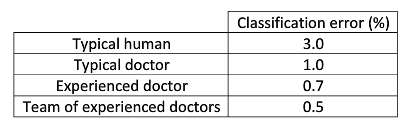
训练集做得很好，因为与人类水平误差只有0.5%的差异。训练集和人的水平误差之间的差异被称为可避免偏差。这里的重点是。减少训练误差与开发误差之差为2%的方差。为了解决这一问题，我们采用了正则化等方差约简技术或进行更大规模的训练集。

**理解人的表现**

人为误差给出了贝叶斯误差的估计。

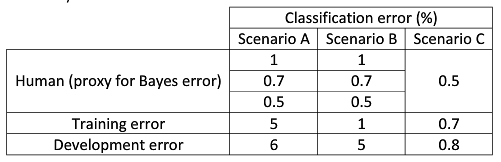
例1:医学图像分类

这是一个医学图像分类的例子，其中输入是放射学图像，输出是诊断分类决策。



人为误差的定义取决于分析的目的，在这种情况下，根据定义贝叶斯误差小于等于0.5%。

例2.误差分析



情形A

在这种情况下，选择人类级别的性能不会产生影响。可避免偏差介于4%-4.5%之间，方差为1%。因此，重点应放在偏置减小技术上。

情形B

在这种情况下，选择人类级别的性能不会产生影响。可避免的偏见介于0 -0。5%两者之间，方差是4%因此，重点应放在方差约简技术上。

情形C

在这种情况下，贝叶斯误差的估计值必须是0.5%，因为不能低于人类水平的性能，否则训练集就会过度拟合。可避免偏差是0。2%方差是0。1%因此，重点应放在偏置减小技术上

对人类水平表现的偏差/差异的总结

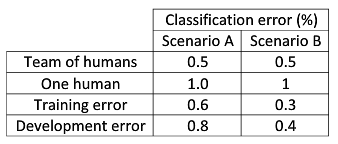
人为误差-代理贝叶斯误差

如果人为误差与训练误差的差值大于训练误差与开发误差的差值，重点应该放在偏压减小技术上

如果训练误差与开发误差的差值大于人为误差与训练误差的差值，重点应放在方差约简技术上

**超过人类水平的表现**

例1：分类任务



情形A在这种情况下，贝叶斯误差是0.5%，因此可避免偏差是01%，方差是0.2%。

情形B在这种情况下，没有足够的信息来知道是否需要对算法进行偏差减少或方差减少。这并不意味着模型不能改进，这意味着传统的方法知道是否偏差减少或方差减少在这种情况下不起作用。

在许多问题中，机器学习显著地超越了人类水平的性能，特别是在结构化数据方面:

在线广告

产品推荐

物流(预测运输时间)

贷款审批......

**改善模型表现**

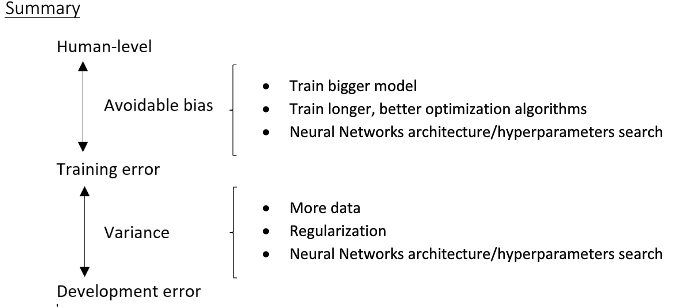
监督学习的两个基本假设监督学习

有两个基本假设。

第一个是低的可避免偏差。这意味着训练集非常适合。第二种是有一个低的或可接受的方差，这意味着训练集的性能可以很好地推广到开发集和测试集。

如果人的层次误差与训练误差的差值大于训练误差与开发误差的差值，那么重点应该放在偏置减小技术上，即训练更大的模型，训练更长时间，或者改变神经网络结构，或者尝试各种超参数搜索。

如果训练误差与开发误差的差值大于人的层次误差与训练误差的差值，则重点研究方差约简技术，即大数据集、正则化或改变神经网络结构或尝试各种超参数搜索。



**第二周**

**误差分析：**

并行评估多个想法

猫检测思路:

修正狗被认为是猫的照片。

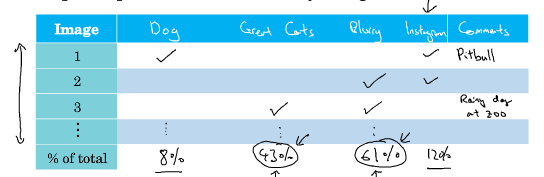
Comment

评论

解决大猫(狮子、美洲豹等)被错认

模糊

提高模糊图像的性能



还看了论文GAN