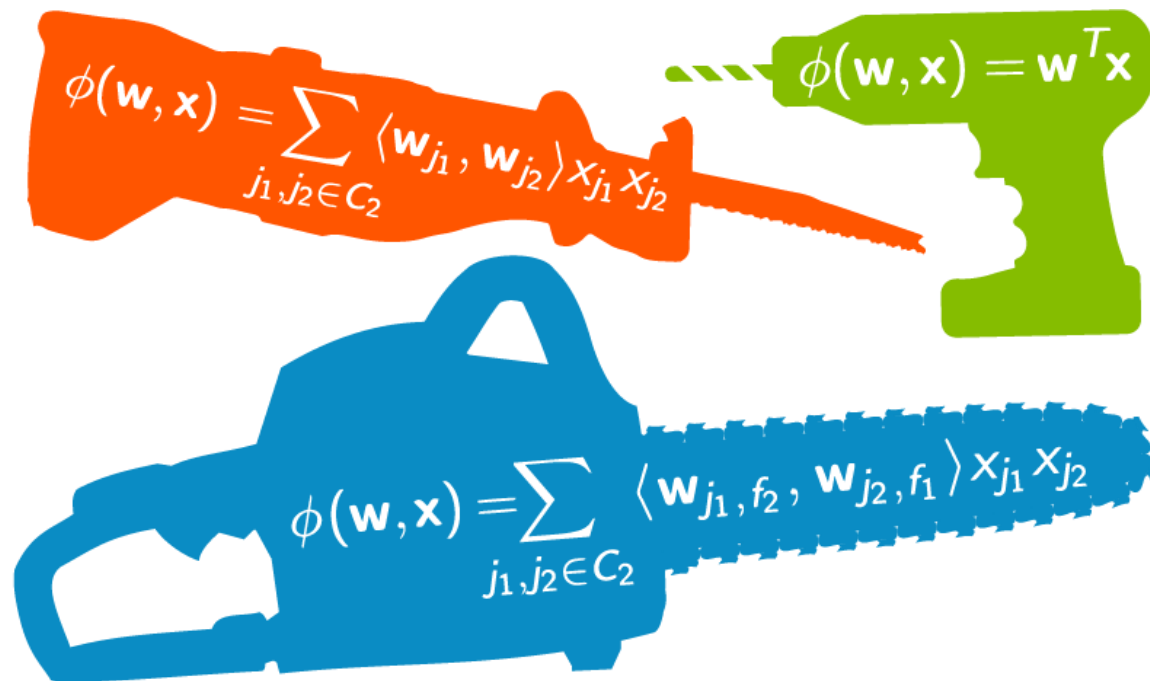


法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。



关注 **小象学院**



特征工程

--Robin

特征预处理

- 特征类别
 - 数值型特征，如：长度、宽度、像素值等
 - 有序型特征，如：等级（A，B，C）；级别（低、中、高）
 - 类别型特征，如：性别（男、女）
- 数值型特征
 - 可直接使用
 - 但是，对于有些模型来说，**数值归一化**（feature normalization）可以提高模型的性能，如：线性回归，kNN，SVM，神经网络等
 - 把不同量纲的东西放在**同一量纲**下比较，即把不同来源的数据统一到一个参考系下，这样比较起来才有意义。

特征预处理

- 数值型特征(续)

- 范围归一化：将所有特征数据按比例缩放到0-1区间（或者-1到1）

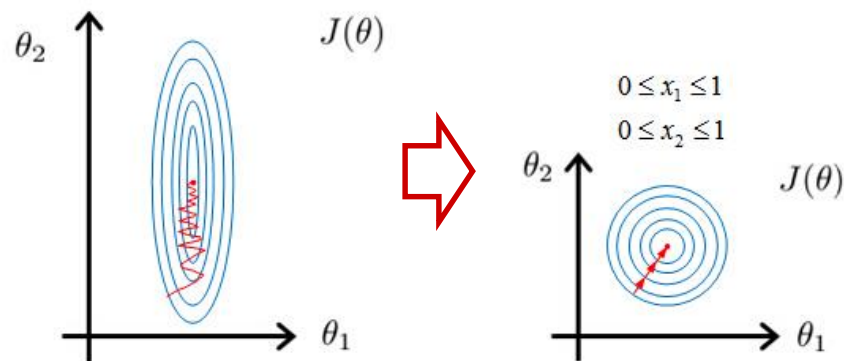
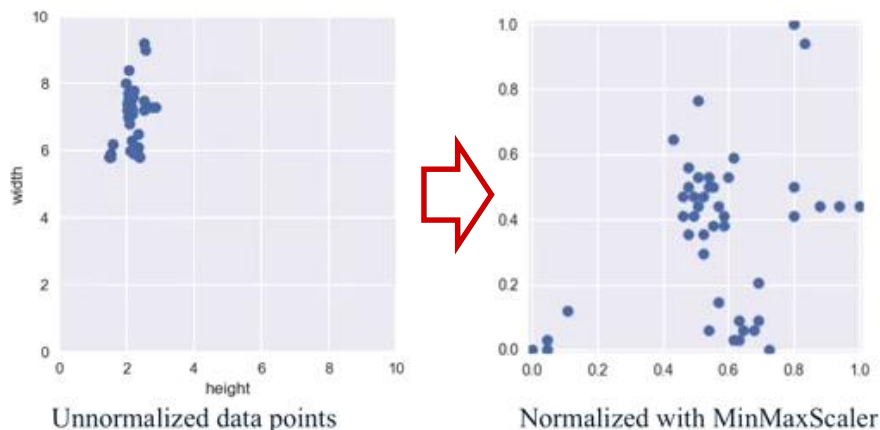
- $$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- `sklearn.preprocessing.MinMaxScaler`

- 标准化：将所有特征数据缩放成 平均值为0, 方差为1


- $$z = \frac{x - \mu}{\sigma}$$

- `sklearn.preprocessing.StandardScaler`



特征预处理

- 有序型特征，如：等级（A, B, C）；级别（低、中、高）
 - 转换成有序数值即可，如A->1, B->2, C->3
 - sklearn.preprocessing.LabelEncoder
- 类别型特征，如：性别（男、女）
 - 独热编码（One-Hot Encoding），如男->0 1, 女->1 0
 - sklearn.preprocessing.OneHotEncoder

ID	Gender		ID	Male	Female	Not Specified
1	Male		1	1	0	0
2	Female		2	0	1	0
3	Not Specified		3	0	0	1
4	Not Specified		4	0	0	1
5	Female		5	0	1	0

- 注意
 - 在测试集上的scaler或encoder和训练集上的scaler或encoder要保持一致
 - 不要在训练集和测试集分别使用不同的scaler或encoder

联系我们

小象学院：互联网新技术在线教育领航者

— 微信公众号：**小象学院**

