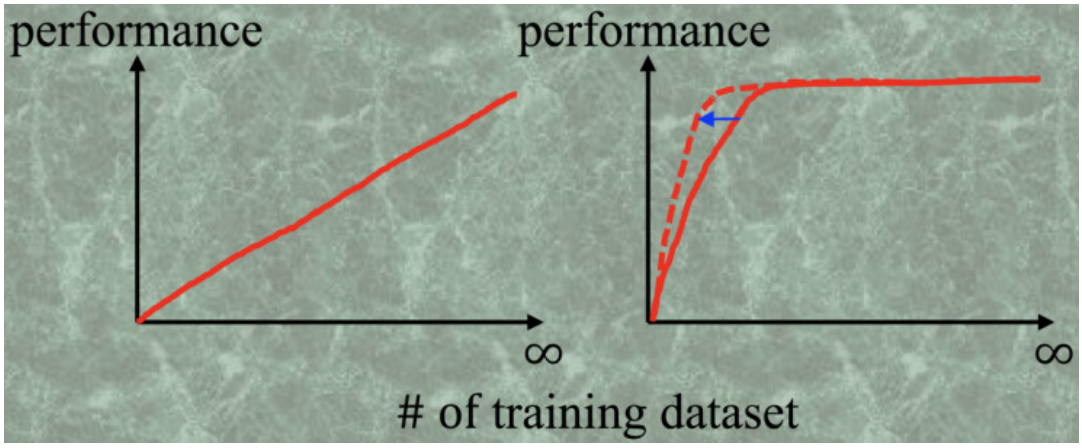
背景（为什么需要主动学习？何时需要主动学习？）

- 虽然说我们现在正处于大数据时代，每天都会产生海量的数据，而大数据正是近些年来，机器学习/深度学习能够迅速发展的原因。



- 但在某些领域，机器学习/深度学习往往会碰到一个大问题：**有标记的数据太少、标注成本太高，**特别是一些非常专业化的问题，比如说判断一个肿瘤的良恶性、小语种间相互的准确翻译、蚂蚁分类问题（世界上有一万多种蚂蚁）
  - 问题重点：标注难，而不是获取数据难
  - 对这些任务所需的大量数据样本进行标注，成本很高。
  - 怎么办：主动学习

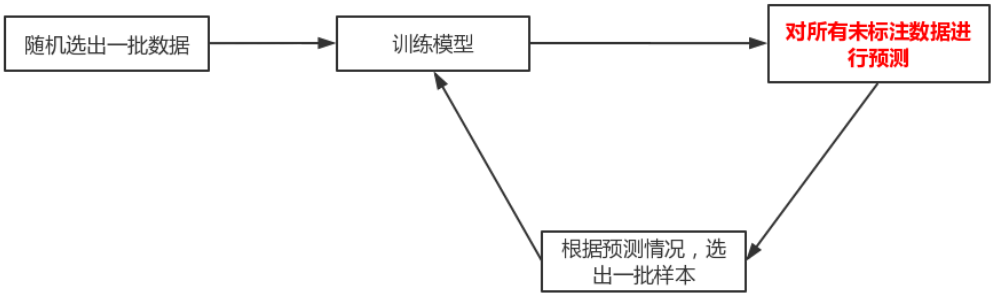
核心目标：

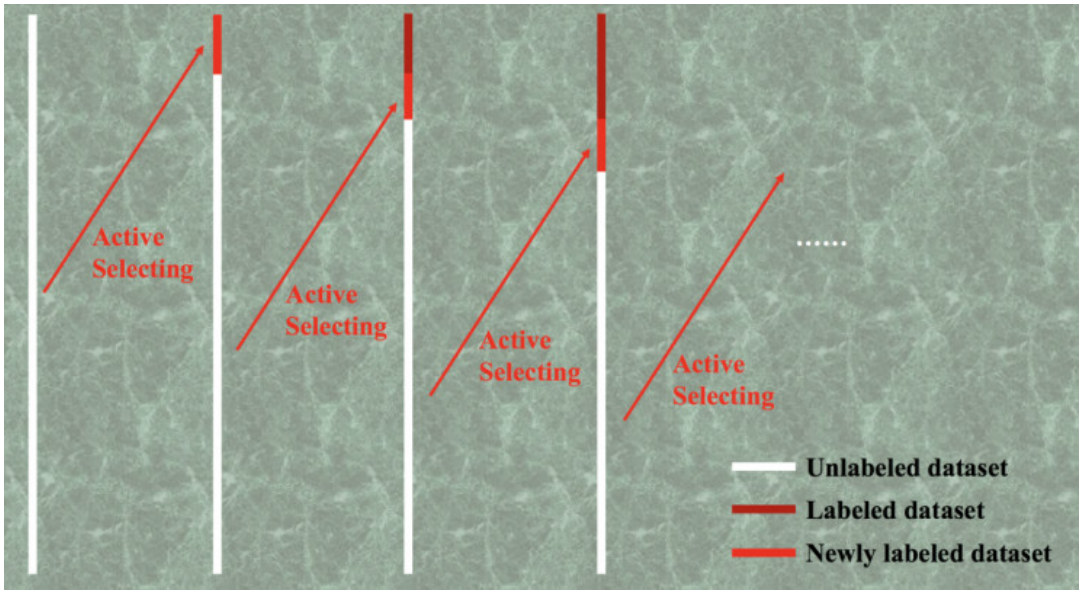
- 对更有价值的数据进行标注
- 用更少的标签数据来训练一个效果理想的分类器

怎么做

主动学习算法主要分为两阶段：

- 第一阶段为初始化阶段：随机从未标注样本中选取小部分样本，由督导者 (比如肿瘤问题中的专家) S标注，作为训练集建立初始分类器模型；
- 第二阶段为循环查询阶段，先让C对未标注样本集U进行预测，并以某种查询标准Q作为选择策略，选取一定的未标注样本让S进行标注，并加到训练样本集L 中，重新训练分类器，直至达到训练停止标准为止。





选择策略

什么是好的选择策略?

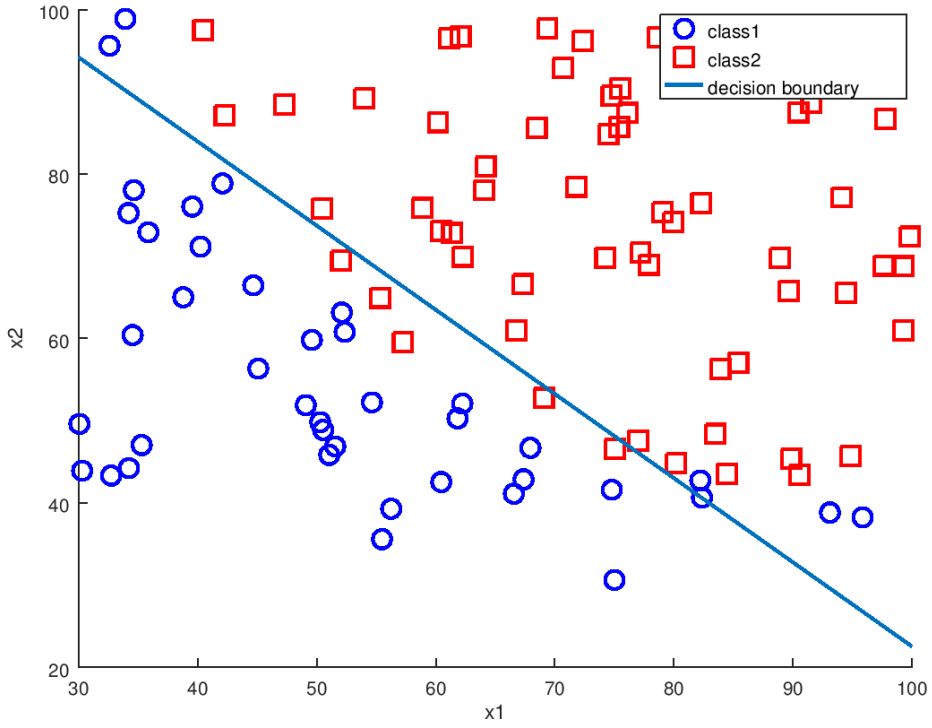
- 选出的样本更有可学习价值：与一般样本相比，在加入到训练集中之后，可以让模型更快地提升性能。

三种选择策略（不同类型的问题，适用不同的策略）：

1. 对样本做出预测的概率最低，Least Confident : argmax_x [1 - P(y^|x)]。
2. 选择在分类边缘的样本（能刻画/代表潜在分布/边缘） - Margin sampling : argmin_x [P(y1^|x) - P(y2^|x)]
3. 比较不确定性/信息量，即比较信息熵的大小。
4. Query-by-committee：同时训练多个模型来组成一个评委会，在选择样本时，让大家共同投票打分决定，不同模型给出的分值可以有不同的权重。

进一步深入两种策略

策略2 - 选择在分类边缘的样本，即下图中靠近分界线附近的样本点



策略3-比较不确定性/信息量：

熵与不确定性：

1. 不确定性和概率的关系：
    - 肯定发生 或 肯定不发生 -----> 十分确定
    - 而不确定的事 -----> 就是不确定它会不会发生
2. 概率和熵的关系

可以通过比较熵的大小来比较不确定性/信息量。

越不确定（0.5） ---> 熵就越大
越不确定（99.9% 或 0.1%） ---> 熵就越小

- 求信息熵的公式:

$$S_{\text{信息熵}} = -\sum_i p_i \log_2 p_i$$