机器学习与数据挖掘实验报告

梁睿凯*

中山大学计算机学院

March 29, 2022

*电子邮件: liangrk5@mail2.sysu.edu.cn，学号: 19335121

目录

1 实验问题 **4**

2 实验要求 **4**

3 实验过程 **4**

3.1 SVM 一般理论 4

3.1.1 简介 4

3.1.2 算法基本思想 4

3.1.3 学习对偶算法 5

3.1.4 软间隔最大化 6

3.1.5 核技巧 6

3.2 采用不同核函数的模型和性能比较及分析 7

3.2.1 线性核函数 7

3.2.2 高斯核函数 7

3.2.3 性能比较及分析 8

3.3 线性分类模型 8

3.3.1 hinge loss 线性分类模型 8

3.3.2 cross-entropy loss 线性分类模型 10

3.3.3 hinge loss 对比 cross-entropy loss 11

3.4 训练过程 12

3.4.1 初始化方法 12

3.4.2 超参数参数选择 12

3.4.3 训练技巧 13

4 实验结果 **13**

4.1 结果及分析 13

4.2 讨论及心得 13

1 实验问题

根据提供的数据，训练一个采用在不同的核函数的支持向量机 SVM 的 2 分类器，并验证其在测试数据集上的性能。

2 实验要求

1. 考虑两种不同的核函数：线性核函数和高斯核函数
2. 可以直接调用现成 SVM 软件包来实现
3. 手动实现采用 hinge loss 线性分类模型和 cross-entropy loss 线性分类模型，并比较他们的优劣。

3 实验过程

3.1 SVM 一般理论

3.1.1 简介

支持向量机 (SVM) 是一种二类分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使他有别于感知机；支持向量机还包含核技巧，这使它成为实质上的非线性分类器，支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

当训练数据线性可分时，通过硬间隔最大化，学习一个线性的分类器，即线性可分支持向量机；当训练数据近似线性可分时，通过软间隔最大化，也学习一个线性的分类器，即线性支持向量机；当训练数据线性不可分时，通过使用核技巧及软间隔最大化，学习非线性支持向量机。

3.1.2 算法基本思想

支持向量机的基本思想是求解能够正确划分训练数据集并且几何间隔最大的分离超平面，对线性可分的训练数据集而言，线性可分分离超平面有无穷多个（等价于感知机，但是几何间隔最大的分离超平面是唯一的。这里的间隔最大化又称为硬间隔最大化。

求得一个几何间隔最大的分离超平面，即最大间隔分离超平面可以表示为以下约束最

优化问题：

$$\begin{aligned} \max_{w,b} \quad & \gamma \\ s.t. \quad & y_i(\frac{\omega}{||\omega||} \cdot x_i + \frac{b}{||b||}) \geq \gamma, \qquad i = 1, 2, \cdots, N \end{aligned} \tag{1}$$

即我们希望最大化超平面 (ω, b) 关于训练数据集的几何间隔 γ，约束条件表示的是超平面 (ω, b) 关于每个训练样本点的几何间隔至少是 γ, 考虑到几何间隔和函数间隔的关系，可将问题改写为

$$\begin{aligned} \max_{w,b} \quad & \frac{\hat{\gamma}}{||\omega||} \\ s.t. \quad & y_i(\omega \cdot x_i + b) \geq \hat{\gamma}, \qquad i = 1, 2, \cdots, N \end{aligned} \tag{2}$$

函数间隔 $\hat{\gamma}$ 的取值并不影响最优化问题的解，因此可以取其为 1 带入，同时最大化 $\frac{1}{||\omega||}$ 和最小化 $\frac{1}{2}||\omega||^2$ 是等价的，于是得到线性可分支持向量机学习的最优化问题

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||\omega||^2 \\ s.t. \quad & y_i(\omega \cdot x_i + b) - 1 \geq 0, \qquad i = 1, 2, \cdots, N \end{aligned} \tag{3}$$

由此即可得到最大间隔分离超平面和分类决策函数。

3.1.3 学习对偶算法

为了求解线性可分支持向量机的最优化问题，将它作为原始最优化问题，应用拉格朗日对偶性，通过求解对偶问题得到原始问题的最优解，这就是线性可分支持向量机的对偶算法。定义拉格朗日函数如下：

$$L(\omega, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i y_i (\omega \cdot x_i + b) + \sum_{i=1}^{N} \tag{4}$$

根据拉格朗日函数对偶性，原始问题的对偶问题是极大极小问题，故为了得到对偶问题的解，需要先求 $L(\omega, b, a)$ 对 ω, b 的极小，再求对 α 的极大，最终得到等价的对偶优化问题:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha \\ s.t. \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \qquad i = 1, 2, \cdots, N \end{aligned} \tag{5}$$

3.1.4 软间隔最大化

线性可分问题的支持向量机学习方法，对线性不可分训练数据是不适用的，因为这时上述方法中的不等式约束并能都成立，此时就需要修改硬间隔最大化使其成为软间隔最大化。线性不可分意味着某些样本点 (x_i, y_i) 不能满足函数间隔大于等于 1 的约束条件。为了解决这个问题，可以对每个样本点引进一个松弛变量 $\xi \geq 0q$，使函数间隔加上松弛变量大于等于 1，这样约束条件变为

$$y_i(\omega \cdot x_i + b) \geq 1 - \xi \tag{6}$$

同时对每个松弛变量 ξ_i，支付一个代价 xi_i，目标函数变为

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i \tag{7}$$

这里，$C > 0$ 称为惩罚参数，一般由应用问题决定，C 值大时对误分类的惩罚增大，C 值小时对误分类的惩罚减小，线性不可分的线性支持向量机的学习问题变成如下凸二次规划问题 (原始问题)：

$$\begin{aligned} \min \quad & \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i \\ s.t. \quad & y_i(\omega \cdot x_i + b) \geq 1 - \xi, \qquad i = 1, 2, \cdots, N \\ & \xi \geq 0, \qquad i = 1, 2, \cdots, N \end{aligned} \tag{8}$$

同理可得对偶问题：

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \\ s.t. \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & C - \alpha_i - \mu_i = 0 \\ & \alpha_i \geq 0 \\ & \mu_i \geq 0, \qquad i = 1, 2, \cdots, N \end{aligned} \tag{9}$$

3.1.5 核技巧

对解线性分类问题，线性分类支持向量机是一种非常有效的方法。但是有时分类问题是非线性的，这时可以使用非线性支持向量机。

核技巧应用到支持向量机，其基本想法就是通过一个非线性变换将输入空间对应于一个特征空间，使得在输入空间中的超曲面模型对应于特征空间中的超平面模型，这样分类问题的学习任务通过在特征空间中求解线性支持向量机就可以完成。常用的核函数包括：

表 1: 常用核函数

名称	表达式	参数
线性核	$k(x_i, x_j) = x_i^T x_j 1$	
多项式核	$k(x_i, x_j) = (x_i^T x_j)^d$	$d \geq 0$ 为多项式的次数
高斯核	$k(x_i, x_j) = \exp(-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2})$	$\sigma > 0$ 为高斯核的带宽
拉普拉斯核	$k(x_i, x_j) = \exp(-\frac{\|\|x_i - x_j\|\|}{\sigma})$	$\sigma > 0$
sigmod 核	$k(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$	tanh 为双曲正切函数，$\beta > 0, \theta < 0$

3.2 采用不同核函数的模型和性能比较及分析

3.2.1 线性核函数

采用线性核函数模型进行分类得到结果如下图。

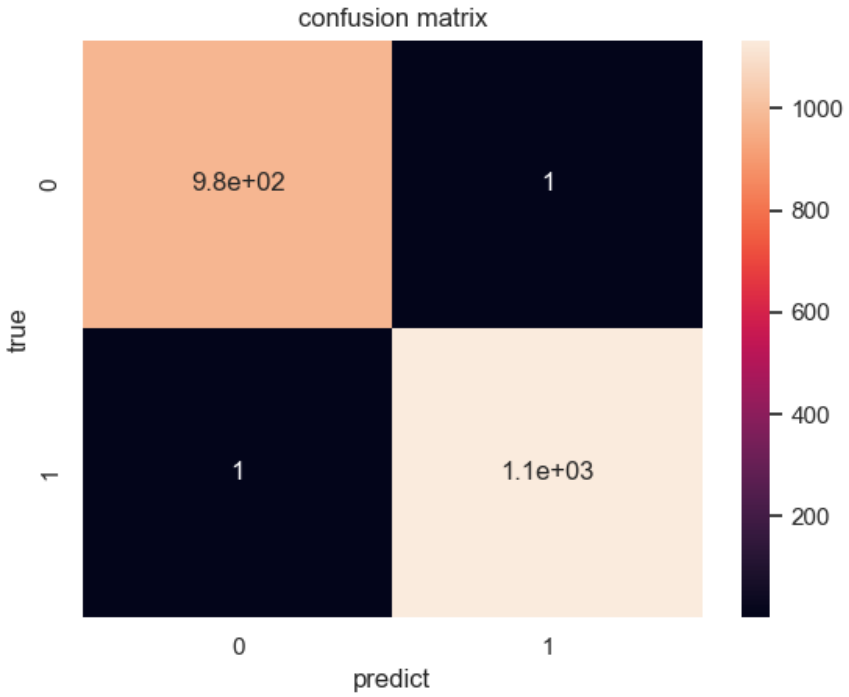


图 1: 线性核函数结果

3.2.2 高斯核函数

采用高斯核函数模型进行分类得到结果如下图。

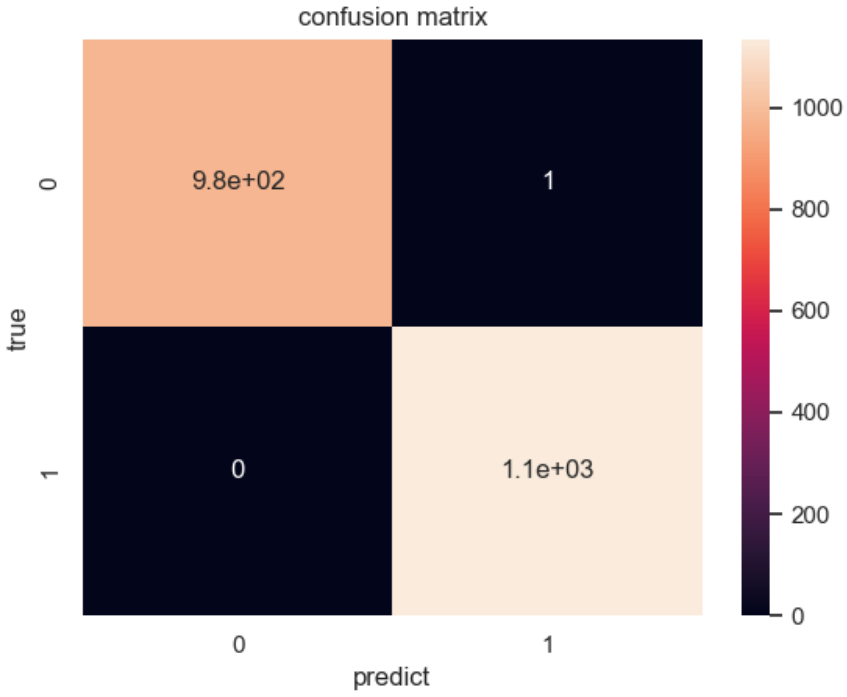


图 2: 高斯核函数结果

3.2.3 性能比较及分析

通过比较可以发现采用线性核函数和高斯核函数均可以得到近乎完美的效果，高斯核函数略优于线性核函数，因此可以判断数据集线性可分，分析是因为图片数据是 28×28，特征维数较高，在这种情况下往往线性可分。

对于这两种不同的核函数，通常情况下，线性核函数用于线性可分的情形，它具有参数少，速度快的优点，对于一般的数据，分类效果已经很理想。而高斯核函数主要用于线性不可分的情形，参数多，分类效果依赖于参数的设置。在选择核函数的问题上要根据具体情况，多尝试不同的核不同的参数。

3.3 线性分类模型

3.3.1 hinge loss 线性分类模型

对于线性支持向量机学习来说，其模型为分离超平面 $w^* \cdot x + b^* = 0$ 及决策函数 $f(x) = sign(w^* \cdot x + b^*)$，其学习策略为软间隔最大化，学习算法为凸二次规划，线性支持

向量机学习还有另一种解释，就是最小化以下目标函数：

$$\sum_{i=1}^{N}[1-y_i(\omega\cdot x_i+b)]_+ + \lambda||\omega||^2 \tag{10}$$

目标函数的第一项是经验损失或经验风险，函数

$$L(y(\omega\cdot x+b)) = [1-y(\omega x+b)]_+ \tag{11}$$

称为折页损失函数 (hinge loss function)。下标“+”表示以下取正值的函数

$$[z]_+ = \begin{cases} z, & z>0 \\ 0, & z\leq 0 \end{cases} \tag{12}$$

这就是说，当样本点被正确分类且函数间隔（确信度）$y_i(\omega\cdot x_i+b)$ 大于 1 时，损失是 0，否则损失是 $1-y_i(\omega\cdot x_i+b)$。

折叶损失函数的图形如图3所示，横轴是函数间隔 $y(\omega\cdot x+b)$，纵轴是损失，由于函数形状像一个折页，故名折页损失函数。图中还画出了 0-1 损失函数，可以认为他是二分类问题的真正的损失函数，而折页损失函数是 0-1 损失函数的上界，由于 0-1 损失函数不是连续可导的，直接优化由其构成的目标函数比较困难，可以认为线性支持向量机是优化由 0-1 损失函数的上界（折页损失函数）构成的目标函数，这时的上街损失函数又称为代理损失函数。折页损失函数不仅要分类正确，而且确信度足够高时损失才是 0，也就是说折页损失函数对学习有更高的要求。

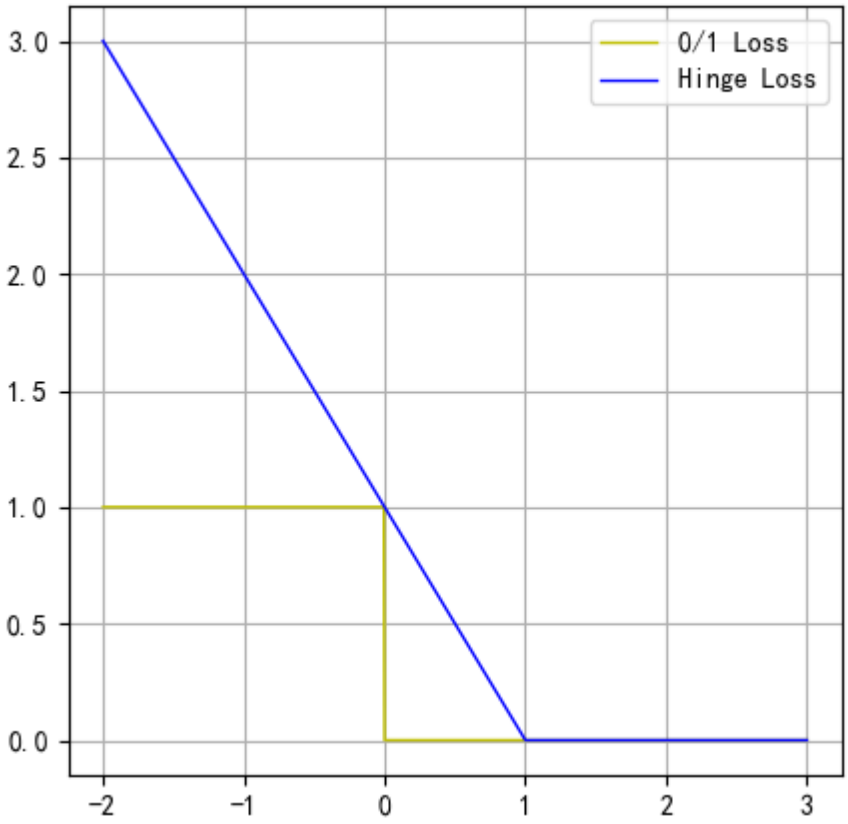


图 3: 折页损失函数

3.3.2 cross-entropy loss 线性分类模型

cross-entropy loss 线性分类模型的预测 $h_\theta(x_i)$ 由 $\sigma(Wx_i+b)$ 给出。$\sigma(Wx_i+b)$ 产生一个介于 0 和 1 之间的值，该值可以解释为示例 x_i 属于哪个类别的概率。如果该概率小于 0.5，我们将其分类为负类，否则将其分类为正类，这意味着我们可以写下观察负面或正面实例的概率：

$$p(y_i=1|x_i)=h_\theta(x_i) \qquad and \qquad p(y_i=0|x_i)=1-h_\theta(x_i) \tag{13}$$

假设数据是独立且均匀分布的，则

$$L(x,y)=\prod_{i=1}^{N}[h_\theta(x_i)]^{y_i}[1-h_\theta(x_i)]^{(1-y_i)} \tag{14}$$

采用上述表达式的对数，并使用 log 的属性进行简化，最后反转整个表达式得到交叉熵损失函数：

$$J = -\sum_{i=1}^{N} y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i)) \tag{15}$$

3.3.3 hinge loss 对比 cross-entropy loss

两种损失函数图像绘制如下：

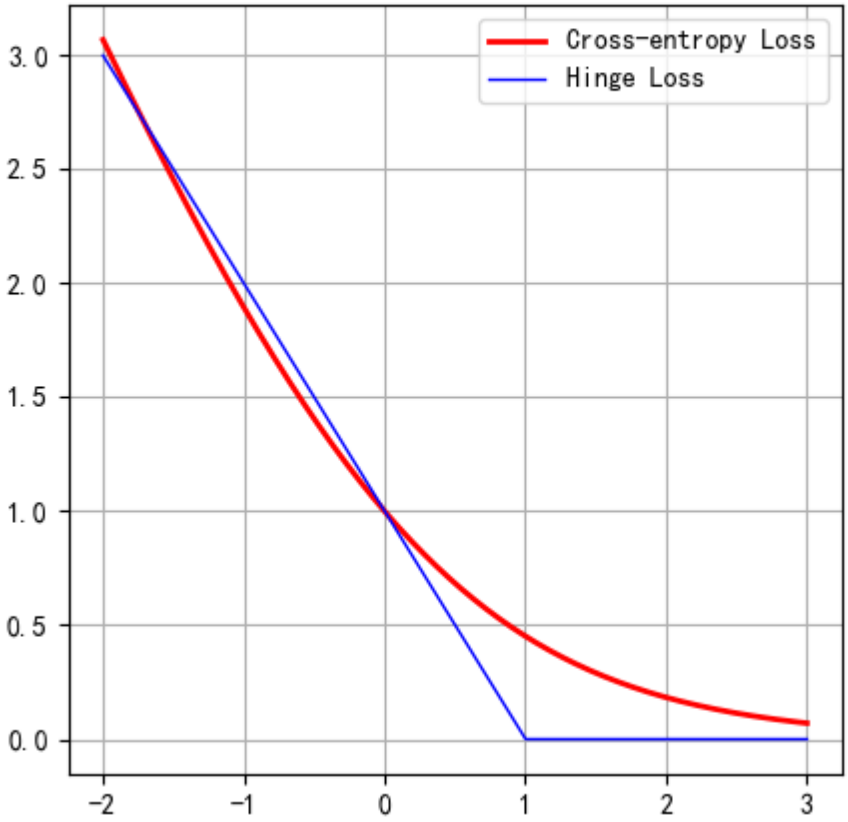
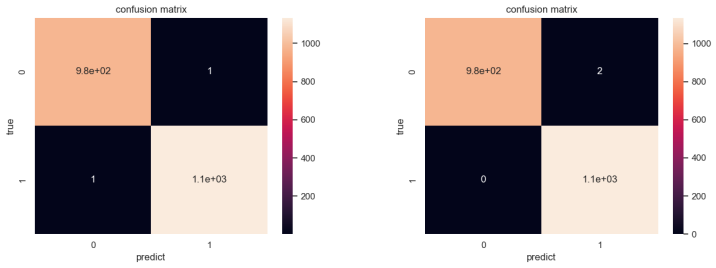


图 4: Hinge loss vs Cross-entropy loss

二者的区别在于前者来自试图最大化决策边界和数据点之间的间隔——从而试图确保对每个点进行正确且可靠的分类，而后者来自模型参数的最大似然估计。softmax 函数的得分由交叉熵函数损失使用，它使我们能够将模型的得分解释为彼此之间的相对概率。例如（未归一化）分数是 $[10, 8, 8]$ 和 $[10, -10, -10]$, 其中第一类是正确的，则交叉熵损失会比 hinge loss 要高的多。实际上，（多类别）hinge loss 会认识到正确的类别分数在边界上已经超过了其他分数，因此它将在两个分数上均给出零损失，一旦满足边界要求，SVM 将不

再优化权重以尝试做得更好。

手动实现线性分类模型，分别采取这两种损失函数，采用 SGD 随机梯度下降法更新参数。batch_size 设置为 100, 迭代次数设置为 2000。结果如下：



(a) hinge_loss (b) cross-entropy_loss

图 5: 分类结果

可以看到二者在其数据集上训练的结果均较为理想，而在训练过程中可以发现采用 cross-entropy loss 的线性分类模型训练时间较长。

3.4 训练过程

3.4.1 初始化方法

从文件中读取训练集和测试集数据，并将其转换为 np.array 形式，将 label 中的 0 标签转变为-1 标签。初始化一个 SVC 对象并选取核函数。

3.4.2 超参数参数选择

可供选择的超参数如下：

- C: 惩罚系数，用来控制损失函数的惩罚系数，类似于 LR 中的正则化系数。C 越大，相当于惩罚松弛变量，希望松弛变量接近 0，即对误分类的惩罚增大，趋向于对训练集全分对的情况，这样会出现训练集测试时准确率很高，但泛化能力弱，容易导致过拟合。C 值小，对误分类的惩罚减小，容错能力增强，泛化能力较强，但也可能欠拟合。

- gamma: 核函数系数，该参数是 rbf，poly 和 sigmoid 的内核系数；默认是'auto'，那么将会使用特征位数的倒数，即 $1/n_features$。（即核函数的带宽，超圆的半径）。gamma 越大，σ 越小，使得高斯分布又高又瘦，造成模型只能作用于支持向量附近，可能导

致过拟合；反之，gamma 越小，σ 越大，高斯分布会过于平滑，在训练集上分类效果不佳，可能导致欠拟合。

C 和 gamma 的最佳组合通常通过在 C 和 gamma 为指数增长序列下网格搜索来选取，例如 $C \in \{2^{-5}, 2^{-3}, \cdots, 2^{1}3, 2^{1}5\}$；$\gamma \in \{2^{-15}, 2^{-13}, \cdots, 2^{1}, 2^{3}\}$。通常情况下，使用交叉验证来检查参数选择的每一个组合，并选择具有最佳交叉验证精度的参数。而在本实验中，通过上述方法发现参数的选择对本次实验结果影响不大（因为在测试集上的表现效果太好）。

3.4.3 训练技巧

shrink：启发式算法。如果能预知那些变量对应着支持变量，则只需要在这些样本上训练就够了，其他样本可不予考虑，这不影响实验结果，但降低了问题的规模并有助于迅速求解。进一步，如果能预知哪些变量在边界上（即 a=C），则这些变量可保持不动，则对其他变量进行优化吗，从而使问题的规模更小，训练时间大大降低，这就是 Shrink 技术。

4 实验结果

4.1 结果及分析

对于不同的惩罚系数 C，不同的核函数系数，采用线性核函数和高斯核函数得到的分类正确率均无限接近 100%（即在训练集上分类正确率为 100%，在测试集上分类错误样本数为 1$\tilde{3}$）。分析原因可能是本次实验的数据集过于完美（线性可分），且训练集与测试集差距不大导致的结果。总结而言，一般来说采用线性核函数的 SVM 在线性可分的数据上进行效果较好，而采用高斯核函数的 SVM 在线性可分的数据上表现并不会优于线性核函数，而在线性不可分的数据上，高斯核函数的表现则明显优于线性核函数。

4.2 讨论及心得

支持向量机（SVM ）是一种比较好的实现了结构风险最小化思想的方法。它的机器学习策略是结构风险最小化原则为了最小化期望风险，应同时最小化经验风险和置信范围。它是专门针对有限样本情况的学习机器，实现的是结构风险最小化：在对给定的数据逼近的精度与逼近函数的复杂性之间寻求折衷，以期获得最好的推广能力。它最终解决的是一个凸二次规划问题，从理论上说，得到的将是全局最优解，解决了在神经网络方法中无法避免的局部极值问题。它将实际问题通过非线性变换转换到高维的特征空间，在高维空间中构造线性决策函数来实现原空间中的非线性决策函数，巧妙地解决了维数问题，并保证了有较好的推广能力，而且算法复杂度与样本维数无关。当然 SVM 也存在观测样本很多时效率不高的缺点。

本次实验学习使用了 SVM 解决二分类问题，对课上学习到的理论知识有了更深入的理解与实践应用，进一步激发了我对机器学习与人工智能的兴趣。

参考文献

[1] https://zhuanlan.zhihu.com/p/49331510

[2] 李航. 统计学习方法 [M]. 北京: 清华大学出版社，2012：95-130