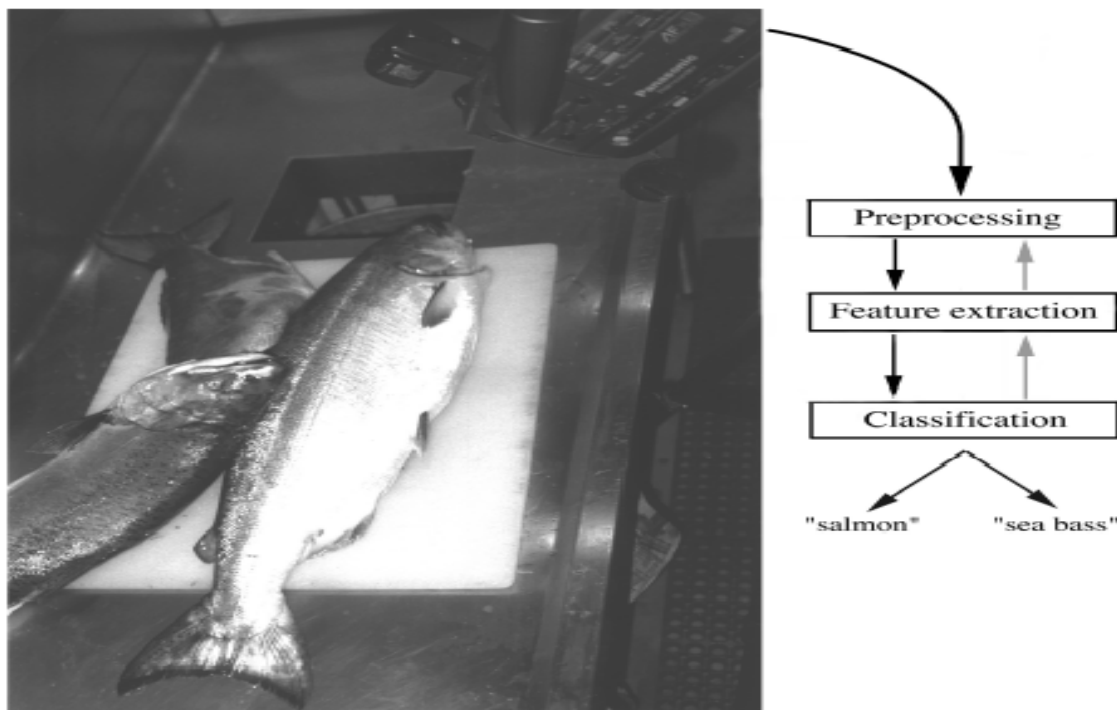


1.3 一个典型的机器学习案例 —— 对鱼进行分类

CSDN学院

► 例：鱼的分类

- 根据一些光学传感器对传送带上的鱼进行分类



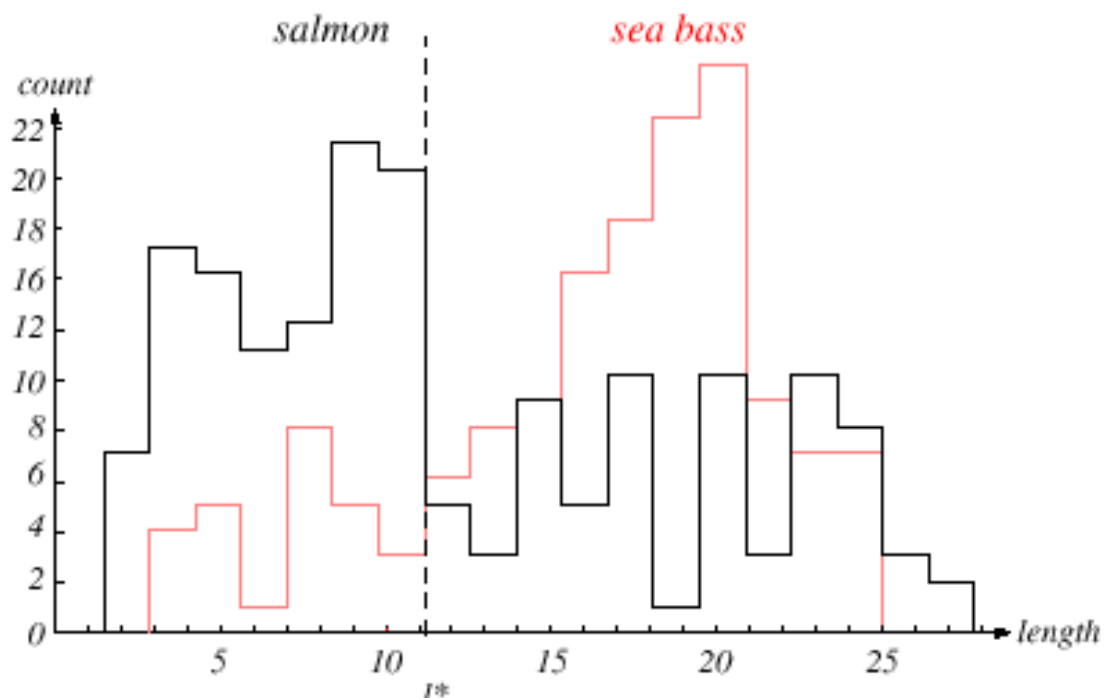
► 形式化为机器学习问题

- 训练数据 : $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$
 - 每条鱼的测量向量 (**特征**) : \mathbf{x}_i (如重量、长度、颜色)
 - 每条鱼的**标签** y_i (如三文鱼/salmon、鲈鱼/sea bass)



- 测试 :
 - 给定一个新的特征向量 \mathbf{x}
 - **预测**对应的标签 y

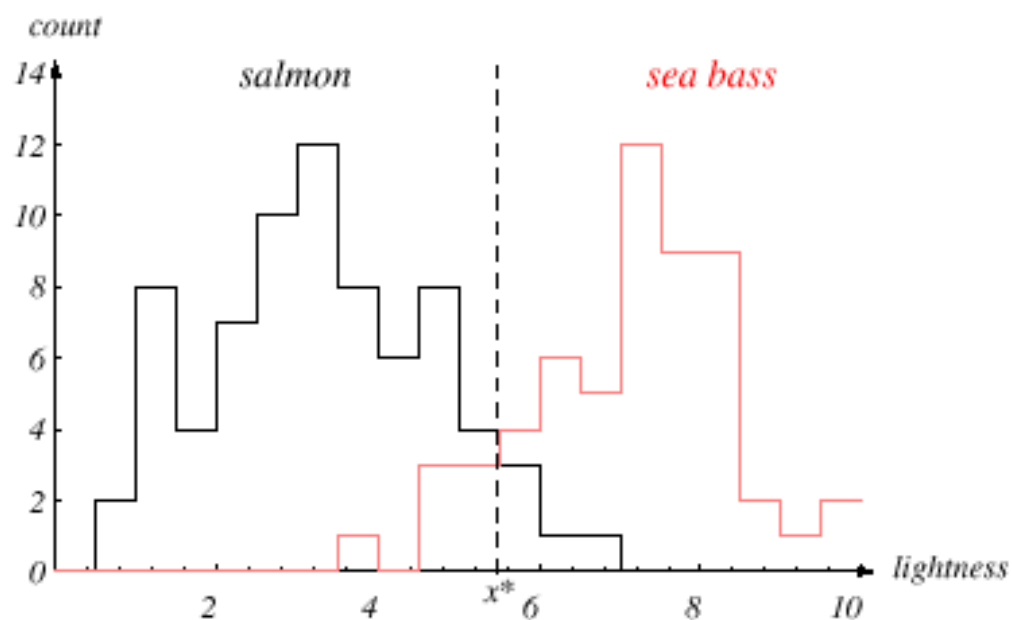
► 将长度作为特征进行分类



- 需要先则一个**决策边界**
 - **最小化平均损失**

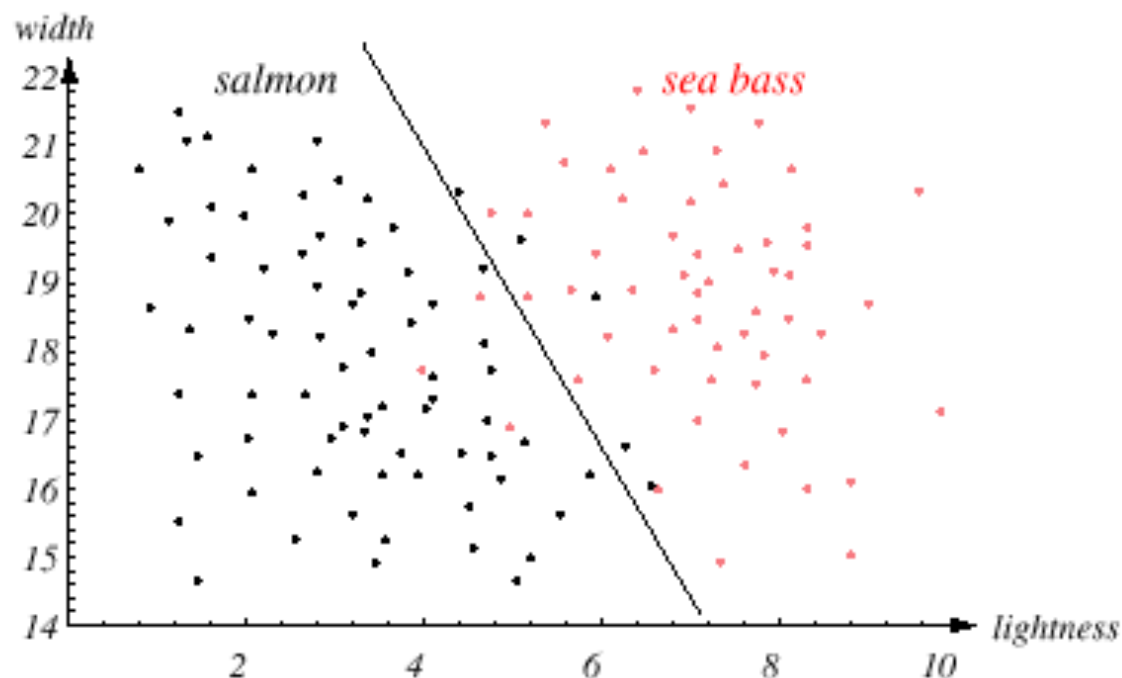
训练误差：90 / 316 = 28%

► 将亮度作为特征进行分类



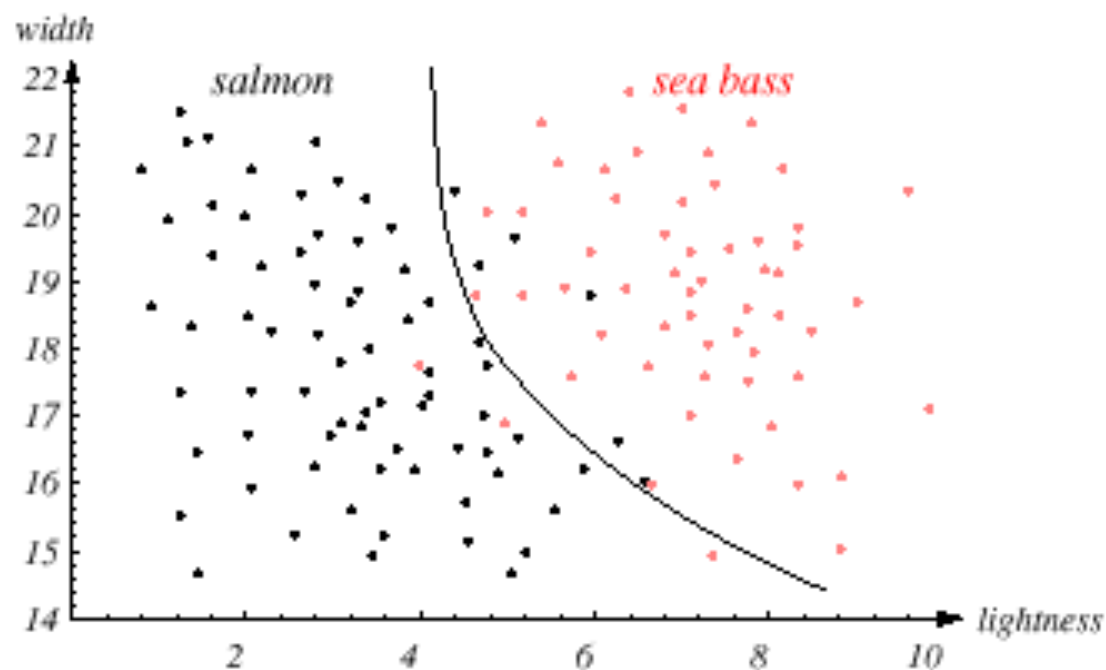
训练误差：16 / 316 = 5%

► 长度和亮度一起作为特征



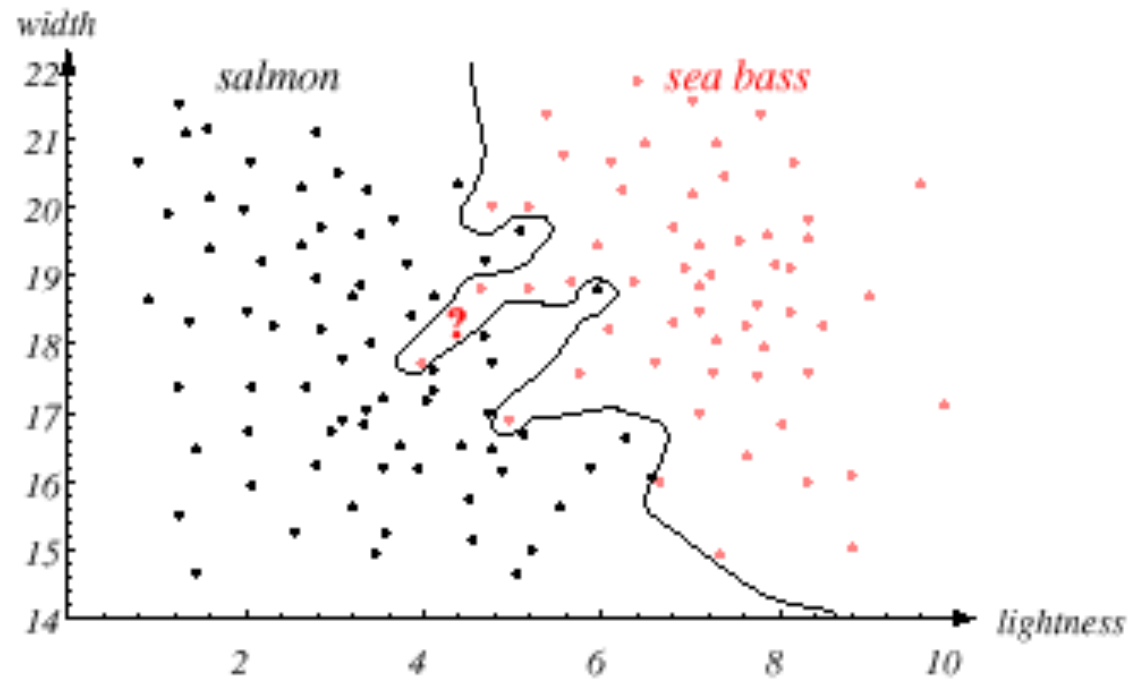
- 线性决策函数

► 更复杂的决策边界



- 二次决策边界函数

► 更复杂的决策边界...



训练误差 : $0 / 316 = 0\%$

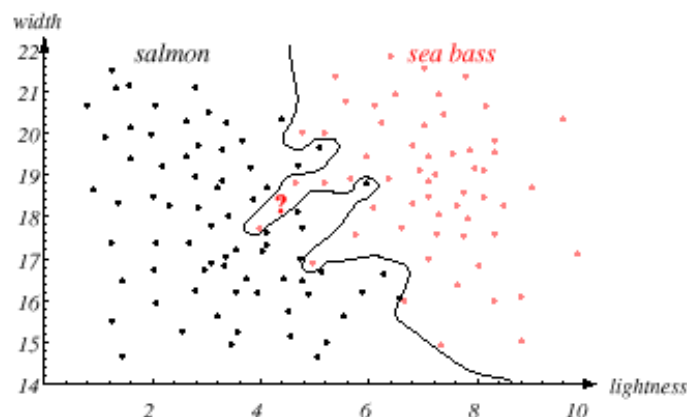
“ideal” classifier”? Is this good?

训练集上的误差 \neq 测试集上的误差

- 分类器应该在新数据上表现好
- 上述“ideal”分类器在新数据集上的错误率：25%

► What's Wrong?

- 推广性(*generalization*)差



- 复杂的决策边界不能泛化/推广到新数据上，根据特定调制得太好，而不是真正将salmon 和sea bass 分开的模型
 - 被称为数据过拟合(*overfitting*)

► 小结: 设计一个鱼分类器

- 选择特征
 - 可能是**最重要的**步骤! (收集训练数据)
- 选择模型 (如决策边界的形状)
- 根据训练数据估计模型
- 利用模型对新样本进行分类