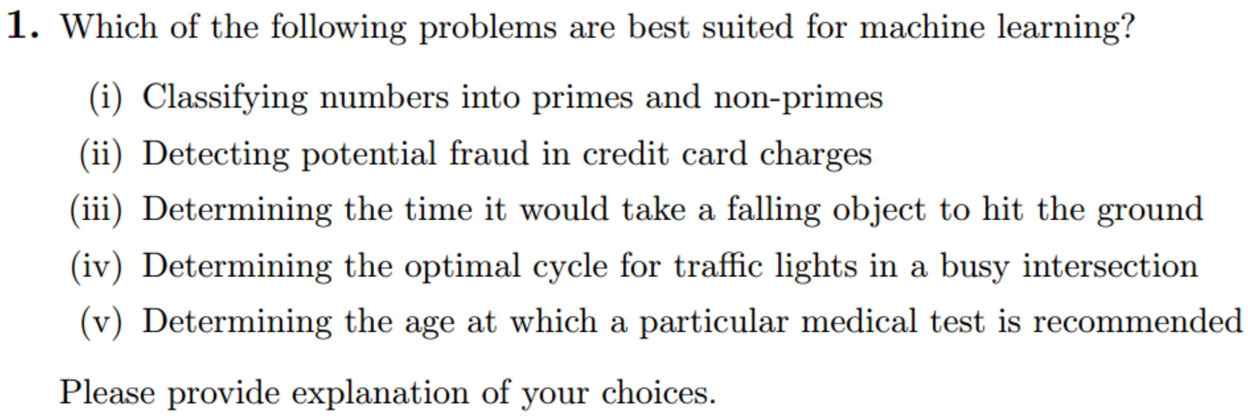
Machine-Learning Foundations hw1

Problem 1



1. Which of the following problems are best suited for machine learning?

(i) Classifying numbers into primes and non-primes

(ii) Detecting potential fraud in credit card charges

(iii) Determining the time it would take a falling object to hit the ground

(iv) Determining the optimal cycle for traffic lights in a busy intersection

(v) Determining the age at which a particular medical test is recommended

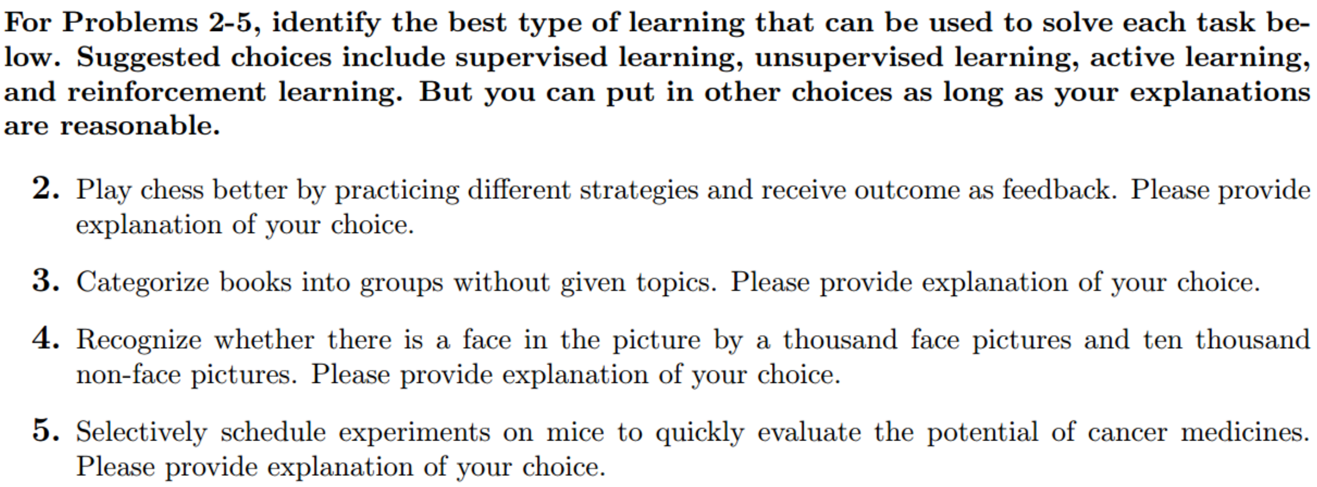
Please provide explanation of your choices.

> 适合使用机器学习的三个关键要素：
>
> - 存在某些潜在的模式或者规则可以学习
> - 没有具体的定义或者规则，来编写程序，或者不容易给出完备的定义或者规则
> - 有大量的可以学习潜在模式或者规则的数据用来学习

Solution 1

- (i) 判断一个数是否为素数。不需要使用机器学习的方法来做，因为已经有明确的规则可以进行判断：若一个整数只能被1和它本身整除，则为素数，否则不是素数；
- (ii) 检测信用卡消费中的潜在欺诈行为。属于使用机器学习中检测异常点检测问题；
- (iii) 确定物体下落所需的时间。不需要使用机器学习，因为有物理公式可以计算；
- (iv) 确定繁忙路口的交通信号灯的最佳周期。可以统计不同路口交通信号灯的周期和拥堵程度的数据，然后使用回归分析，学习信号灯周期与拥堵程度的之间的关系；
- (v) 确定建议进行特定医疗检查的年龄。可以对不同年龄人群生病情况进行统计，进行聚类分析，年龄相似患病相似的人群聚为一类，就可以向这个年龄附近的人推荐相应的医疗检查。

Problem 2-5



For Problems 2-5, identify the best type of learning that can be used to solve each task below. Suggested choices include supervised learning, unsupervised learning, active learning, and reinforcement learning. But you can put in other choices as long as your explanations are reasonable.

2. Play chess better by practicing different strategies and receive outcome as feedback. Please provide explanation of your choice.

3. Categorize books into groups without given topics. Please provide explanation of your choice.

4. Recognize whether there is a face in the picture by a thousand face pictures and ten thousand non-face pictures. Please provide explanation of your choice.

5. Selectively schedule experiments on mice to quickly evaluate the potential of cancer medicines. Please provide explanation of your choice.

> 机器学习分类:
>
> - **监督式学习**：可以从训练集中学到或建立一个模式（函数 / learning model），并依此模式推测新的实例。训练集是由一系列特征（通常是向量）和预期输出所组成。函数的输出可以是一个连续的值（称为回归分析），或是预测一个分类标签（称为分类）
>
> - **非监督式学习**：不需要人力来输入标签。它是监督式学习和强化学习等策略之外的一种选择。在监督式学习中，典型的任务是分类和回归分析，且需要人工预先做好标注。一个常见的非监督式学习是数据聚类。
>
> - **主动学习**：主动学习是半监督学习的一种特殊情况，其中学习算法能够交互地查询用户（或某些其他信息源），以在新的数据点获得所需的输出。
>
> - **强化学习**：强调如何基于环境而行动，以取得最大化的预期利益。其灵感来源于心理学中的行为主义理论，即有机体如何在环境给予的奖励或惩罚的刺激下，逐步形成对刺激的预期，产生能获得最大利益的习惯性行为。环境通常被规范为马可夫决策过程（MDP），所以许多强化学习算法在这种情况下使用动态规划技巧。强化学习和标准的监督式学习之间的区别在于，它并不需要出现正确的输入/输出对，也不需要精

> 确校正次优化的行为。强化学习更加专注于在线规划，需要在探索（在未知的领域）和遵从（现有知识）之间找到平衡。

Solution 2

采用不同的策略下棋，以结果作为回馈，学习更好的下棋策略，属于强化学习

Solution 3

不给定主题对书籍进行分类，属于无监督学习，可以统计书籍的关键词等作为特征，使用聚类分析

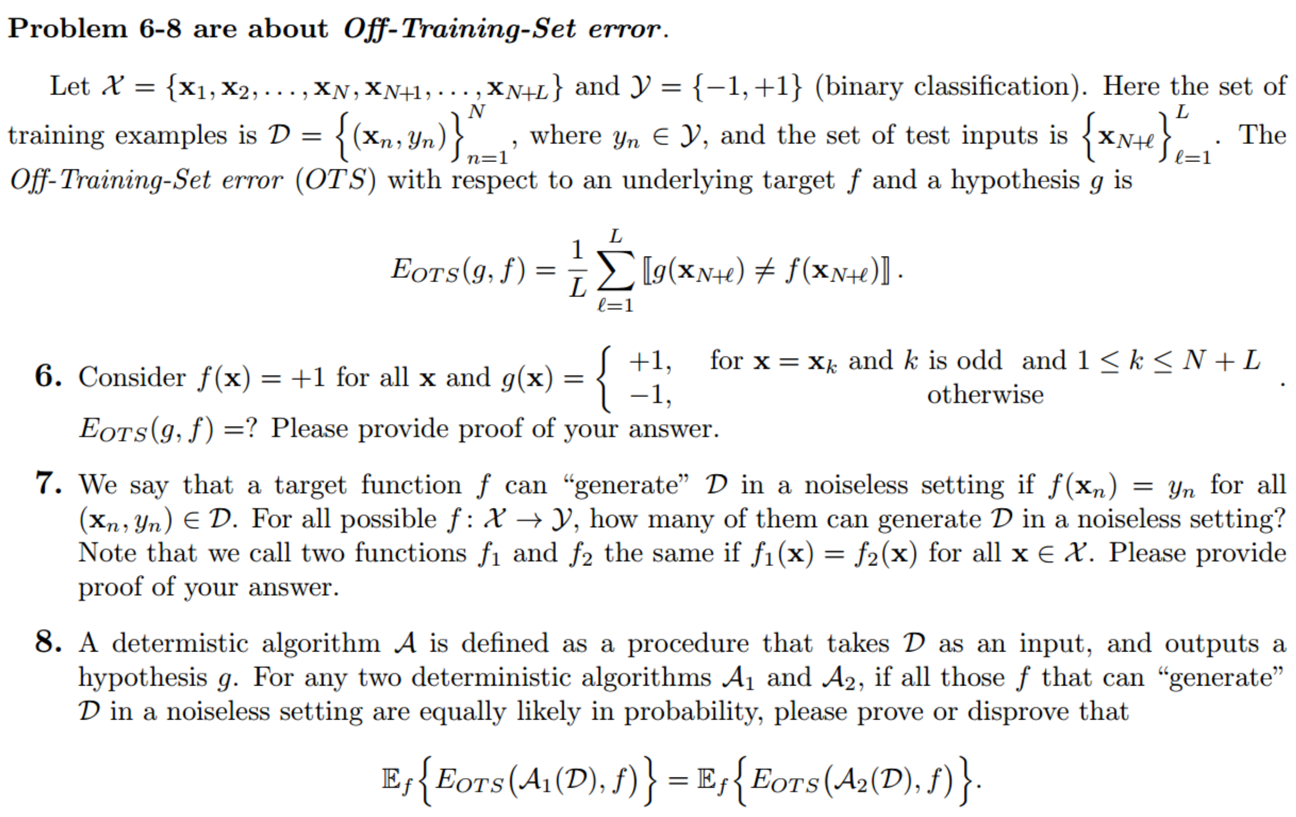
Solution 4

人脸识别，训练数据为1000张标注为 `+1` 的人脸图像和10000张标注为 `-1` 的非人脸图像，属于监督式学习

Solution 5

有选择性的安排实验，来快速的评估抗癌药物的效果，类似主动学习，对每一种药物，由于不知道实际效果，于是进行小白鼠实验，来获得输出

Problem 6-8



Problem 6-8 are about *Off-Training-Set error*.

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+L}\}$ and $\mathcal{Y} = \{-1, +1\}$ (binary classification). Here the set of training examples is $\mathcal{D} = \left\{(\mathbf{x}_n, y_n)\right\}_{n=1}^{N}$, where $y_n \in \mathcal{Y}$, and the set of test inputs is $\left\{\mathbf{x}_{N+\ell}\right\}_{\ell=1}^{L}$. The *Off-Training-Set error* (OTS) with respect to an underlying target f and a hypothesis g is

$$E_{OTS}(g, f) = \frac{1}{L}\sum_{\ell=1}^{L}[\![g(\mathbf{x}_{N+\ell}) \neq f(\mathbf{x}_{N+\ell})]\!]\,.$$

6. Consider $f(\mathbf{x}) = +1$ for all $\mathbf{x}$ and $g(\mathbf{x}) = \begin{cases} +1, & \text{for } \mathbf{x} = \mathbf{x}_k \text{ and } k \text{ is odd and } 1 \le k \le N+L \\ -1, & \text{otherwise} \end{cases}$.
$E_{OTS}(g, f) = ?$ Please provide proof of your answer.

7. We say that a target function f can "generate" $\mathcal{D}$ in a noiseless setting if $f(\mathbf{x}_n) = y_n$ for all $(\mathbf{x}_n, y_n) \in \mathcal{D}$. For all possible $f\colon \mathcal{X} \to \mathcal{Y}$, how many of them can generate $\mathcal{D}$ in a noiseless setting? Note that we call two functions f_1 and f_2 the same if $f_1(\mathbf{x}) = f_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Please provide proof of your answer.

8. A determistic algorithm $\mathcal{A}$ is defined as a procedure that takes $\mathcal{D}$ as an input, and outputs a hypothesis g. For any two deterministic algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$, if all those f that can "generate" $\mathcal{D}$ in a noiseless setting are equally likely in probability, please prove or disprove that

$$\mathbb{E}_f\left\{E_{OTS}(\mathcal{A}_1(\mathcal{D}), f)\right\} = \mathbb{E}_f\left\{E_{OTS}(\mathcal{A}_2(\mathcal{D}), f)\right\}.$$

Solution 6

f 对任何输入都输出 $+1$，g 对偶数下标的样本输出 -1，所以只需要判断 $N+1$ 到 $N+L$ 中有多少个偶数即可，可以分不同情况讨论：

- 若 L 为偶数，则肯定有 $\frac{L}{2}$ 个偶数；
- 若 L 为奇数，$N+1$ 也为奇数，则奇数占多数，故有 $\frac{L-1}{2}$ 个偶数；
- 若 L 为奇数，$N+1$ 为偶数，则偶数占多数，故有 $\frac{L+1}{2}$ 个偶数。

最后可整理为一个式子:

$$\frac{1}{L}\times(\lfloor\frac{N+L}{2}\rfloor-\lfloor\frac{N}{2}\rfloor)$$

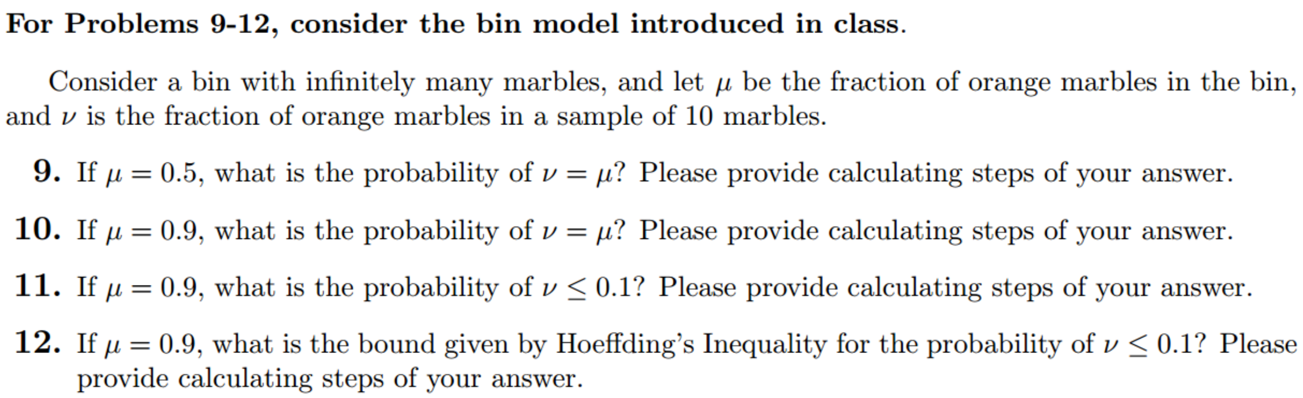
Solution 7

题目意思是 f 在训练集上完全正确，则有多少种不同的 f，由于测试集上的每一个样本都有两种可能输出，所以不同的 f 有 2^L 种

Solution 8

未完全理解

Problem 9-12



For Problems 9-12, consider the bin model introduced in class.

Consider a bin with infinitely many marbles, and let μ be the fraction of orange marbles in the bin, and ν is the fraction of orange marbles in a sample of 10 marbles.

9. If $\mu = 0.5$, what is the probability of $\nu = \mu$? Please provide calculating steps of your answer.

10. If $\mu = 0.9$, what is the probability of $\nu = \mu$? Please provide calculating steps of your answer.

11. If $\mu = 0.9$, what is the probability of $\nu \leq 0.1$? Please provide calculating steps of your answer.

12. If $\mu = 0.9$, what is the bound given by Hoeffding's Inequality for the probability of $\nu \leq 0.1$? Please provide calculating steps of your answer.

> 题意理解：从罐子中随机抓取一个小球为橘色的概率为 u，现从中随机抓取 10 个小球，则橘色小球个数为 $10v$ 的概率是多少

Solution 9

题意理解：$u = \frac{1}{2}$，则随机抓取的10个小球中有5个是橘色的概率：

$$\binom{10}{5} \times (\frac{1}{2})^5 \times (\frac{1}{2})^5 = \frac{63}{256}$$

Solution 10

题意理解：$u = \frac{9}{10}$，随机抓取的10个小球中有9个是橘色的概率：

$$\binom{10}{9} \times (\frac{9}{10})^9 \times (\frac{1}{10})^1 = (\frac{9}{10})^9$$

Solution 11

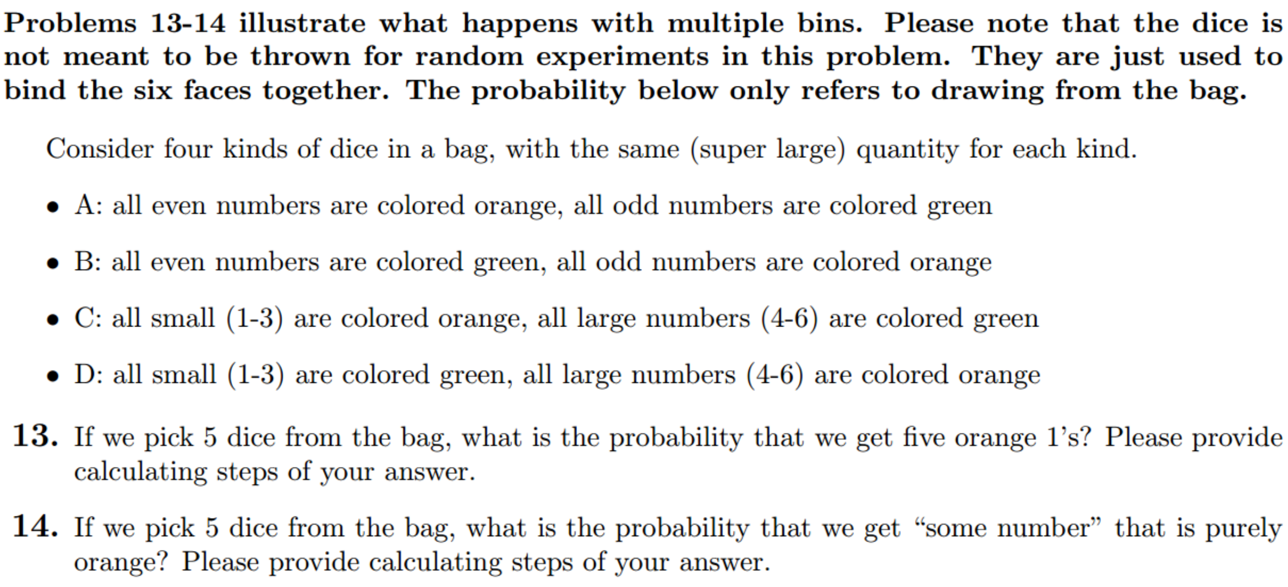
题意理解：$u=\frac{9}{10}$，随机抓取的10个小球中橘色小球个数不大于1的概率：

$$\binom{10}{1}\times(\frac{9}{10})^1\times(\frac{1}{10})^9+(\frac{1}{10})^{10}=\frac{91}{10^{10}}$$

Solution 12

由 $u=0.9, v\leq 0.1$ 可得 $|u-v|\geq 0.8$, 即$\epsilon=0.8$，由 $Hoeffding$ 不等式 $P(|u-v|\geq\epsilon)\leq 2e^{-2N\epsilon^2}$ ，带入计算得 5.5215×10^{-6}

Problem 13-14



Problems 13-14 illustrate what happens with multiple bins. Please note that the dice is not meant to be thrown for random experiments in this problem. They are just used to bind the six faces together. The probability below only refers to drawing from the bag.

Consider four kinds of dice in a bag, with the same (super large) quantity for each kind.

- A: all even numbers are colored orange, all odd numbers are colored green
- B: all even numbers are colored green, all odd numbers are colored orange
- C: all small (1-3) are colored orange, all large numbers (4-6) are colored green
- D: all small (1-3) are colored green, all large numbers (4-6) are colored orange

13. If we pick 5 dice from the bag, what is the probability that we get five orange 1's? Please provide calculating steps of your answer.

14. If we pick 5 dice from the bag, what is the probability that we get "some number" that is purely orange? Please provide calculating steps of your answer.

Solution 13

AC两种骰子的1是橘色的，因此随机抽取到1个骰子是橘色1的概率为 0.5，所以抽取到5个橘色1的概率为 $(\frac{1}{2})^5=\frac{1}{32}$

Solution 14

分析可知，AB两种骰子涂色方式完全相反，CD两种骰子也是完全相反，所以要想抽到的骰子中存在某一点数是都是同一个颜色，那么就不能有两种以上的骰子的组合。最简单的，只抽到一种骰子，那肯定符合题意，抽到两种骰子，那么不能是AB，也不能是CD，这样一来可能的组合为：A、B、C、D、AC、AD、BC、BD，它们的概率都可以求出来，前四种的概率都相等，为 $(\frac{1}{4})^5$，后四种的也都相等，为 $(\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4})\times(\frac{1}{4})^5$，所以最后的结果为：

$$4\times(\frac{1}{4})^5+4\times(\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4})\times(\frac{1}{4})^5=\frac{31}{64}$$

Problem 15-20

> 由于 Problem 15-20 都是 PLA 或 PLA-pocket 的实现，大部分代码都是可以共用的，因此放在同一个类 `PLA` (pla.py) 中。有如下一些方法：
>
> `read_file(fp)`
> *fp* 为文件存放路径，将文件中的除最后一列外的列读入 numpy 矩阵 matrix，并且为了方便计算，按照题目中的建议，在 matrix 的第一列之前添加了一列 1；将文件最后一列读入一个 numpy 数组 labels，最后返回 matrix 和 labels
>
> `sign(x)`
> x 小于等于0时返回-1，大于0时返回+1
>
> `train(train_X, train_Y, random=False, eta=1, silent=True)`
> 输入由 *read_file* 方法读取到的 X 和 Y，PLA 算法迭代更新参数 w，直至全部分类正确，保存结果 w，迭代轮数 num_of_updates，和最后一次分类错误的样本的下标 index_of_last_mistake
>
> `calc_err_rate(w, X, Y)`
> 以参数 w 计算数据集 (X Y) 的错误率

> `train_pocket(train_X, train_Y, iteration=50, silent=True)`
> 用 pocket 算法迭代 iteration 轮，保存直接迭代的结果 w，和装袋的结果 w_pocket
>
> `verify(test_X, test_Y, pocket=True)`
> 由 *train_pocket* 方法训练得到参数 w 和 w_pocket，验证其在测试集上的表现。pocket=True 时用 w_pocket 计算错误率；pocket=False 时用 w 计算错误率

以下实际运行的代码和图例在 'pla.ipynb' 文件中

Solution 15

直接使用默认参数调用 *train* 方法

Solution 16

调用 *train* 方法，设置参数 'random=True'

Solution 17

调用 *train* 方法，设置参数 'random=True, eta=0.5'

Solution 18

调用 *train_pocket* 和 *verify_pocket* 方法 2000 次，统计平均错误率

Solution 19

调用 *train_pocket* 和 设置了 'pocket=False' 的 *verify_pocket* 方法 2000 次，统计平均错误率

Solution 20

调用 设置了 'iteration=100' 的 *train_pocket* 方法 和 *verify_pocket* 方法 2000 次，统计平均错误率

【注：如有错误，望不吝指正，欢迎交流学习。邮箱：3051266672@qq.com】

参考

- http://blog.csdn.net/a1015553840/article/details/50986313
- http://blog.csdn.net/a1015553840/article/details/50979434
- http://blog.csdn.net/a1015553840/article/details/50979640
- https://zh.wikipedia.org/wiki/%E5%BC%BA%E5%8C%96%E5%AD%A6%E4%B9%A0
- https://en.wikipedia.org/wiki/Active_learning_(machine_learning)