**一、关于期末测验题的说明：**

**1、每个测验题基础分为20分**

**2、每个测验题都有难度系数，用1到4颗“\*“来表示，对应难度系数1.0到4.0**

**3、每个测验题总分为该题基础分乘以难度系数**

**4、以期末项目小组方式进行测验题解答，完成后，按照小组提交“期末测验题答卷.doc”，完整代码以.ipynb文件单独提供**

**5、每个期末项目小组须至少选总分为100分的题（超过100分则按最高得分计算）进行详细回答（包括原理阐述，数据描述，模型设计，代码实现，实际数据集上测试情况等）**

**6、期末项目小组推荐4人一组，对于多于4人的组，每增加一个人须在“总分100分”基础上额外增加一个测验题**

**二、实施方式与注意事项：**

1、采用分组方式（具体分组名单见“20计科智能算法期末项目组队.xlsx”），**相互学习、优势互补、团队协作**

2、每组原则上为四人，合理分工：按照题目分工，或者按照步骤分工（比如数据及软硬件环境、模型设计及结果分析、编码实现及调试、撰写解答等）

3、每组选出一个组长，负责分工与协调，并对每个组员在项目中的表现打分(满分为100分)

4、**17周周一发布，提交截止时间为18周周日晚12点**

5、可以查阅相关文献（在参考文献中列出），但千万不要简单拷贝，一定要有自己的理解和验证，一定要用自己的语言表述，更要敢于质疑和批判

**特别注意：发现简单拷贝者视具体情况扣分！！**

6、各个组必须按照要求独立、认真完成

**特别注意：发现相互抄袭者均记为不及格！！**

第一组测验题：

1-1、（\*\*\*）请推导对率回归的解析解，并在实际数据集上验证。

1-2、（\*\*）基于1-1，请在“病马预测数据集”上比较对率回归的解析解和梯度下降的解，谁的训练误差最低，谁的测试误差最低。

1-3、（\*\*）基于1-2，请思考：对率回归的解析解一定是我们在机器学习任务上期望得到的最优解吗？为什么？

1-4、（\*\*）关于对率回归的求解，请在1-1、1-2、1-3的基础上，进一步思考梯度下降和解析解相比有什么特点和优势，为什么？

1-5、（\*\*\*）关于对率回归，如果没有偏置项会有什么影响？为什么？请在实际数据集上验证你的分析和判断。

1-6、（\*）用梯度下降求解对率回归模型的过程中，参数的初值我们是如何设置的呢？可以设置为其它值吗？为什么？请在实际数据集上验证你的分析和判断。

1-7、（\*\*）用梯度下降求解对率回归模型的过程中，学习率我们是如何设置的呢？为什么这么设置呢？能够搜索到最佳学习率吗？请在实际数据集上验证你的分析和判断。

1-8、（\*）用随机梯度下降求解对率回归模型的过程中，“小batch”的大小和一遍（epoch）迭代的抽样次数我们应该如何设定？请在实际数据集上尝试设置不同的“小batch”大小，并对结果进行比较。

第二组测验题：

2-0、（\*\*）对率回归模型是关于参数w的线性模型吗？为什么？

2-1、（\*\*）对率回归模型需要对数据做归一化处理吗？为什么？请在实际数据集上验证你的分析和判断。

2-2、（\*）kNN模型需要对数据做归一化处理吗？为什么？请在实际数据集上验证你的分析和判断。

2-3、（\*\*\*）思考如何用对率回归模型解决多分类问题，并在二值手写数字识别数据集上进行验证和测试。

2-4、（\*\*）对于机器学习问题，我们真正关心的是测试集上的精度，因此在训练集上最优不一定在测试集上最优，那么我们应该如何选出在测试集上最优的模型呢？请用对率回归模型和梯度下降方法在实际数据集上验证你的想法。

2-5、（\*\*\*）说明如何将对率回归模型推广到能解决多分类问题的Softmax模型，并在二值手写数字识别数据集上进行验证和测试。

第三组测验题：

3-1、（\*\*\*\*）思考如何用多个SVM解决多分类问题，并在二值手写数字识别数据集上进行验证和测试。然后，与kNN模型和对率回归模型的多分类结果进行比较，并给出分析和解释。

3-2、（\*\*）（代码分析与纠错）以下代码定义的函数calcWs是SVM代码的一部分，请阅读相关代码回答以下问题：

def calcWs(alphas,dataArr,classLabels):

    X = mat(dataArr); labelMat = mat(classLabels).transpose()

    m,n = shape(X)

    w = zeros((n,1))

    for i in range(m):

        w += (alphas[i]\*labelMat[i])\*X[i,:].T

    return w

（1）参数alphas、dataArr、classLabels各自的具体意义、数据类型和维数？

（2）每一行语句的具体作用？

（3）m、n、w分别表示什么？

（4）返回值w的数据类型和维数？

（5）这段代码有错吗？如果有错请把具体错误指出来，并完成纠错。

3-3、（\*\*\*）请尝试用线性核SVM求解“病马预测数据集”，并与对率回归模型（包括解析解和梯度下降解）比较训练误差和测试误差、并分析比较结果。

第四组测验题：

4-1、（\*\*\*）全连接神经网络模型中，我们是如何在网络的输出层，用多个对率回归模型解决的Mnist数据集上的多分类问题呢？还有其它解决方案吗？哪个方案更优呢？

4-2、（\*\*）为什么对率回归模型中，我们要增加一列取值固定为1.0的特征，而在SVM和NN模型中我们并不需要这样做呢？

4-3、（\*）就梯度归一化比较对率回归模型和NN模型。

4-4、（\*\*\*）全连接神经网络模型可由对率回归单元堆叠而成，那么基于对率回归的解析解（问题1-1）同样能够求解全连接神经网络模型吗？为什么？

4-5、（\*\*）全连接神经网络模型中的对率回归单元都是二分类器吗？为什么？

第五组测验题：

5-1、（\*\*\*\*）核函数是如何应用在“局部加权线性回归”模型中的？可否在对率回归模型中类似应用？请完成编码并在实际数据集上验证。

提示：第二问请与问题1-1结合起来考虑。

5-2、（\*\*）阅读和执行以下代码：

>>> xArr,yArr=regression.loadDataSet('ex0.txt’)

>>> ws = regression.standRegres(xArr,yArr)

>>> xMat=mat(xArr)

>>> yMat=mat(yArr)

>>> yHat = xMat\*ws

（1）xArr, yArr, ws, xMat, yMat, yHat各自的具体意义、数据类型和维数？

（2）说明每一行语句的具体作用？

（3）这几行代码都能顺利执行吗？如果能执行，请给出执行结果；如果不能执行，请分析出错原因，并加以修改和完善，使之能够正确执行。

5-3、（\*\*）线性回归模型必须是关于输入数据x的线性模型吗？为什么？请举例说明。

5-4、（\*\*\*\*）请首先在实际数据集上验证“决策树模型容易过拟合”；在此基础上，查阅参考书和相关资料，说明有哪些避免决策树过拟合的方法；最后，选取一种避免决策树过拟合的方法，在实际数据集上进行验证。

第六组测验题：

6-1、（\*）“diffMat = np.tile(inX, (dataSetSize,1)) - dataSet”可以替换为“diffMat = inX - dataSet”吗？为什么？

6-2、（\*）“normDataSet = dataSet – np.tile(minVals, (m,1))

normDataSet = normDataSet/np.tile(ranges, (m,1))”可以替换为“normDataSet = (dataSet - minVals)/ranges”吗？为什么？

6-3、（\*\*）kNN算法的分类决策规则除了基本的“少数服从多数”之外，还可能有哪些变种？请选取其中的一种变种，在实际数据集上进行验证。

6-4、（\*\*\*）请以实际数据集为例，分析kNN算法的计算复杂度；进一步查阅参考书和相关资料，说明有哪些加速算法，这些算法各自有什么优缺点和适应场景。

6-5、（\*\*）说明“多变量决策树”与基本的“单变量决策树”的不同之处，并举例说明“多变量决策树”的优势在哪里。