# 使用sklearn进行标签编码和独热编码

在使用python进行机器学习时，常常需要将非数值型特征转成标签编码和古道独热编码，以下面的例子为例，说明这两种编码的生成方法。

需要注意的是，在处理数据前，先将数据转成numpy.array类型，在进行OneHot编码前，先将数据转成Label编码，使用concatenate来插入列，使用delete方法来删除整列。

|  |
| --- |
| #coding=utf-8  **import** numpy **as** np  **from** sklearn**.**preprocessing **import** **\***  **def** loadData**():**  data **=** **[**  **[**'1th'**,** '20'**,** 'male'**],**  **[**'2th'**,** '21'**,** 'female'**],**  **[**'3th'**,** '32'**,** 'male'**],**  **[**'3th'**,** '21'**,** 'female'**],**  **[**'2th'**,** '56'**,** 'male'**],**  **[**'1th'**,** '43'**,** 'male'**],**  **[**'2th'**,** '29'**,** 'female'**],**  **[**'3th'**,** '12'**,** 'male'**],**  **[**'2th'**,** '8'**,** 'female'**],**  **[**'1th'**,** '43'**,** 'male'**],**  **]**  label **=** **[**  **[**1**],[**0**],[**1**],[**0**],[**1**],**  **[**1**],[**0**],[**1**],[**1**],[**1**]**  **]**  **return** np**.**array**(**data**),** np**.**array**(**label**)**  #载入数据  data**,** label **=** loadData**()**  num\_of\_rows **=** data**.**shape**[**0**]**  #1.处理性别数据  enc **=** LabelEncoder**()**  integer\_encoder **=** enc**.**fit**(**data**[:,**2**])**  class\_number **=** len**(**list**(**set**(**data**[:,**2**])))**  SexFeature **=** integer\_encoder**.**transform**(**data**[:,**2**])**  #查看编码与原特征的对应关系  **print(**integer\_encoder**.**classes\_**)**  **print(**integer\_encoder**.**transform**(**integer\_encoder**.**classes\_**))**  data**[:,**2**]=**SexFeature  #2.处理等级数据  #2.1首先，将非数字类别转换成数字类别  enc **=** LabelEncoder**()**  integer\_encoder **=** enc**.**fit**(**data**[:,**0**])**  class\_number **=** len**(**list**(**set**(**data**[:,**0**])))**  integer\_classes **=** integer\_encoder**.**transform**(**integer\_encoder**.**classes\_**).**reshape**(**class\_number**,**1**)**  tmp\_feature **=** integer\_encoder**.**transform**(**data**[:,**0**]).**reshape**(**num\_of\_rows**,** 1**)**  #2.2将数字类别转成OneHot编码  enc **=** OneHotEncoder**()**  oneHot\_encoder **=** enc**.**fit**(**integer\_classes**)**  ClassFeature **=** oneHot\_encoder**.**transform**(**tmp\_feature**)**  #2.3将编码转成numpy.array并替换原特征  data **=** np**.**concatenate**([**data**,** ClassFeature**.**toarray**()],** axis **=** 1**)**  data **=** np**.**delete**(**data**,** **[**0**],** axis**=**1**)**  feature\_names **=** **[**'Age'**,** 'Sex'**,** 'first\_class'**,** 'second\_class'**,** 'third\_class'**]**  data **=** data**.**astype**(**float**)**  **print(**feature\_names**)**  **print(**data**)** |

输出：

|  |
| --- |
| ['female' 'male']  [0 1]  ['Age', 'Sex', 'first\_class', 'second\_class', 'third\_class']  [[ 20. 1. 1. 0. 0.]  [ 21. 0. 0. 1. 0.]  [ 32. 1. 0. 0. 1.]  [ 21. 0. 0. 0. 1.]  [ 56. 1. 0. 1. 0.]  [ 43. 1. 1. 0. 0.]  [ 29. 0. 0. 1. 0.]  [ 12. 1. 0. 0. 1.]  [ 8. 0. 0. 1. 0.]  [ 43. 1. 1. 0. 0.]] |