代码使用说明及模型的介绍

1. 代码文件夹及其代码文件的介绍

代码文件夹内容介绍：

代码文件夹中共有 1 个文件夹，分别是 data。

data 文件夹下存放的是训练集数据 data.xlsx 和测试集数据 100_test.xlsx（是最新的测试集数据）。其中 data.xlsx 的 sheet 名称为 train_400，即 400 个训练数据，100_test.xlsx 的 sheet 名称为 valid_data，即 100 个测试数据。

代码介绍：

由于模型训练较快，因此将模型的训练和测试统一写成了 run_model.m 脚本文件。其中第一种模型对应的函数脚本文件为 svr_train.m, svr_predict.m, gaussKernel.m；第二种模型对应的函数脚本文件为 fmincg.m, linearRegCostFunction.m, trainLinearReg.m。

2. 运行代码方法

- 直接打开 matlab 软件运行 run_model.m 脚本文件。
- 运行脚本文件后，命令行终端输出选择模型，共有三个模型提供选择，分别对应数字 1、2、3，比如要选择第一个模型，则在终端中输入 1。
- 程序就会根据输入的模型进行训练，最后输出验证集的预测准确率。

3. 验证集准确率介绍

- 将验证集准确率定义为模型输出和验证集输出值偏差在 5%范围内的数据占比。
- 模型输出和验证集输出值偏差定义为二者的绝对值之差与标准输出的占比。
- 比如模型输出为 90，标准输出为 96，则偏差为(96-90)/96 = 0.0625,即 6.25%。
- 而验证集中数据有 30 个满足以上偏差范围，则准确率为 30/40=75%。

4. 模型介绍

- 模型 1：模型使用 SVR 方法(支持向量回归机)，是支持向量机的回归形式，使用带有松弛变量的最小化数据点距平面的距离和。使用梯度下降求解平面法向量，得到分界平面，根据不同的核函数可以求解不同分界曲面。

- 模型 2：使用线性映射（非线性化）的线性回归模型，线性映射是将 6 个特征投影到高维的空间进行非线性化，然后再高维空间中使用梯度下降法进行线性回归拟合，其中添加正则化项用参数 λ 进行控制，参数 λ 的大小意味着模型对于正则化项的重视程度。该正则化项防止过拟合，而使模型不具有很好的泛化性。
  - 模型 2 使用的线性映射有变化，映射方式为特征 2、4、6 分别平方，其他特征不变。
  - 正则化项系数改变为 0.0002。

- 模型 3：模型的结构为 3 层 BP 网络结构。其中输入层有 6 个结点，隐藏层分别有 200 个结点，最后输出层有 1 个输出结点。隐藏层使用对数 S 型激活函数，输出层使用双曲正切激活函数。

5. 模型结果分析

- 经过多次训练后固定参数权重得出，模型 1 的训练集和测试集准确率都为 61.0%；模型 2 的训练集准确率为 85.0%，测试集准确率为 87%；模型 3 的训练集的准确率为 98.0%，测试集准确率为 87%。
- 因此，可以得出模型 2、3 的预测准确率最高，模型性能最好。
- Bp 算法性能能提升的原因：
  - 网络层宽度就是结点数能减少过拟合的可能
  - 特征个数，这里道个歉，总共有 6 个特征，输入节点应该要有 6 个，原始代码一直以 5 个神经元作为输入节点，因此性能不佳。

6. 模型训练参数

- 模型 1 训练参数
  + 正规化因子 C：40，又称惩罚系数
  + 松弛变量：0.001
  + γ 核函数系数：0.150
  + 核函数：gaussian Kernel

- 模型 2 训练参数
  + 正规化因子 λ：0.0002

- 模型 3 训练参数
  + 学习率：0.01
  + 训练最大回合数：1000
  + 动态因子：0.9
  + 最小平方误差目标：10^-7
  + 梯度下降法：trainlm