用梯度下降法拟合多项式

- 算法说明

n次多项式函数的形式为$P(x;\theta)=\sum_{i=0}^{n}a_i\times x^i$，其中$\theta=(a_0,a_1,a_2,\ldots,a_n)^T$是多项式的参数向量，拟合多项式就是确定多项式参数的过程。假设观测到多项式曲线上的m个点：$D=(x^{(j)},y^{(j)})_{j=1}^{m}$,其中$y^{(j)}=P(x^{(j)};\theta)+\epsilon^{(j)}$，即，第$j$个观测点的函数值$y^{(j)}$收到一个小噪声$\epsilon^{(j)}$的污染，因此$y^{(j)}\neq P(x^{(j)};\theta)$,为了找到合适的参数$\theta$，我们采用最小二乘准则，即，找一个合适的参数向量$\theta$使得$P(x;\theta)$在样本集$D$上的均方误差(Mean Squared Error)$l(\theta;D)$最小，均方误差定义为：

$$l(\theta;D)=\frac{1}{m}\sum_{j=1}^{m}(P(x^{(j)};\theta)-y^{(j)})^2 \tag{1}$$

把上面的公式展开：

$$l(\theta;D)=\frac{1}{m}\sum_{j=1}^{m}(\sum_{i=0}^{n}a_i\times(x^{(j)})^i-y^{(j)})^2 \tag{2}$$

引入特征变换：

$$\begin{aligned}&x\to z=(z_0,z_1,z_2,\ldots,z_n)^T\\&z_i=x^i,i=0,1,\ldots,n\\&z=\phi(x)=(1,x,x^2,\ldots,x^n)^T\in R^{n+1}\end{aligned} \tag{3}$$

那么，多项式可以写为:$P(z;\theta)=\theta^Tz$，上述均方误差可以改写为：

$$l(\theta;D)=\frac{1}{m}\sum_{j=1}^{m}(\theta^Tz^{(j)}-y^{(j)})^2=\frac{1}{m}\sum_{j=1}^{m}(\sum_{i=0}^{n}a_i\times z_i^{(j)}-y^{(j)})^2 \tag{4}$$

用样本集D拟合$P(x;\theta)$的梯度下降过程是一个迭代过程：

0. 在第$t=0$步，初始化参数$\theta^{(0)}=\mathbf{0}:a_i=0,i=0,1,\ldots,n$

1. 在第$t+1$步，计算$l(\theta;D)$相对于参数$\theta^{(t)}=(a_0^{(t)},a_1^{(t)},\ldots,a_n^{(t)})$的梯度：

$$\Delta a_k=\frac{\partial l(\theta^t;D)}{\partial a_k^{(t)}}=\frac{2}{m}\sum_{j=1}^{m}[(\sum_{i=0}^{n}a_i^{(t)}\times z_i^{(j)}-y^{(j)})\times z_k^{(j)}]$$

   在编码实现时，可以先计算$P(x;\theta^{(t)})$的预测误差：$e^{(j)}=P(x^{(j)};\theta^{(t)})-y^{(j)}=\sum_{i=0}^{n}a_i^{(t)}z_i^{(j)}-y^{(j)}$；然后计算梯度：$\Delta a_k=\frac{2}{m}\sum_{j=1}^{m}e^{(j)}\times z_k^{(j)}$。

2. 更新参数：

$$a_k^{t+1}=a_k^t-\alpha\times\Delta a_k$$

   其中，α是一个常数，表示梯度更新的速率，可以自己调节，你可以试一下$\alpha=0.01$。

3. 如果收敛，结束，否则转1。 判断是否收敛可以通过比较$l(\theta^t;D)$与$l(\theta^{t+1};D)$之间的差异实现，如果$|l(\theta^{t+1};D)-l(\theta^t;D)|<\epsilon$，就可以认为算法已经收敛，停止迭代。$\epsilon$是一个很小的值，比如$\epsilon=1e-6$。

- 特征缩放

上面的算法中，不同次幂的数值范围相差非常大，给算法带来很大的困难，因此要做一些预处理。普遍的做法是对输入向量$z=(z_0,z_1,\ldots z_n)^T$的各个维度按如下方式做归一化，然后用归一化的样本$(z^{(j)},y^{(j)})_{j=1}^m$拟合多项式。

$$
\begin{aligned}
&\mu_i = \frac{1}{m}\sum_{j=1}^{m} z_i^{(j)} \\
&\sigma_i^2 = \frac{1}{m}\sum_{j=1}^{m}(z_i^{(j)} - \mu_i)^2 \\
&z_i^{(j)} \leftarrow \frac{z_i^{(j)} - \mu_i}{\sigma_i + \eta}
\end{aligned} \tag{5}
$$

注意：

1) 不要对z_0做归一化，因为$z_0^j = 1, \forall j$;

2) 公式(4)中的η是一个很小的正实数，比如$\eta = 1e-3$，因为 σ_i可能为0，为避免出现被0除，引入该常数。

- 生成随机样本

  本作业要求大家首先设定一个已知参数的多项式$Q(x)$，然后生成一些随机数$x^{(j)}$以及小幅度的噪声$\epsilon^{(j)}$，得到$y^{(j)} = Q(x^{(j)}) + \epsilon^{(j)}$,假设$\epsilon^{(j)}$服从0均值的高斯分布，噪声方差由大家自己设定，不宜过大。

  得到样本后，计算各个次幂，把$x^{(j)}$转换为$z^{(j)}$，并做特征缩放处理。然后使用上述迭代算法拟合多项式参数。

- 计算拟合误差

  使用拟合后的多项式系数θ，按照公式(4)计算$l(\theta; D)$，得到拟合误差，误差的大小反映了拟合结果的好坏。