PRML Assignment1 Report

Part I 数据生成

通过*numpy.random*中的*multivariate_normal*即可在给定的高斯分布上取样。

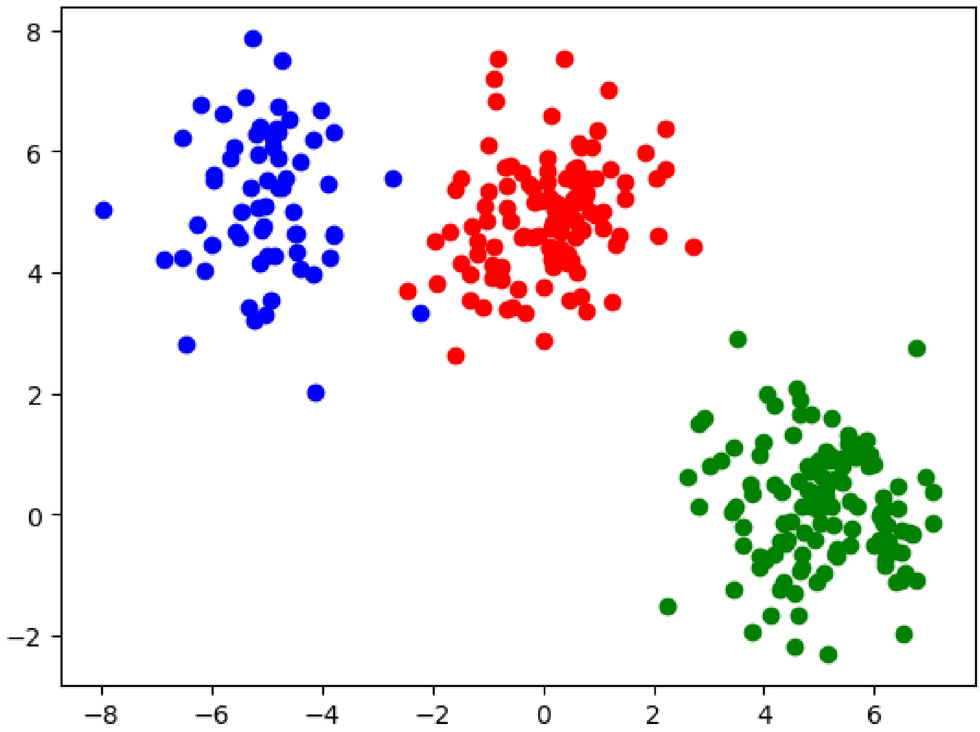
三个高斯分布的均值分别为

$$mean1 = (-5, 5)$$
$$mean2 = (0, 5)$$
$$mean3 = (5, 0)$$

三个高斯分布的协方差矩阵均设置为

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

分别从这三个高斯分布中进行采样，分别采样700，1100，1200个，按9:1分成训练集和测试集



三个高斯分布的采样如上图所示。

Part II 模型

判别式模型

使用判别式模型的时候，我们希望对于每个类别y直接得到$P(y|x)$，从而选择使得概率$P(y|x)$最大的y作为答案。

这里是线性的判别式模型，即$f(X) = WX + b$，$f(X) \in R^{N,3}$，其中N为数据组数。

为了更简洁的表示，给X和W增加一个常数维度，省去b，变成

$$f(X) = WX$$

实际实现中，构造的数据$X \in R^{N,2+1}$，为了方便计算，将模型改写为

$$f(X) = XW$$

$f(X)$的意义是，中的每个坐标，对于3个组别分别的得分，进而可以得到

$$P(y|X) = softmax(f(X))$$

其中

$$softmax(x) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

训练方法

使用交叉熵损失函数

$$L = -\sum \log P(y|x)$$

其中y是对于x的正确分类，$P(y|x)$是上述学到的$softmax(f(X))$。

采用sgd来训练该模型，那么大致的思路为

1. 从输入数据(X, y)中采样得到(X_s, y_s)
2. 计算$L = -\sum \log P(y_s, X_s)$
3. 计算参数W关于L的梯度 $grad = \frac{dL}{dW}$
4. 更新$W = W - lr \cdot grad$

以上步骤重复若干次。

对于一组数据(x, y)

$L = -\log P(y|x) = -\log softmax(xW) = -(xW)_y + \log \sum_i e^{(xW)_i}$

$\frac{dL}{dW_{i,j}} = \frac{dL}{d(xW)} \cdot \frac{d(xW)}{dW}$

当$j = y$时，$\frac{dL}{d(xW)_j} = -1 + \frac{e^{(xW)_j}}{\sum_k e^{(xW)_k}}$

当$j \neq y$时，$\frac{dL}{d(xW)_j} = \frac{e^{(xW)_j}}{\sum_k e^{(xW)_k}}$

即

$$\frac{dL}{d(xW)_j} = -[j = y] + \frac{e^{(xW)_j}}{\sum_k e^{(xW)_k}}$$

其中M为种类数

$$\frac{d(xW)_j}{dW_{i,j}} = x_i$$

$$

合起来可以得到梯度

$$\frac{dL}{dW_{i,j}} = (-[j=y] + \frac{e^{(xW)_j}}{\sum_k e^{(xW)_k}}) \cdot x_i$$

对于多组数据，将梯度累加起来即可。

生成式模型

这里我们用二维的高斯分布作为分布模型，对于每一个类别建立一个高斯分布。

根据贝叶斯公式：$P(y|x) \propto P(y) \cdot P(x|y)$

通过每个类别我们自己建立的高斯分布，可以得到对于分布y，x的似然函数 $P(x|y)$

通过对输入数据的统计，我们可以统计每个类别的占比，将此作为每个类别的先验概率$P(y)$

那么$argmax_y P(y|x)$就认为是x的类别。

具体的，对于每个高斯分布，由平均值向量$mean$和协方差矩阵cov组成

$mean$由对应标号的坐标求平均值代替，cov由对应标号的坐标集合求协方差代替

即

$$mean_c = \frac{\sum_{y_j=c} Xj}{\sum_{y_j=c} 1}$$

$$cov_c = \frac{\sum_{y_j=c} (X_j - mean_c)^T (X - mean_c)}{\sum_{y_j=c} 1}$$

对比

生成式模型：

正确率：99.6%

需要有一个符合输入数据的分布模型

可以生成新的数据

参数较少

判别式模型

正确率：99.0%

不需要一个输入数据的分布模型

只能做分类问题，不能生成新的数据

参数更多更复杂

由于三个高斯分布距离较远，判别的正确率都非常高，没有区分度。

Part III 调整分布

调整1

调整均值为

$$
\begin{aligned}
mean1 &= (-2, 2) \\
mean2 &= (0, 2) \\
mean3 &= (2, 0)
\end{aligned}
$$

其他不变，得到结果

生成式模型：86.0%

判别式模型：81.0%

结论：

当重叠的部分增多的时候，生成式模型和判别式模型的差距开始拉开，生成式模型在有重叠的情况下的表现更好。

调整2

调整均值为

$$
\begin{aligned}
mean1 &= (0, 0) \\
mean2 &= (0, 0) \\
mean3 &= (0, 0)
\end{aligned}
$$

调整协方差为

$$
cov1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}
$$
$$
cov2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
$$
$$
cov3 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}
$$

生成式模型：76.7%

判别式模型：34.3%

结论

当均值相同，只有协方差不同的时候，可以知道判别式模型的效果非常差，几乎和随便猜的期望正确率相近，而生成式模型还有一定的判别能力，可以做到76.7%的正确率。

代码运行

```
python source.py
```