

# Answer of Assignment 6

## PCA & SVM

2020E8017782032\_ 蒲尧

### 第一部分：计算与证明

1. 有  $N$  个样本  $x_1, \dots, x_N$ ，每个样本维数  $D$ ，希望将样本维数降低到  $K$ ，请给出 PCA 算法的计算过程。
2. 根据自己的理解简述结构风险最小化与 VC 维。
3. 请推导出 Hard-Margin SVM 的优化目标。
4. 请解释出 Hinge Loss 在 SVM 中的意义。
5. 简述核方法的基本原理。

### 第二部分：计算机编程

6. 从 MNIST 数据集中任意选择两类，对其进行 SVM 分类，可调用现有的 SVM 工具如 LIBSVM，展示超参数  $C$  以及核函数参数的选择过程。

### 第一部分回答

1. PCA 算法的计算过程：

(1) 计算数据的均值： $\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$ ;

(2) 计算样本的协方差矩阵： $S = \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T$ ;

(3) 做  $D \times D$  维矩阵  $S$  的特征值分解;

(4) 提取其中最大的  $K$  个特征值对应的最大的  $K$  个特征向量  $\vec{u}_1, \dots, \vec{u}_K$ , ( $s.t. \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{K-1} \geq \lambda_K \geq 0$ ), 得到  $D \times K$  大小的映射矩阵  $U = [\vec{u}_1, \vec{u}_2, \dots, \vec{u}_K]$ ;

(5) 将每个样本进行映射： $z_n = U^T x_n$ ，得到的  $z_n$  是  $K \times 1$  的向量。

2. (1) 结构风险最小化：在未知的测试集上的错误率达到最小化。 $Test\ error\ rate \leq train\ error\ rate + f(N, h, p)$
- (2) VC 维：某个空间中给  $n$  个样本点随机打标签，如果某个模型足够强大能够将其分开，则增加样本个数为  $n+1$ ，直到不能分开  $n+1$ ，则这个最大的样本点数  $n$  为该空间的 VC 维。

### 3. Hard-Margin SVM 的优化目标

点  $\vec{x}_i$  到直线  $\vec{w}\vec{x} + \vec{b} = 0$  的距离为:

$$d(\vec{x}_i) = \frac{|\vec{w}\vec{x}_i + \vec{b}|}{\sqrt{\|\vec{w}\|_2^2}}$$

$$\begin{cases} \vec{w}\vec{x}_i + \vec{b} \geq 0, & y_i = 1; \\ \vec{w}\vec{x}_i + \vec{b} \leq 0, & y_i = -1 \end{cases}$$

我们的目标是最大化到直线最近点的距离  $d$ , 从而推出参数  $\vec{w}, \vec{b}$ :

$$\begin{aligned} & \max_{\vec{w}, \vec{b}} \min_{\vec{x}_i \in \mathbf{D}} d \\ &= \max_{\vec{w}, \vec{b}} \min_{\vec{x}_i \in \mathbf{D}} \frac{|\vec{w}\vec{x}_i + \vec{b}|}{\sqrt{\|\vec{w}\|_2^2}} \\ & s.t. \forall \vec{x}_i \in \mathbf{D} : y_i (\vec{w}\vec{x} + \vec{b}) \geq 0 \end{aligned}$$

我们采用如下策略:

$$\forall \vec{x}_i \in \mathbf{D} : |\vec{w}\vec{x} + \vec{b}| \geq 1$$

我们可以推出:

$$\min_{\vec{x}_i \in \mathbf{D}} \frac{|\vec{w}\vec{x}_i + \vec{b}|}{\sqrt{\|\vec{w}\|_2^2}} \geq \min_{\vec{x}_i \in \mathbf{D}} \frac{1}{\sqrt{\|\vec{w}\|_2^2}} = \frac{1}{\sqrt{\|\vec{w}\|_2^2}} = \frac{1}{\|\vec{w}\|}$$

直线两侧都有点, 距离  $\times 2$ , 得到最终优化目标:

$$\max \frac{2}{\|\vec{w}\|}$$

### 4. 对于 Soft-margin SVM, 优化目标可以写成:

$$\min_{\vec{w}, b} \frac{\|\vec{w}\|^2}{2} + C \sum_{i=1}^m \mathcal{L}_{0/1}(y_i (\vec{w}^T \vec{x}_i + b) - 1)$$

其中,  $C > 0$  是一个常数, 作为惩罚因子,  $\mathcal{L}_{0/1}$  是 “0/1 损失函数”:

$$\mathcal{L}_{0/1}(z) = \begin{cases} 1, & \text{if } z < 0; \\ 0, & \text{otherwise} \end{cases}$$

Hinge Loss:

$$\mathcal{L}_{hinge}(z) = \max(0, 1 - z)$$

Hinge Loss 的零区域对应的正是非支持向量的普通样本, 从而所有的普通样本都不参与最终超平面的决定; 只考虑支持向量。这正是支持向量机最大的优势所在, 对训练样本数目的依赖大大减少, 从而提高了训练效率。

### 5. 核方法: 回顾前面得到的公式:

$$\begin{aligned} & \max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^n \alpha_i \alpha_j y_i y_j x_i x_j \\ & s.t. C \geq \alpha_i \geq 0, \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

我们看到计算目标函数时，需要知道  $x$  的内积。我们引入核函数，一个可以应用于一对输入数据以计算对应特征空间中的内积的函数。对于不能线性分割的数据集，核函数可以将数据升高维度，进行高纬度的内积，寻找高纬度数据的相似性和高纬度的线性分界面。

## 第二部分回答

代码见如下文件 [Python 文件](#)。由图可知：C 过小，欠拟合，准确度较小；C 过大，过拟合，Test 和 Train 的准确率相差会变大。gamma 过小准确度较小，gamma 越大准确度相对变大。

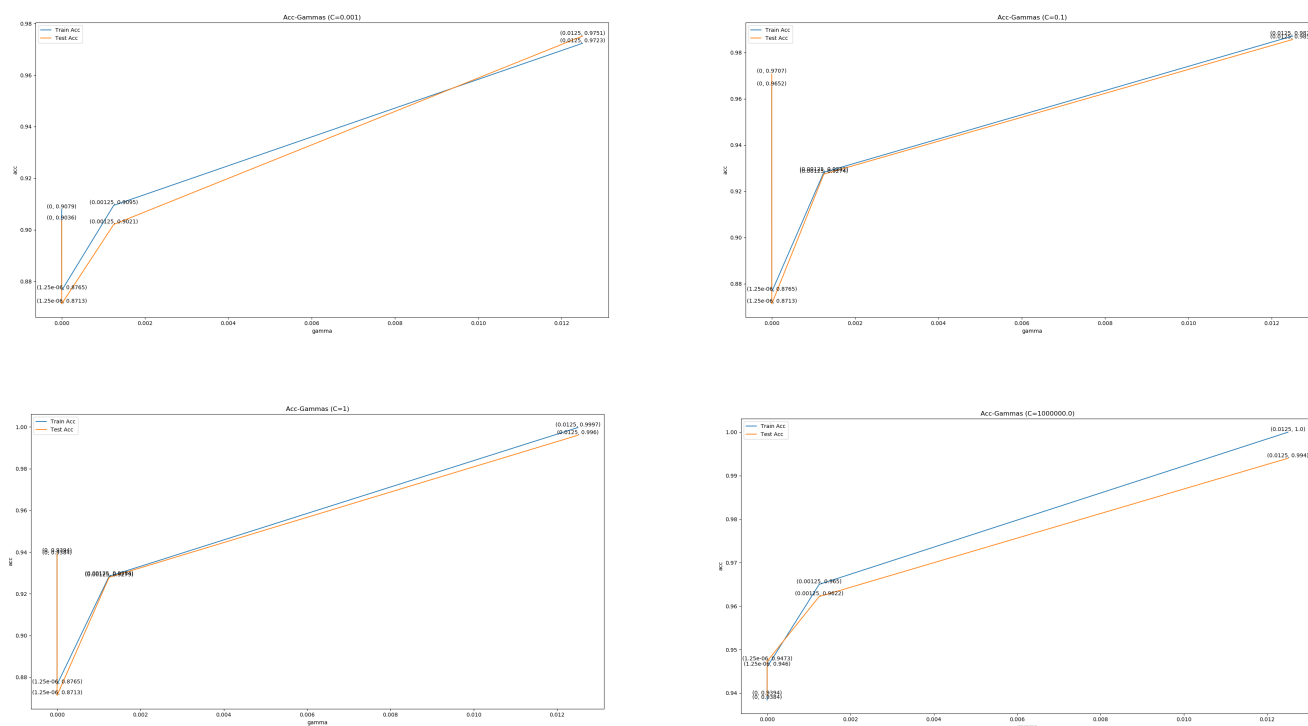


图 1: Acc-gamma-C