李航统计学习：

第二章讨论部分：

读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：

问题1:p37页说损失函数的自然选择是误分类点的总数，但是为什么说这样的损失函数不是参数w，b的连续可导函数？

讨论后的理解：

损失函数自然选择很好的度量条件就是误分类点的总个数，但是误分类点的总 个数很可能是离散的，比如说对于相同的w，b+1和b是两种完全不同的总数，那么这样的损失 函数必定不是连续可导函数。

1. 提出的问题2：

p44页，为什么说当n=1时，表示第i个实例点由于误分而进行更新的次数？

讨论后的理解：

因为afa=ni\*n，所以当n等于1的时候afa=ni的叠加综合，所以当n=1时，表示第i个实例点由于误分而进行更新的次数。实力点更新的次数越多，意味着它距离分离超平面越近，也就越难正确分类。

二、（必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：

问题3：感知机学习中，在定义损失函数时，为什么可以直接忽略w的L2范数的倒数从而得到该损失函数？

自己的理解：

泛化误差是检验训练后的模型具不具备代表性，用偏差（bias）和方差（variance）来描述。如果一种方法学习的模型比另外一种方法学习的模型拥有更小的泛化误差上界，那么这种方法就更有效。泛化误差上界是样本容量的函数，当样本容量增加时，泛化误差上界趋于0；泛化误差上界也是假设空间容量的函数，假设空间容量越大，模型越难学，泛化误差上界就越大。

问题4：感知机对偶形式相对于普通形式的优点？

自己的理解：

为什么使用对偶形式：本质上来说对偶形式与原始形式在模型定义和参数更新上没有区别，唯一的不同就是两种方法计算参数的方式不同。如果设  为模型递归总次数，则在原始形式下，模型在参数优化过程中总共需要进行  次向量数乘和  次向量加法，在模型预测时需要进行1次向量点积；而在对偶形式下，模型在参数优化过程中总共需要进行  次数值加法，在模型预测时需要进行  次向量点积和  次数值加法，其中  为训练数据个数。

这样看来，对偶形式在参数优化过程中所需的计算资源远远小于原始形式，但预测时消耗的计算资源很大。为了减少对偶形式在预测时所消耗的计算资源，我们可以在模型训练之前就将所有训练数据间的点积计算好，生成一个矩阵，这样在对偶形式进行预测时便可以直接按照索引在矩阵中找寻对应的点积结果，这样一来对偶形式无论在参数优化还是模型预测时，其所需的计算资源都要比原始形式要小。

问题5：感知机的普通形式和对偶形式有什么区别？换句话说为什么要有对偶模式？我感觉对偶形式不就是批次训练么？

自己的理解：

根据查阅资料，普遍的观点是以下俩点：

1.从对偶形式学习算法过程可以看出，样本点的特征向量以内积的形式存在于感知机对偶形式的训练算法中，凡是涉及到矩阵，向量内积的运算量就非常大（现实中特征维度很高），这里我们如果事先计算好所有的内积，存储于Gram矩阵中，以后碰到更新的点，直接从Gram矩阵中查找即可，相当于我就初始化运算一遍Gram矩阵，以后都是查询，大大加快了计算速度。

1. 跟SVM的对偶形式其实有类似之处（后面讲到SVM的时候再说明）

问题6：感知机与SVM的区别是什么？

自己的理解：

1、相同点

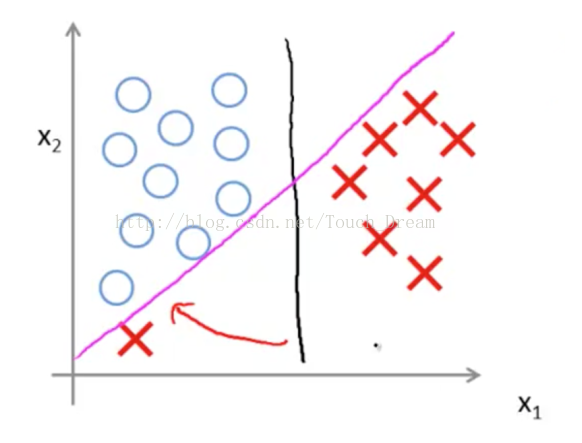
都是属于监督学习的一种分类器（决策函数）。

2、不同点

感知机追求最大程度正确划分，最小化错误，效果类似紫线，很容易造成过拟合。

支持向量机追求大致正确分类的同时，一定程度上避免过拟合，效果类似下图中的黑线。

感知机使用的学习策略是梯度下降法，而SVM采用的是由约束条件构造拉格朗日函数，然后求偏导令其为0求得极值点。这里特别说明下一般我们的拉格朗日函数是符合凸函数的，因此对于凸函数一定存在极值点，也是唯一的最优解。而一般的非凸函数，只好采用梯度下降法一步一步的求得极值点，如果非凸函数还是采用求导令为0，可能找不到极值点！因为鞍点也是导数为，但却不是极值点的特例，如y = x^3函数。导数为0是函数极值点的必要条件。

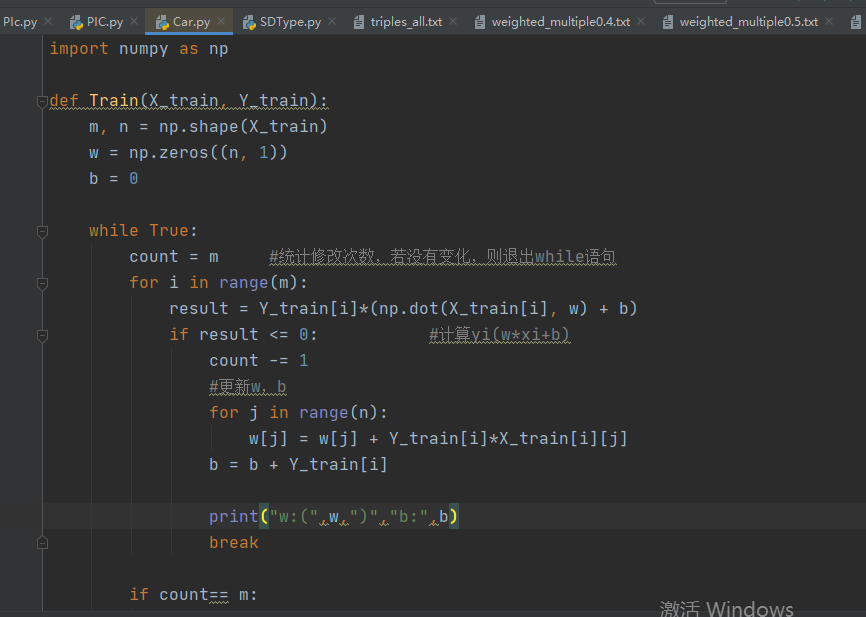


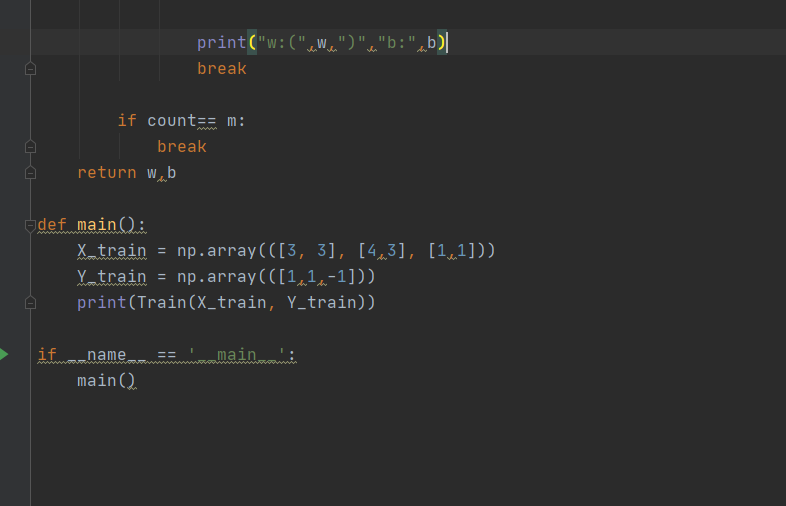
1. （必填）读书计划
2. 本周完成的内容章节：

（1）第二章完成并且梳理知识点，寻找问题，自己思考并且于小组会之前完成了自己的思考。

2、下周计划：第三章的阅读.

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）





数据结果：

