感知机算法与代码实现

> 统计学习方法感知机章节笔记

> 感知机是二分类模型，输入为特征向量，输出为实例的类别，取+1和-1二值。感知机学习旨在求出将训练数据进行线性划分的分离超平面，为此，导入基于误分类的损失函数，利用梯度下降法对损失函数进行极小化，求得感知机模型。感知机学习算法分为原始形式和对偶形式。

1. 感知机模型

定义： 假设输入空间$X \subseteq \mathbb{R}^n$，输出空间$Y=\{+1,-1\}$。输入x表示实例的特征向量，对应于输入空间的点，输出y表示实例的类别。由输入空间到输出空间的如下函数：

$$f(x)=sign(w \cdot x+b)$$

称为感知机。其中，w和b为模型参数。感知机模型的假设空间是定义在特征空间中的所有线性分类模型或线性分类器，即函数集合$\{f|f(x)=w \cdot x+b\}$

2. 感知机学习策略

假设训练数据集是线性可分的，感知机学习的目标是求得一个能够将训练集正实例点和负实例点完全正确分开的分离超平面。为了找出这样的超平面，即确定感知机模型参数w,b，需要确定一个学习策略，即定义(经验)损失函数并将损失函数极小化。
损失函数一个自然的选择是误分类点的总数。 但是，这样损失函数不是参数w,b的连续可导函数，不易优化。*损失函数的另一个选择是误分类点到超平面的总距离，这是感知机所采用的。* 为此:
首先，写出输入空间中任一点x_0到超平面的距离：

$$\frac{1}{||w||}|w \cdot x_0+b|$$

其次，对于误分类的数据(x_i,y_i)来说，

$$-y_i(w \cdot x_i+b)>0$$

成立。因为当$w \cdot x_i + b > 0$时，$y_i = -1$，而当$w \cdot x_i + b < 0$时，$y_i = +1$。因此，误分类点x_i到超平面的距离是：

$$-\frac{1}{||w||} y_i (w \cdot x_0 + b)$$

不考虑$\frac{1}{||w||}$，就得到感知机学习的损失函数。
给定训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

感知机$sign(w \cdot x + b)$学习的损失函数定义为

$$L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$$

其中，M是误分类点的集合。这个损失函数就是感知机学习的经验风险函数。

3. 感知机学习算法

3.1. 感知机学习算法的原始形式

感知机原始形式的算法实现

感知机学习算法是对一下最优化问题的算法。给定一个训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

求参数w,b，使其为一下损失函数极小化问题的解：

$$min_{w,b} L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$$

其中M为误分类点的集合。
感知机学习算法是误分类驱动的，具体采用随机梯度下降法。首先，任意选取一个超平面w_0, b_0，然后用梯度下降法不断地极小化目标函数。极小化过程中不是一次使M中所有误分类点的梯度下降，而是一次随机选取一个误分类点使其梯度下降。
假设误分类点集合M是固定的，那么损失函数L(w,b)的梯度由

$$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$$
$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$

给出。
随机选取一个误分类点(x_i, y_i)，对w,b进行更新：

$$w \leftarrow w + \eta y_i x_i$$
$$b \leftarrow b + \eta y_i$$

式中]η是步长，在统计学习中又称为学习率。这样，通过迭代可以期待损失函数$L(w,b)$不断减小，直到为0.综上，得到以下算法：

> 输入：训练数据集$T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$，学习率$\eta$;
> 输出：w,b；感知机模型$f(x) = sign(w \cdot x + b)$
> (1) 选取处置w_0, b_0
> (2) 在训练集中选取数据(x_i, y_i)
> (3) 如果$y_i(w \cdot x_i + b) \leq 0$
>
> $$w \leftarrow w + \eta y_i x_i$$
> $$b \leftarrow b + \eta y_i$$
>
> (4) 转至(2)，直至训练集中没有误分类点

直观上这种算法解释如下：当一个实例点被误分类，即位于分类超平面的错误一侧时，则调整w,b的值，使分离超平面向该误分类点的一侧移动，以减少该误分类点与超平面的距离，直至超平面越过该误分类点使其被正确分类。

3.2. 算法的收敛性