Assignment-1 Report

一、Assignment 概述

1. 概述

本次 assignment 在三个二维正态分布上分别随机抽取样本数据点，并标以相应的标签名，构成数据集。为简单实现 linear generative model，三个二维正态分布取相同的协方差矩阵。分别训练 linear generative model 和 linear discriminative model，对测试数据进行分类，观察准确率。

2. 文件解读

source.py：顶层文件，调用 data.py 和 model.py
model.py：定义 linear generative model 和 linear discriminative model
data.py：定义如何生成数据、读取数据

3. 使用说明

1）直接运行 python source.py 可自动生成训练数据、测试数据，完成模型的训练和数据测试，并输出测试准确率。
2）在 source.py 中可以修改传入参数的值，来调整模型的超参数
3）通过修改 data.py 的 generate_data()函数中 mean 和 cov 值，可以改变样本数据服从的高斯分布。

二、Linear Generative Model

根据贝叶斯理论，对于多分类问题，我们有给定数据点 x 时，其属于第 k 类的条件概率如下，其中$a_k = \ln[p(x|C_k)p(C_k)]$

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_k)}$$

假设数据点服从高斯分布，且有着相同的协方差 S，于是有：

$$a_k(x) = w_k^T x + b$$

其中$w_k = S^{-1}\mu_k$，$b = -\frac{1}{2}\mu_k S^{-1}\mu_k + \ln[p(C_k)]$

由最大似然方法，我们可以有如下估计：

$$\mu_k = \frac{1}{N_k}\sum x_k$$

$S = p(C_k) * S_k$，其中$S_k = \frac{1}{N_k}\sum(x_k - \mu_k)(x_k - \mu_k)^T$

于是我们可以对每个数据点 x，估计出每一个类别的条件概率$p(C_K|x)$
对于多分类问题，我们采取 argmax 策略，取条件概率最大的类别。

三、Linear Discriminative Model

使用 softmax 回归如下：

$$p(y=c|x)=\text{softmax}(w_c^T x)=\frac{\exp(w_c^T x)}{\sum \exp(w_c^T x)}$$

使用小批量梯度下降法，使用交叉熵损失函数对参数进行学习，迭代公式如下：

$W_{t+1}=W_t+\alpha(\frac{1}{N}\sum x(y-y')^T$，*其中$y'$是估计出的条件概率向量*

同样，对于多分类问题，我们采取 argmax 策略，取条件概率最大的类别。

四、测试结果

1. 训练数据集规模的影响

对于 linear discriminative model，规定学习率为 0.8，ecpoch=5，batch_size 则随训练数据集规模相应变化。改变训练数据集大小，两种模型的准确率如下图所示。

```
test 0
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.7666666666666667
test 1
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.7666666666666667
test 2
------Generative Model Accuracy:  0.43333333333333335
------Discriminative Model Accuracy:  0.7
```

size=3*10

```
test 0
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.7166666666666667
test 1
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.7633333333333333
test 2
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.71
```

size=3*100

```
test 0
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.6756666666666666
test 1
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.7766666666666666
test 2
------Generative Model Accuracy:  1.0
------Discriminative Model Accuracy:  0.6756666666666666
```

size=3*1000

有如下发现：

1）对于 generative model，其需要学习大量参数。因此当数据集较小时，参数估计不准确，其准确率将产生较大摆动。而当数据集较大时，它可以体现出良好的准确率。

2）对于 discriminative model，可以发现，数据规模的变化，对分类的准确率的影响比 generative model 要小。不足的是，该模型的准确率较低，大约在$\sqrt{2}/2$左右起伏。分析认为，对于二维高斯分布数据点的分类问题，在原数据空间中决策边界是非线性的，在分类时应当使用基函数 Φ(x)来将决策边界改变成线性的

超平面。而在本次 assignment，为方便处理，并没有使用基函数，从而使得准确率较低。此外，存在着过拟合现象，可通过正则化方式来限制过拟合现象。

2. 协方差不一致的影响

在 generative model 中，为简单起见，我们假定三个二维正态分布有着相同的协方差。在构造数据集时，我们也遵循了这一假设。改变协方差，使得三者不一致，我们观察两种模型的准确率，取训练数据集大小为 3*1000，结果如下图。

```
test 0
------Generative Model Accuracy:  0.3333333333333333
------Discriminative Model Accuracy:  0.743
test 1
------Generative Model Accuracy:  0.3333333333333333
------Discriminative Model Accuracy:  0.67
test 2
------Generative Model Accuracy:  0.3333333333333333
------Discriminative Model Accuracy:  0.6963333333333334
```

可见，generative model 的准确率是 1/3，几乎相当于随机分类的平均准确率。因此，协方差一致的假设下的解并不适用于协方差不一致的情况，需要重新计算参数学习的函数。对于 discriminative model，准确率同协方差不一致时则无明显差距。

五、总结

1. 对于 generative model，需要足够的数据来训练大量参数，且良好估计数据的分布情况，此时准确率高。
2. 对于 discriminative model，参数较少，受训练数据集较小的不良影响较小。在此次 assignment 的基础上，需要引入基函数改善决策边界和正则化限制过拟合，以提高准确率。