实验一　线性回归与决策树

一、实验目的

线性回归是机器学习中有监督学习的一个重要方法，用于预测输入量和输出量之间的线性关系。特别是当输入量的值发生变化时，输出量的值随之发生线性变化。线性回归模型可以表示为从输入量到输出量之间映射的线性函数。

决策树也是一种机器学习中有监督学习方法。决策树是一种树形结构，其中每个内部节点表示一个属性（或特征）上的判断，每个分支代表一个判断结果的输出，每个叶节点代表一种分类结果。决策树的生成算法有 ID3、C4.5、C5.0、CART 等。

决策树就是给出一个样本集合，其中每个样本都有一组特征和一个分类标签，也就是分类结果已知。通过学习这些样本数据来建立一个决策树。这个决策树能够对新的样本给出正确的分类。

具体目的有：

（1）掌握线性回归算法、决策树算法 ID3 的原理；

（2）学会线性回归算法、决策树算法 ID3 的实现和使用方法。

二、实验类型

验证型、设计型。

三. 实验要求与支撑的课程目标

1. 实验要求

（1）课前充分预习理论基础知识，理解实验内容，做好实验准备；

（2）安装 Python3 和 Anaconda3；

（3）使用 Python 实现线性回归算法、决策树算法 ID3；

（4）针对具体的任务和数据集，运行 Python 程序；对结果进行可视化和分析，

得出结论。

（5）根据实验要求，做好预习，完成实验记录，提要规定的实验报告内容。

2.支撑的课程目标

本实验主要是线性回归算法、决策树算法 ID3 的实现和使用，因此主要支撑课程目标 2。

课程目标 2.能够根据实际数据分析的需要，选择和应用现有的学习算法，能够设计实验方案，实现经典的机器学习算法，开展实验和分析实验结果。（支持指标点 4-1）

四、实验平台

1. 操作系统：Windows 系统（64bit）；

2. 开发环境：Anaconda 3（64bit）的 Jupyter Notebook；
3. 开发语言：Python3。

五、实验内容

1. 一元线性回归模型实验

（1）假设 line-ext.csv 是对变量 y 随变量 x 的变化情况的统计数据集。请根据教材的公式（3.7，3.8），使用 Python 语言编程计算线性回归模型的系数，建立一个线性回归模型。要求如下：

1）计算出回归系数，输出模型表达式；绘制散点图和回归直线，输出均方误差。

参考代码：

i）数据预览

```
# 导入第三方模块
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# 导入数据集
income = pd.read_csv(r'line-ext.csv')
# 绘制散点图
sns.lmplot(x = 'YearsExperience', y = 'Salary', data = income, ci = None)
# 显示图形
plt.show()
```

ii）简单线性回归模型的参数 w , b 求解

```
# 样本量
n = income.shape[0]
# 计算自变量、因变量、自变量平方、自变量与因变量乘积的和
sum_x = income.YearsExperience.sum()
sum_y = income.Salary.sum()
sum_x2 = income.YearsExperience.pow(2).sum()
xy = income.YearsExperience * income.Salary
sum_xy = xy.sum()
# 根据公式计算回归模型的参数
w = (sum_xy-sum_x*sum_y/n)/(sum_x2-sum_x**2/n)
b = income.Salary.mean()-w*income.YearsExperience.mean()
```

```
# 打印出计算结果
print('回归参数 w 的值：',w)
print('回归参数 b 的值：',b)
print('模型表达式：f(x)=',w,'x+',b)
# 打印出均方误差
```

2）请给出当自变量 x =0.8452 时，因变量 y 的预测值。

```
# 请给出代码
# 打印出变量 x =0.8452 时，变量 y 的预测值
```

（2）对于上面的数据集 line-ext.csv，可以使用第三方模块 statsmodels 中的函数 ols( )来计算线性回归模型的系数，建立线性回归模型，并验证上面计算结果及预测结果。也可以使用第三方模块 sklearn 模块中的类 LinearRegression 来完成这个任务，请分别给出 Python 代码。

参考代码：

1）调用第三方模块 statsmodels 计算

```
# 请给出代码
# 导入第三方模块
import statsmodels.api as sm
# 利用收入数据集，构建回归模型
fit = sm.formula.ols('Salary ~ YearsExperience', data = income).fit()
# 返回模型的参数值
fit.params
```

2）调用第三方模块 sklearn 模块中的类 LinearRegression 计算

```
# 请给出代码
```

2. 决策树算法实验

（1）隐形眼镜数据集 glass-lenses.txt 是著名的数据集。它包含了很多患者眼部状况的观察条件以及医生推荐的隐形眼镜类型。要求：

1）使用 Python 语言建立决策树模型 ID3，划分 25%的数据集作为测试数据集。使用 Graphviz 工具，将此决策树绘制出来。 此小题代码由指导教师提供，必须运行出结果。

数据集的属性信息如下：

-- 4 Attributes

1. age of the patient: (1) young, (2) pre-presbyopic (for short, pre), (3)

presbyopic

2. spectacle prescription: (1) myope, (2) hypermetrope (for short, hyper)

3. astigmatic: (1) no, (2) yes

4. tear production rate: (1) reduced, (2) normal

-- 3 Classes

1 : the patient should be fitted with hard contact lenses,

2 : the patient should be fitted with soft contact lenses,

3 : the patient should not be fitted with contact lenses.

2）请使用测试数据集进行测试，输出测试准确率。

请给出代码

（2）对于数据集 glass-lenses.txt，使用第三方模块 sklearn 中的类 DecisionTreeClassifier（其中，criterion='entropy'，表示采用信息增益法选择特征）来建立决策树模型 ID3，重复 2（1）中的操作，验证上面计算结果。

1）数据预览

```
# 导入第三方包
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn import ensemble
from sklearn import metrics
from sklearn import tree
# 读入数据
fr = open('glass-lenses.txt')
lenses = [inst.strip().split('\t') for inst in fr.readlines()]
lensesLabels = ['age','prescript','astigmatic','tearRate','type']
lens = pd.DataFrame.from_records(lenses, columns=lensesLabels)
lens
```

2）数据预处理

```
#哑变量处理
dummy = pd.get_dummies(lens[['age','prescript','astigmatic','tearRate']])
# 水平合并数据集和哑变量的数据集
lens = pd.concat([lens,dummy], axis = 1)

# 删除原始的 age, prescript, astigmatic 和 tearRate 变量
```

```
lens.drop(['age','prescript','astigmatic','tearRate'], inplace=True, axis = 1)
lens.head()
```

3）将数据集拆分为训练集和测试集，且测试集的比例为 25%

```
X_train, X_test, y_train, y_test = model_selection.train_test_split(lens.loc[:,'
age_pre':'tearRate_reduced'],lens.type,test_size = 0.25, random_state= 1234)
```

4）构建分类决策树，请给出代码

5）输出预测准确率，请给出代码

3. 多元线性回归模型实验（选做，加分题）

以某产品的销售利润数据为例，该数据集（Predict to Profit.xlsx）包含 5 个变量，分别是产品的研发成本、管理成本、市场营销成本、销售市场和销售利润。数据集的部分截图如下所示。请根据此数据集建立一个预测利润的多元线性模型。

RD_Spend	Administration	Marketing_Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77
123334.88	108679.17	304981.62	California	149759.96
101913.08	110594.11	229160.95	Florida	146121.95
100671.96	91790.61	249744.55	California	144259.4
93863.75	127320.38	249839.44	Florida	141585.52

图 1 产品的利润数据集

提示：可以使用第三方模块 sklearn 实现多元线性回归。

六、参考资料

1. 周志华，机器学习，清华大学出版社，2016 年　第 3，4 章。