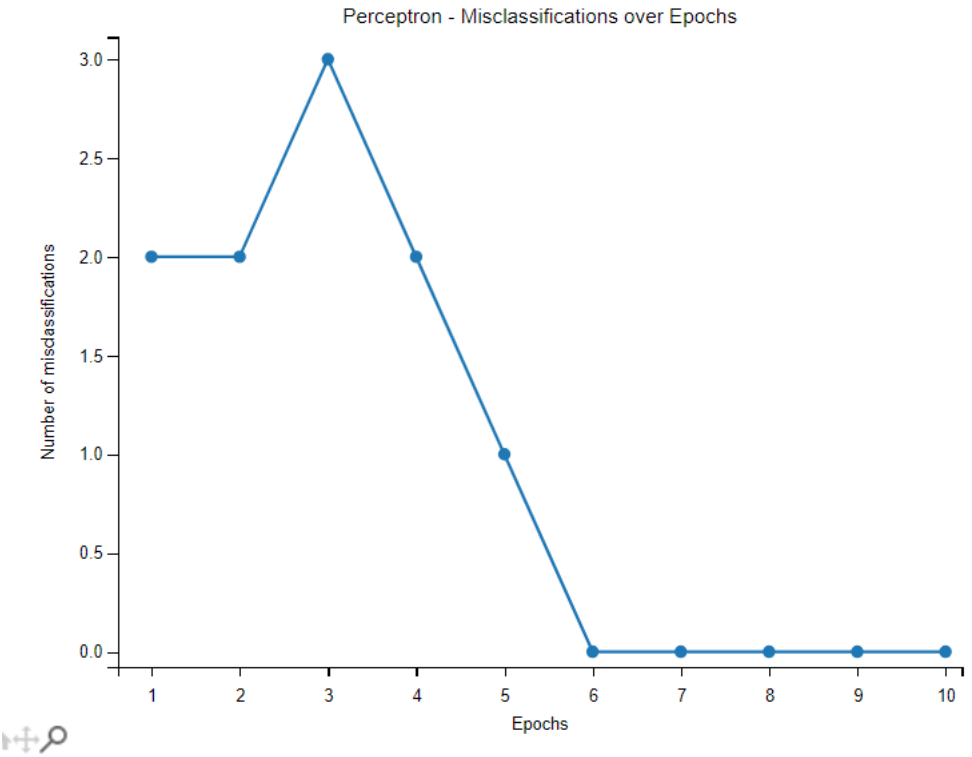
practice on perception



上图为使用用 Python 实现感知器训练算法，并在鸢尾花数据集（仅用两类）上进行二分类后感知器训练过程中每轮（epoch）的误分类数量变化。

X 轴（Epochs）：训练的轮数。每个 epoch 表示模型完整遍历了训练数据一次。

Y 轴（Number of misclassifications）：每个 epoch 中被错误分类的数据点数量。

曲线变化趋势：

- 初始时误分类数量较高；
- 随着训练进行，误分类数量逐渐减少；
- 最终达到某个较低的稳定值，甚至可能变为 0（即完全分类正确）。

说明感知器训练成功的迹象：

- 图像中误分类数量**呈下降趋势**，说明感知器在不断学习并改进分类能力。
- 如果最后几轮误分类数为 0，说明数据被**完全线性分割**，即模型已收敛。

选取前两类数据原因以及感知器局限性

鸢尾花（Iris）数据集有三类：Setosa、Versicolor、Virginica。我们只选取前两类（Setosa 和 Versicolor），因为它**们是线性可分**的，这使得感知器可以正常收敛。

原因：Setosa 和 Versicolor 是线性可分的；而 Versicolor 和 Virginica 有重叠，感知器将无法正确收敛。

局限性：

1. 感知器仅适用于**线性可分数据**；
2. 对噪声敏感；
3. 学习率和迭代次数难调；
4. 不支持概率输出（不像逻辑回归）；
5. 对特征缩放敏感。

若学习率 η 过大（如 0.5），训练过程会如何变化？

现象：训练过程可能出现**震荡**、**不收敛**，甚至不断跳过最优解。

原因：每次更新的步长太大，可能导致错过收敛点，像“跳过山谷”。

为什么感知器无法解决异或（XOR）问题？

1. 数学解释：

XOR 的输入输出如下：

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0

- 无法通过一条直线将 y=1 和 y=0 分开 → 线性不可分。

2. 几何视角解释：

- 在二维平面中，(0,1) 和 (1,0) 需要在同一侧，而 (0,0) 和 (1,1) 在另一侧。

- 任意一条直线都无法满足这一条件。

解决方法：

使用多层感知器（MLP）或非线性变换特征空间（如核方法），才能拟合 XOR。

感知器做垃圾邮件分类（词频向量）会遇到哪些挑战？

挑战包括：

1. **数据稀疏维度高**：
    - 词频向量通常是高维稀疏向量；
    - 感知器不具备自动特征选择能力。
2. **非线性关系处理能力差**：
    - 垃圾邮件的判定可能涉及多个词联合出现（例如"free" + "win"）；
    - 感知器无法学习非线性特征交互。
3. **类不平衡问题**：
    - 垃圾邮件占比可能远小于正常邮件，导致偏向多数类。
4. **过拟合与泛化问题**：
    - 特征选择不当或词频波动大，会导致模型不稳定。
5. **缺乏概率输出或置信度**：
    - 感知器只能输出 +1 或 -1，不能衡量“多大可能性是垃圾邮件”。