桂林電子科技大學
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

《机器学习》实验指导书

计算机与信息安全学院

教师：孙晋永，刘斯韵

2023 年 3 月

实验要求和注意事项（根据学校相关教学管理规定）

1、实验报告：按照学校的实验教学管理规定，学生完成每个实验后要交一份实验报告。大家每做完一个实验，回去后立即利用课余时间写好实验报告。实验报告的每个部分所论述的内容如下：

一、 实验目的与内容。

二、 实验方案、方法和过程等分析与设计内容。

三、 实验结果与分析。

四、 问题与总结：所遇到的具体问题和分析及解决过程，尚存在的一些问题，所获得的与课程相关、技术相关的心得与体会。

五、 附录：视需要附录代码或其它相关资料。

注意：最终的实验报告可打印，但必须使用正规实验报告纸或 A4 大小的纸张。实验报告雷同者，成绩为不及格！

2、课堂纪律：请勿迟到与早退。上课时请不要听歌或聊天，不要带耳塞，不要在实验室内吃东西和扔任何形式的废弃物。未经允许不能随便调整实验批次，有客观原因不能上课者必须有医院、年级主任等单位或领导的书面证明和签字盖章，随意口头请假者一般均视为无效。

目 录

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）

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。

实验二 神经网络

一、实验目的

神经网络（Neural Networks，NNs）也称为人工神经网络（Artificial Neural Networks，简写为 ANNs）。它是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。

BP 神经网络由一个输入层、一个输出层和一个或多个隐层构成。它的激活函数采用 sigmoid 函数，采用 BP 算法训练多层前馈神经网络。

BP 算法全称是误差反向传播(error Back Propagation，或者也称为误差逆传播)算法。基本思想：在 BP 神经网络中，输入信号经输入层输入，通过隐层计算后由输出层输出。把输出值与标记值比较，若有误差，将误差反向由输出层向输入层传播。在这个过程中，利用梯度下降算法对神经元权值进行调整。

具体目的是：

1. 掌握神经网络的 BP 算法原理与实现方法。
2. 神经网络的构建、训练和测试方法。

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）实验前充分准备，准备好要处理的数据，设计好相应的处理方法；

（2）在网上查找资料安装所需的软件环境。

（3）按实验老师要求提交实验报告。

2. 支撑的课程目标

本实验既要掌握 BP 算法的基本原理，也要把这些基本原理应用到实际问题的解决，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） Jupyter Notebook；
3. 开发语言：Python3；
4. 开发语言：TensorFlow 2.6.0 或其他版本；Keras，2.0 或其他版本。

5. 开发语言：Pytorch 1.4.0 或其他版本。

五、实验内容

1.使用 Python 语言编程实现标准 BP 算法和累积 BP 算法，在 wine 数据集（wine_data-2.csv）上分别使用这两个算法训练一个单隐层网络（如，13×50×1），并进行比较。教师给出两种算法的部分代码。要求：1）学习率 e 在[0.001, 0.2]内，分析 e 的大小对算法性能的影响。2）绘制均方误差随训练轮数的变化曲线。3）改变隐层神经元的个数，观察网络的性能，进行分析。4）输出混淆矩阵和准确率。

说明：

（1）wine 数据集的最后一列为 wine 的类别：0, 1。

（2）使用尝试法设置适当的训练轮数。

2. 使用 TensorFlow（或 Pytorch）建立一个 3 层神经网络（如，13×50×1），对 wine 数据集（wine_data-2.csv）进行分类。教师给出部分代码。要求：1）学习率 e 在[0.01, 0.05]内调整，分析 e 的大小对算法性能的影响。2）绘制均方误差、准确率随训练轮数的变化曲线。3）改变隐层神经元的个数，观察网络的性能，进行分析。4）输出混淆矩阵和准确率。

3.（可选，加分）使用 TensorFlow 构建一个卷积神经网络（CNN）对手写体数字数据集（mnist）进行分类，要求将标签为 0 和 1，2 和 3，4 和 5，6 和 7，8 和 9 的数字分别分为一类（共 5 类）。使用 5 个神经元输出结果，比如第 1 类输出[1,0,0,0,0]，第 2 类输出[0,1,0,0,0]，第 3 类输出[0,0,1,0,0]，第 4 类输出[0,0,0,1,0]，第 5 类输出[0,0,0,0,1]。并输出这个卷积神经网络的测试精度。

六、参考资料

1. 周志华，机器学习，清华大学出版社，2016 年　第 5 章.

实验三 支持向量机

一、实验目的

支持向量机（Support Vector Machine, SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器。其决策边界是对学习样本求解的最大边距超平面。支持向量机的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机。SVM 还包括核技巧，这使它成为实质上的非线性分类器。SVM 的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。

本实验的目的是训练学生掌握支持向量机模型 SVM 的原理和使用方法。

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）实验前充分准备，准备好要处理的数据，设计好相应的处理方法；

（2）按实验老师要求提交实验报告。

2. 支撑的课程目标

本实验既要掌握支持向量机的基本原理，也要把这些基本原理应用到实际问题的解决，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） Jupyter Notebook；
3. 开发语言：Python3。

五、实验内容

1. Python 的 sklearn 扩展包中包含了“威斯康星州乳腺癌”数据集，其中详细记录了威斯康星大学附属医院的乳腺癌测量数据。数据集包括 569 行和 31 个特征。可以使用这个数据集训练一个支持向量机模型，参考代码如下。教师给出部分代码。请输出测试精度。

参考代码：

```
from sklearn.svm import SVC
```

```
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap

# 导入肺癌数据集
data = load_breast_cancer()
X = data['data']
y = data['target']
print(X.shape)

x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
# 建立支持向量机模型
clf1 = SVC(kernel='linear')          #线性核函数
clf2 = SVC(kernel='rbf', C=10, gamma=0.0001)     #采用高斯核函数
clf1.fit(x_train, y_train)
clf2.fit(x_train, y_train)

# 输出两个 SVM 模型的精度。 此处填入你的代码。
```

2. 保持公司的员工满意的问题是一个长期存在且历史悠久的挑战。如果公司投入了大量时间和金钱的员工离开，那么这意味着公司将不得不花费更多的时间和金钱来雇佣其他人。以 IBM 公司的员工流失数据集（HR-Employee-Attrition.csv）作为处理对象，使用第三方模块 sklearn 中的相关类来建立支持向量机模型，进行 IBM 员工流失预测。教师给出部分代码。要求：1）对数据集做适当的预处理操作，2）划分 25%的数据集作为测试数据，3）输出支持向量，4）输出混淆矩阵，计算查准率、查全率和 F1 度量，并绘制 P-R 曲线和 ROC 曲线。

说明：数据集的第 2 列（Attrition）为员工流失的类别。

3. （可选，加分）手写字母的识别问题。对于手写英文字母，可以根据写入字母的特征信息（字母的宽度、高度、边际）来判断其属于哪一种字母。手写体字母数据集一共包含 20000 个样本，每个样本有 17 个特征，其中 letter 为类别，具体值就是 20 个英文字母。请利用对该数据集训练一个 SVM 模型，进行分类判断。

	letter	xbox	ybox	width	height	onpix	xbar	ybar	x2bar	y2bar	xybar	x2ybar	xy2bar	xedge	xedgey	yedge	yedgex
0	T	2	8	3	5	1	8	13	0	6	6	10	8	0	8	0	8
1	I	5	12	3	7	2	10	5	5	4	13	3	9	2	8	4	10
2	D	4	11	6	8	6	10	6	2	6	10	3	7	3	7	3	9
3	N	7	11	6	6	3	5	9	4	6	4	4	10	6	10	2	8
4	G	2	1	3	1	1	8	6	6	6	6	5	9	1	7	5	10

图 2　手写体字母数据集的前 5 行预览

六、参考资料

1. 周志华，机器学习，清华大学出版社，2016 年　第 6 章.

实验四 集成学习

一、实验目的

集成学习(Ensemble learning)将多个基学习器进行结合，通常可以获得比单一学习器更加显著的泛化性能。这对“弱学习器”尤为明显。因而集成学习的理论研究都是针对弱学习器进行的，而基学习器有时也被直接称为弱学习器。

AdaBoost 是一种典型的集成学习算法，其核心思想是针对同一个训练集训练不同的分类器（弱分类器），然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。

随机森林就是通过集成学习思想将多棵决策树集成的一种算法。它的基学习器是决策树，属于一种集成学习（Ensemble Learning）方法。

本实验的具体目的是：

（1）掌握 AdaBoost 算法、随机森林算法的基本原理；

（2）掌握 AdaBoost 算法实现和使用方法、以及随机森林算法的使用方法。

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）实验前充分准备，准备好要处理的数据，设计好相应的处理方法；

（2）按实验老师要求提交实验报告。

2. 支撑的课程目标

本实验既要掌握数据挖掘的基本原理和算法，也要把这些基本算法应用到实际问题的解决，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） Jupyter Notebook；
3. 开发语言：Python3.7。

五、实验内容

1. 使用 Python 语言实现 AdaBoost 算法，在马氙气数据集（horseColicTest.txt，

horseColicTraining.txt）上训练一个集成分类器，估计马疝气的死亡率。教师给出部分代码。要求：输出混淆矩阵，计算查准率、查全率和 F1 度量，并绘制 P-R 曲线。

说明：数据集 horseColicTraining.txt 的最后一列为马的类别：1—仍存活，0—未能存活。

2. 使用第三方模块 sklearn 中的随机森林分类器 RandomForestClassfier 为两栖动物数据集（Electrical-Grid-Data.csv）实现一个分类模型，并完成数据分析任务。教师给出部分代码。要求：1）使用数据集的 75%作为训练数据，25%作为测试数据；2）对基础决策树的个数（假定为 50，100, 300）和每个基础决策树的深度（假定为 3, 5, 7, 9）使用网格搜索方法（使用类 GridSearchCV）；3）输出混淆矩阵，计算查准率、查全率和 F1 度量，并绘制 P-R 曲线和 ROC 曲线。

说明：数据集的最后一列为电网稳定类别。

3.（可选，加分）现有一个来自 UCI 网站的信用卡数据集。它一共包含有 30000 条记录和 25 个变量，其中自变量包含客户的性别、受教育水平、年龄、婚姻状况、信用额度、6 个月的历史还款状态、账单金额以及还款金额。标签 y 表示用户在下个月的信用卡还款中是否存在违约的情况（1 表示违约，0 表示不违约）。请使用集成学习算法 AdaBoost 和 GBDT（梯度提升树）训练一个分类模型，并给出测试精度。

六、参考资料

1. 周志华，机器学习，清华大学出版社，2016 年　第 8 章.