Machine Learning Exercise 1.1
姓名：沈金龙
学号：18214806

实验题目：
编写程序：模拟仿真多项式回归
要求如下：

（1） 生成正弦序列 s(n);

（2） 使用噪声函数对正弦序列加噪 x(n)=s(n)+w(n);

（3） 使用多项式回归模型对 x(n)进行拟合，并分析过拟合和欠拟合情况

实验过程：
本实验采用 **Octave4.4.1** 完成仿真任务；
1、假设正弦序列 s(n) = sin(2*pi*n)，利用 linspace 函数在[0,1]区间取 10 个均值点作为 s(n)函数的输入；噪声函数 w(n) = 0.25*randn(1,N)，噪声符合正态分布，加噪之后的函数假设为 x(n)，则 x(n) = s(n) + w(n)；
2、假设多项式函数 y =w_0 + w_1*x + … + w_m*x^M，其中 M 为多项式的最高阶数。本次实验利用 polyfit 以及 polyval 函数根据给定训练集(x, xn)和指定的多项式阶数 M(M=0,1,3,9)来求得指定阶数多项式的参数 w，以此研究不同阶数多项式过拟合以及欠拟合的情况；

部分代码截图：
>模拟仿真正弦函数以及增加噪声后的函数：

```
% -----模拟仿真正弦函数以及增加噪声后的函数----%
N = 10;
NN = 666;
x = linspace(0, 1, N);   % 均分指令,产生0到1之间的N个行矢量
x_fits = linspace(0,1,NN);
sn = sin(2*pi*x); % sn 代表正弦函数
sn_fits = sin(2*pi*x_fits);   % sn_fits用于画出光滑的sn曲线
xn = sn + 0.25*randn(1,N);   % xn代表sn加噪后得到的函数,所加噪声符合高斯分布
                             % randn(1,N)表示生成1*N的,期望为0,标准差为1的正态分布量.
plot(x_fits,sn_fits, 'g', x, xn, 'bo', 'LineWidth',2);   %
legend('s(n)', 'x(n)');   % 标识图例
% set(get(gca,'title'),'fontname','宋体')
% title('模拟仿真正弦函数以及增加噪声后的函数')   % 注:中文标题会乱码
```

>模拟仿真实现不同阶数的多项式拟合：

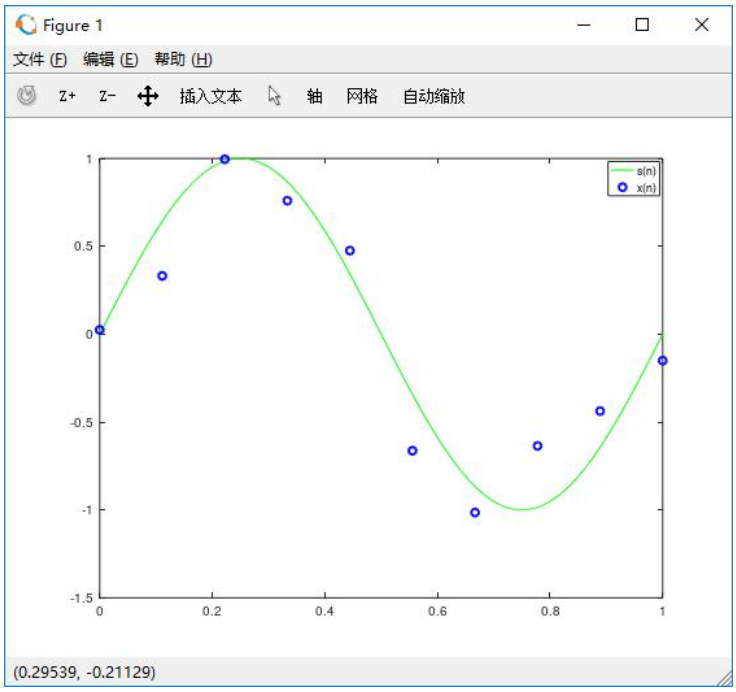
```
% -----仿真实现不同阶数的多项式拟合-----%
i=1;
figure;
for M=[0 1 3 9]
    % w代表M阶多项式的系数
    w = polyfit(x, xn, M);  % 返回w为幂次从高到低的多项式系数向量w
    y = polyval(w, x_fits);  % 返回对应自变量x_fits在给定系数w的多项式的值

    subplot(2,2,i);    % 生成2*2大小的合并子图
    i=i+1;
    plot(x_fits, sn_fits, 'g', x, xn, 'bo', x_fits, y, 'r', 'LineWidth',2);
    str = ['M=' mat2str(M)];  % 标识M的阶数,mat2str将矩阵转化为字符串
    text(0.6, 0.8, str);  % 在图中指定位置显示字符串str
    legend('s(n)', 'x(n)', 'y');  % 标识图例
end
```

>模拟仿真 9 阶多项式拟合不同数据集的表现：

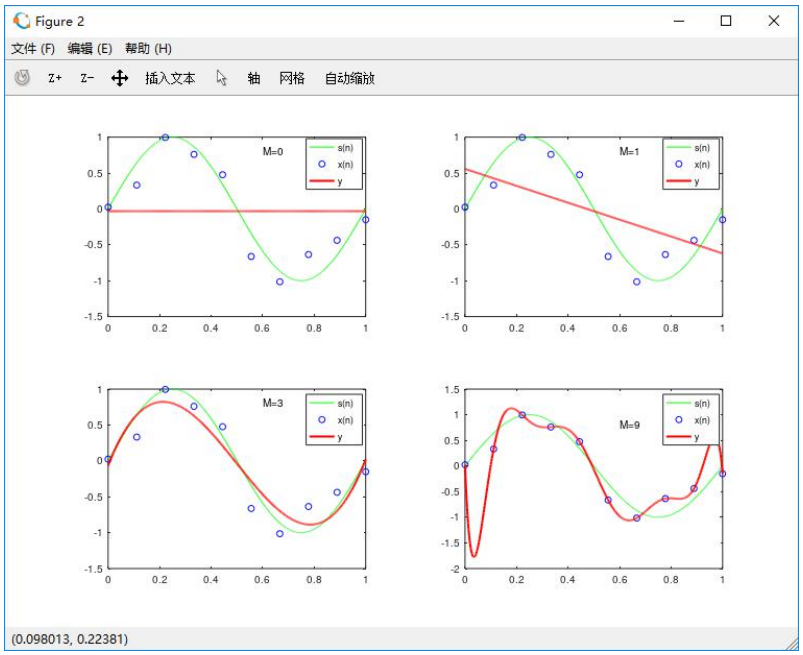
```
% -----模拟仿真9阶多项式拟合不同数据集的表现-----%
i = 1;
figure;
for N=[10,20,50,100]
    x = linspace(0, 1, N);
    sn = sin(2*pi*x);
    xn = sn + 0.25*randn(1,N);
    w = polyfit(x, xn, 9);
    y = polyval(w, x_fits);

    subplot(2,2,i);
    i = i+1;
    plot(x_fits, sn_fits,'g', x, xn, 'bo', x_fits, y, 'r', 'LineWidth',2);
    str = ['N=' mat2str(N)];
    text(0.6, 0.8, str);
    legend('s(n)', 'x(n)', 'y');
end
```

实验结果截图：

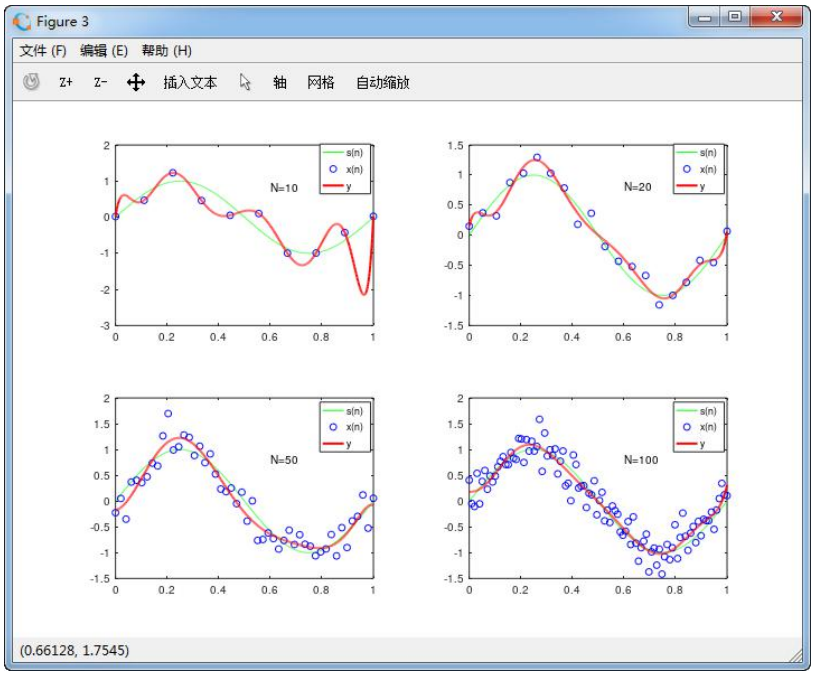
>正弦序列 s(n)和增加噪声后的序列 x(n)的图像如下图所示：



>不同阶数多项式 y 拟合 x(n)图像如下图所示：



>保持 M=9 不变，增加数据集的大小，拟合图像如下图所示：



实验分析：

1、 当 M 比较小时，如 M=0 或者 M=1 时，会出现欠拟合现象，多项式曲线 y 无法很好的拟合数据集(x,xn)，主要原因是模型太简单，没有很好地捕捉到数据特征，不能够很好地拟合数据。

2、 随着 M 的增加，模型逐渐变得复杂，y 对数据集的拟合效果也越来越好。

3、 当 M=9 时，出现了过拟合现象，主要原因是训练的数据量太小或者说训练数据占总数据的比例过小。

4、 当多项式的阶数固定为 M=9 时，随着训练集的增加，模型的过拟合问题得到了解决。