作业5：支持向量机

1. 分析图片并回答问题

1.判断核函数

首先，由分类面形状可以判断出图(c)是线性分类面，对应A线性核。

径向基核的核函数为$K(x,x_i)=exp\{-\frac{|x-x_i|^2}{\sigma^2}\}$，可以实现将低维空间的样本点投影到高维空间，属于局部核函数，其特点为支持向量分布较广，参数σ控制函数径向作用范围。由于图像(b)、(d)、(f)存在远离图中分类面的支持向量，说明核函数将样本点投影到高维空间，使用的是径向基函数。由径向基函数性质，图(d)将两类样本分离的更为细致，径向基半径更小，对应核函数F径向基(σ=0.1)；图(f)对应核函数E径向基(σ=0.5)；图(b)对应核函数D径向基(σ=1)

多项式核函数为$K(x,x_i)=((x\cdot x_i)+1)^d$，多项式函数也可以实现将低维空间的样本点投影到高维空间，属于全局核函数，其特点为支持向量距离分界面较集中，分界曲线为d次多项式函数。进一步仔细观察分类面形状及分类效果，可以看出，图(a)是二次函数分类面，对应B二次多项式核；图(e)对应C三次多项式核，分类效果比图(a)更好。

图片	(a)	(b)	(c)	(d)	(e)	(f)
核函数	二次多项式核	径向基(σ=1)	线性核	径向基(σ=0.1)	三次多项式核	径向基(σ=0.5)

2.核函数选择

我认为针对这个数据集选择线性核更合适。首先该问题是线性可分的，选用线性核可以取得较好的分类效果；其次，采用非线性变换时将产生额外的变换计算，使得空间维数升高，且模型复杂易出现过拟合现象，因此我认为选取线性核更合适。

2. 多项式核支持向量机简单分类

1.采用拉格朗日方法，问题可表示为：

$$min_{\omega,\omega_0}max_{\alpha_i}L(\omega,\omega_0,\alpha)=\frac{1}{2}(\omega\cdot\omega)-\sum_{i=1}^{2}\alpha_i\{y_i[(\omega^T\cdot\phi(x_i)+\omega_0]-1\}$$

分别对ω,ω_0求导有：

$$\frac{\partial L}{\partial\omega}=\omega-\sum_{i=1}^{2}\alpha_i y_i\phi(x_i)=0$$

$$\frac{\partial L}{\partial\omega_0}=\sum_{i=1}^{2}\alpha_i y_i=0$$

即：

$$\omega = \alpha_1 y_1 \phi(x_1) + \alpha_2 y_2 \phi(x_2)$$
$$\alpha_1 y_1 + \alpha_2 y_2 = 0$$

将$y_1 = -1, y_2 = 1$代入有：

$$\alpha_1 = \alpha_2$$

不妨设$\alpha_1 = \alpha_2 = \alpha$，则与最优向量$\omega$平行的向量可表示为$[0, \sqrt[4]{2}\alpha, 2\alpha]^T$

故其中一个与最优向量ω平行的向量为$[0, \sqrt[4]{2}, 2]^T$

2.将约束条件代入，最优化函数可表示为：

$$max_\alpha Q(\alpha) = \sum_{i=1}^{2} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{2} \alpha_i \alpha_j y_i y_j (\phi(x_i), \phi(x_j))$$

将数据代入后有:

$$max_\alpha Q(\alpha) = 2\alpha - [\alpha^2 - (5+\sqrt{2})\alpha^2 + 2\alpha^2] = 2\alpha - \frac{1}{2}(4+\sqrt{2})\alpha^2$$

可知当$Q(\alpha)$取最大值时

$$\alpha = \frac{2}{4+\sqrt{2}}$$

将α代入$\omega = \alpha(y_1\phi(x_1) + y_2\phi(x_2))$可得

$$\omega = [0, \frac{2^{\frac{5}{4}}}{4+\sqrt{2}}, \frac{4}{4+\sqrt{2}}]^T$$

可求得SVM的margin

$$margin = 1/||\omega|| = \frac{1}{2}\sqrt{4+\sqrt{2}}$$

(注：由于此时两个样本均为支持向量，也可由两样本映射至高维空间后，由空间距离的一半求得margin，即$margin = \frac{||(\omega^T \cdot \phi(x_1) + \omega_0) - (\omega^T \cdot \phi(x_2) + \omega_0)||}{2||\omega||} = \frac{1}{2}\sqrt{4+\sqrt{2}}$)

3.由第2问可知

$$\omega = [0, \frac{2^{\frac{5}{4}}}{4+\sqrt{2}}, \frac{4}{4+\sqrt{2}}]^T$$

4.对于支持向量，则有

$$y_i[(\omega^T \cdot \phi(x_i) + \omega_0] = 1$$

由于两样本均为支持向量，不妨把(x_1, y_1)代入

解得

$$\omega_0 = -1$$

故判别面方程关于x的显示表达式为

$$f(x) = \frac{2^{\frac{5}{4}}}{4+\sqrt{2}}\sqrt{x} + \frac{4}{4+\sqrt{2}}x^2 - 1$$

3. 利用支持向量机对MNIST数据集进行分类

a.数据预处理

```
#维数转换与归一化
dims = X_train.shape[0]
x_train=X_train.reshape(dims,1,784)
x_train=x_train.astype(np.float32)
x_train/=255.0

dims_test = X_test.shape[0]
x_test=X_test.reshape(dims_test,1,784)
x_test=x_test.astype(np.float32)
x_test/=255.0
#挑选数字"4"和"9"
x_data=[]
y_data=[]
for i in range(len(x_train)):
    if y_train[i]==4:
        x_data.append(x_train[i])
        y_data.append(1)
    if y_train[i]==9:
        x_data.append(x_train[i])
        y_data.append(-1)
for i in range(len(x_test)):
    if y_test[i]==4:
        x_data.append(x_test[i])
        y_data.append(1)
    if y_test[i]==9:
        x_data.append(x_test[i])
        y_data.append(-1)
x_data_new=np.array(x_data)
y_data_new=np.array(y_data)
#将30%作为测试集合
from sklearn.model_selection import train_test_split
x_train_new, x_test_new, y_train_new, y_test_new = train_test_split(x_data_new,
y_data_new, test_size=0.3, random_state=1)
```

将数字“4”标签处理为1，数字“9”标签处理为-1，得到新的数据集。将新的数据集70%划分为训练集，30%划分为测试集，对模型进行训练与评估。

b.训练集正确率

```
#使用线性核
from sklearn.svm import SVC
clf = SVC(kernel='linear',verbose=1)
clf.fit(x_train_new, y_train_new)

y_predition_train=clf.predict(x_train_new)
acc = sum(y_predition_train == y_train_new)/float(len(y_train_new))
print('linear train accuarcy: ' + str(acc))
```

此时训练集上准确率为98.39%，支持向量个数分别为377，381

```
#使用二次多项式核
clf = SVC(kernel='poly', degree=2,gamma=10,verbose=1)
```

得到训练集上准确率为100%，支持向量个数分别为393，363

```
#使用三次多项式核
clf = SVC(kernel='poly', degree=3,gamma=10,verbose=1)
```

得到训练集上准确率为100%，支持向量个数分别为426，429

```
#使用径向基核
clf = SVC(kernel='rbf', gamma=0.01,verbose=1)
```

得到训练集上准确率为99.12%，支持向量个数分别为652，638

```
#使用Sigmoid核
clf = SVC(kernel='sigmoid', gamma=0.005,verbose=1)
```

得到训练集上准确率为95.93%，支持向量个数分别为924，928

c.测试集正确率

```
y_predition_test=clf.predict(x_test_new)
acc = sum(y_predition_test == y_test_new)/float(len(y_test_new))
print('linear test accuarcy: ' + str(acc))
```

使用线性核时，测试集正确率为96.81%

使用二次多项式核时，测试集正确率为98.86%

使用三次多项式核时，测试集正确率为98.96%

使用径向基核时，测试集正确率为98.60%

使用Sigmoid核时，测试集正确率为96.01%

整理使用不同核函数时训练集与测试集表现如下:

核函数	训练集正确率	测试集正确率
线性核	98.39%	96.81%
二次多项式核	100%	98.86%
三次多项式核	100%	98.96%
径向基核	99.12%	98.60%
Sigmoid核	95.93%	96.01%

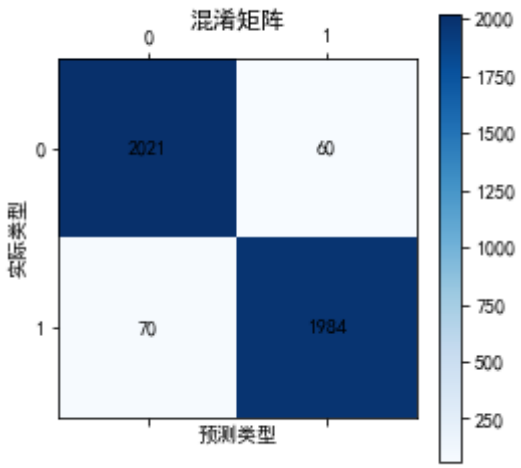
可以看出通过合理的参数设置，使用不同核函数时支持向量机方法在训练集合与预测集合上均能取得不错的分类效果，测试集合准确率在 96% 到 99% 之间，其中三次多项式核分类效果最好，二次多项式核与径向基核在测试集合上效果也不错。

通过核函数选取和参数设置可以发现，支持向量机在训练集上正确率普遍比测试集上略高，且支持向量数目的越多，模型训练与预测的时间越长。当使用径向基核将参数gamma设置为1时，模型在训练集上准确率为100%，但在测试集上准确率仅有50%。此时由于支持向量个数过多，模型会出现明显的过拟合现象。因此在选择模型与参数时，要综合考虑模型的复杂程度，减少不必要的计算开销，缩短模型训练时间且避免出现过拟合现象。

d.三种方法比较

挑选数字“4”与数字“9”组成新集合后，选取70%作训练集合，30%作测试集合。分别使用三种模型在相同的测试集合上测试，比较模型的分类效果。

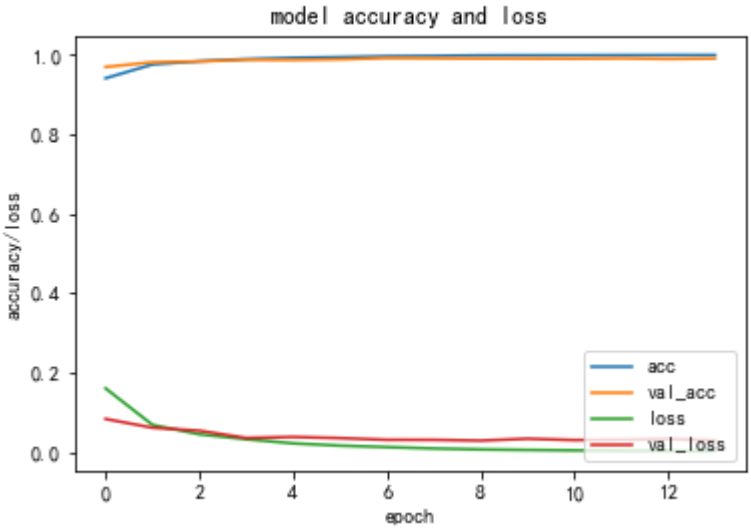
使用Logistic Regression模型用训练集合进行训练 、用测试集合进行评估，得到测试集合上混淆矩阵如下：



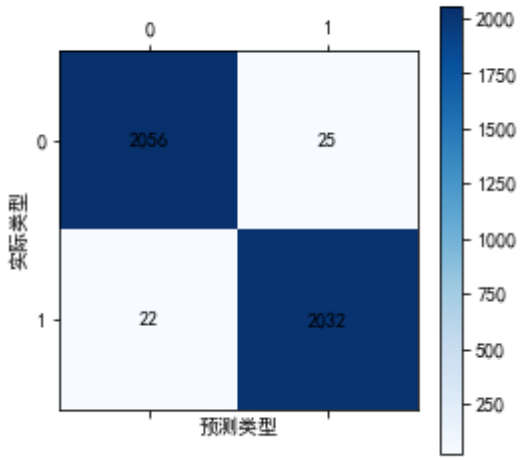
测试集合的正确率为96.92%

由第4次作业知，隐层节点数为100时神经网络分类效果最好，因而选用隐层节点为100的单隐层神经网络完成二分类任务。此时需要将标签处理为one-hot形式，并在训练集中划分一部分数据作为验证集。采用Early stop监视模型训练过程以防过拟合，当验证集损失在5个epochs内不再下降时停止训练。

模型学习曲线如下：



测试集合混淆矩阵如下：



此时测试集合的正确率为98.88%，可见神经网络具有较好的分类效果。

结合第3问对支持向量机方法进行的评估可知，对数字“4”“9”进行分类时，使用径向基核、二次多项式核、三次多项式核的支持向量机模型与神经网络模型的测试集合正确率很高，能达到98%以上；线性核、sigmoid核支持向量机模型与 Logistic Regression模型在测试集合上正确率相近，在 96% 以上。相比较而言，针对该分类问题，非线性方法比线性分类方法正确率更高 。

比较三类监督学习方法可知，Logistic Regression是线性分类方法，使用简单、模型可解释性强，但适用范围受限；神经网络是非线性分类器，功能强大，但模型可解释性差，需要大量训练数据，且一般需要设置验证集以防出现过拟合现象；支持向量机根据核函数的不同，可为线性或非线性分类器，具有良好的理论支撑，但训练数据过多时会导致计算开销过大。

官方文档说明由于支持向量机的拟合时间复杂度大于样本数的二次方，因此很难使用训练样本数大于10000的数据集。实际使用三种模型对数据进行分类时，使用支持向量机所消耗的时间远大于另外两种模型，且核函数越复杂耗时越长，很好地体现了支持向量机模型的特点。由于支持向量机的分类效果与核函数的选取密切相关，实际使用时应选择合适的核函数以取得较好的分类效果。