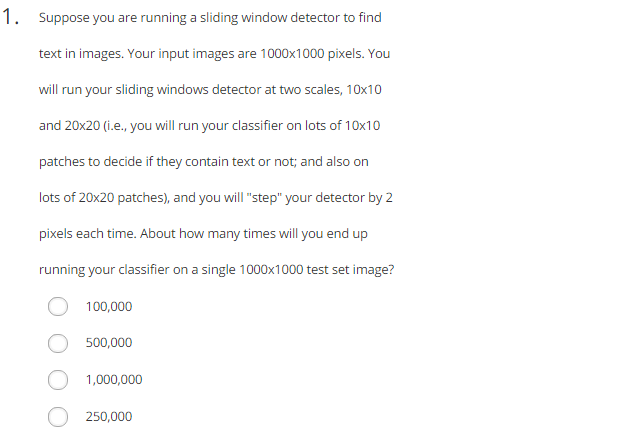
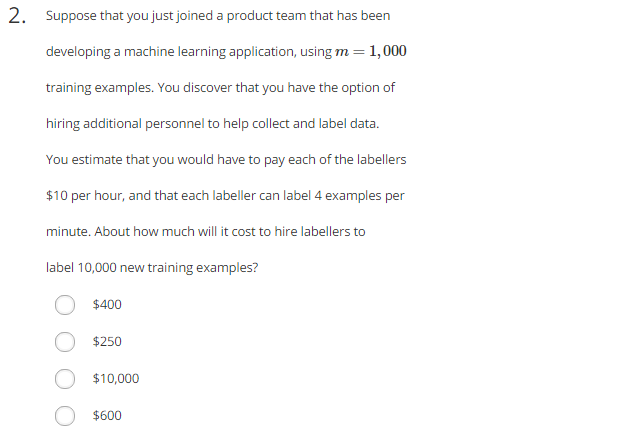
**Week 11:Application：Photo OCR课后习题解答**

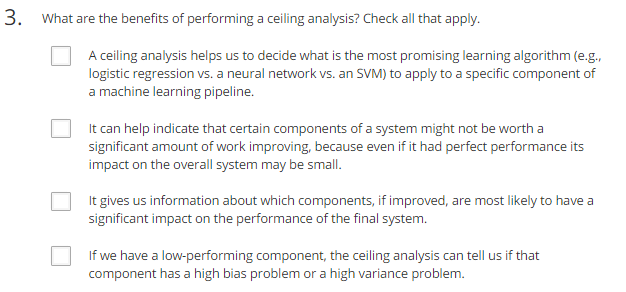
2016年03月20日 22:31:00

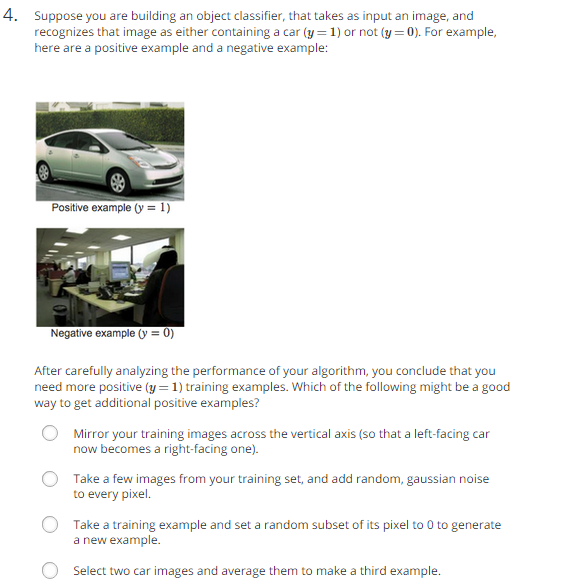
大家好，我是Mac Jiang,今天和大家分享的是Coursera-Stanford University-Machine Learning-Week 11:Application:Photo OCR的课后习题解答。虽然我的答案通过了系统测试，但是有些分析可能是错误的。各位博友如果发现错误，请留言纠正，谢谢。希望我的博客能给大家带来一些学习上的帮助！

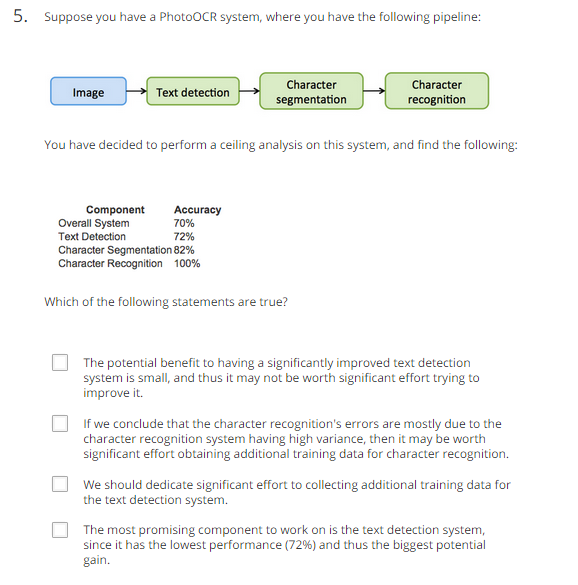
这周的课程中，老师主要讲了四个方面的内容：   
（1）照片光字符识别-Photo OCR。给定一张图片，要求识别图片中的字符，可以分为三步：字符信息定位，字符切割，字符识别。这三个部分形成一个机器学习流水线（pipeline），可以让不同开发小组进行不同的分工。对于字符信息定位和字符切割，提供了一种滑动窗的方法。   
（2）滑动窗（sliding windows）。滑动窗是对图片中物体定位的一种方法，他的思想是在图片中截取固定大小的块，导入到已经学习好的系统中，判断这块有没有需要定位的物体。当检测完当前块后，可以滑动窗口，进行下一块的检测。华东窗的大小可以变化，但比例一般不变，对于不同大小的滑动窗，只要把它压缩到我们学习算法所需要的图片大小就可以了。   
（3）人工数据合成（artificial data synthesis）。对于低偏差的机器学习算法，我们需要大量数据对系统进行训练，所以我们需要大量样本数据，这就可以用人工数据合成的方法。人工数据合成的方法可以分为两类，第一类：无任何起始数据，可以通过从网上下载不同数据库，加不同背景的方法实现；第二类：有少量数据，可以通过对已有数据进行扭曲，变换，加高斯随机噪声等方法得更多数据。   
（4）上限分析（Ceiling analysis）。当我们的系统表现不好的时候，我们想要改进，但是不知道流水线的哪一段的有较大的改进空间，可以用上限分析法进行分析。通过人工处理流水线每段的方法查看系统正确率的提升效果，提升最大的就是最有潜力改进的位置。上限分析的方法可以帮助寻找最优提升空间的位置，避免不必要的工作量。

1.第一题   
  
（1）题意：你在利用滑动窗法寻找图片中的文字位置，输入图片为1000\*1000像素，你在滑动窗10\*10和20\*20下运行滑动窗，你每次移动步长为2像素，则需要多少步？   
 1.100，000   
 2.500，000   
 3.1，000，000   
 4.250，000   
（2）分析：当滑动窗为10\*10时，每行需要滑动（1000-10）/2+1=496次，共有（1000-10）/2+1=496行，共约250，000次   
当滑动窗为20\*20时，每行需要华东（1000-20）/2+1=491次，共有（1000-20）/2+1=491行共约250，000次。共500，000次   
（3）答案：2

2.第二题   
  
（1）题意：假设你所在的产品开发小组在开发一个机器学习应用，有1000个训练样本。你必须雇佣工人进行样本分类，每个工人$10/hour，每人一分钟能分类4个样本。分类10，000个训练样本要多少钱?   
（2）分析：小学数学题，每个工人一小时可分240个样本，那么每个样本需要10/240美元，10，000个样本需要10，000\*10/240=416美元   
（3）答案：1

3.第三题   
  
（1）题意：上限分析的好处是什么，选出所有正确的。   
 1.上限分析可以帮助我们选取合适的机器学习算法（逻辑回归，神经网络，SVM），应用到机器学习流水线的特点部分。   
 2.可以考察流水线的某个部分是否花大量时间改进的价值，应为即使某个部分有完美的表现他对整个系统提高的作用是微乎其微的。   
 3.告诉我们提升流水线哪个部分将会对系统最后结果有显著提升。   
 4.如果流水线的某个部分表现糟糕，上限分析可以告诉我们他是高偏差的还是高方差的。   
（2）分析：上限分析的作用就是告诉我们流水线的哪个部分改进可以提升系统表现，避免我们盲目选择，做无用功。它不能帮助我们选择学习算法，也不能帮助我们判断问题是高偏差还是高方差。   
（3）答案：2,3

4.第四题   
  
（1）题意：我们建立一个物体分类器，输入一张图片，确定图片中是否有汽车（有y=1;无y=0）。当分析系统表现后，你发现需要更多y=1样本。下列哪个方法是得到新y=1样本的好方法。   
 1.把图形做镜面变化，颠倒左右   
 2.从训练样本中取一些图片，对每个像素加高斯随机噪声。   
 3.以一个训练实例，将其像素的一个随机子集设置为0，以生成一个新的例子。   
 4.选择两张汽车图片，平均它们为第三张图   
（2）答案：1

5.第五题   
  
（1）题意：选出正确描述   
 1.有一个显着的改进的文本检测过程的潜在好处是小的，因此，它可能不值得显着的努力，试图提高它。   
 2.如果我们得出这样的结论：字符识别的错误大多是由于字符识别系统具有很高的方差，那么它可能是值得显着的努力获得额外的训练数据的字符识别。   
 3.我们应该致力于收集额外训练样本用于提升文本检测   
 4.最有潜力的流水线位置为文本检测，应为他的表现为72%，最低   
（2）分析：1.正确，对文本检测改进后系统正确率由70-72，提升不大，因此不需要花大量时间。   
 2.正确，高方差可以通过增加训练样本的方法改进   
 3.错误，又没说文本检测是高方差的   
 4.错误，72%是系统正确率，对文本检测改进后系统正确率由70-72，是不值得提高的   
（3）答案：1，2