天津大学

模式识别与深度学习课程

实验 1、SVM 分类算法实验报告

学　　院　智能与计算学部

专　　业　计算机与科学技术

学　　号　3019244140

姓　　名　郭思齐

1．实验目标

作为模式识别与深度学习课程的第一个实验-SVM 分类，实验整体要求我们能够熟练掌握 SVM 分类算法原理及代码实现，能够对已实现算法进行灵活调用获取分类结果；能够熟练掌握线性和非线性 SVM 算法，并且根据实验数据的特性调节线性函数或核函数。具体下分两个小实例任务：

1. 在理解 SVM 鸢尾花分类算法原理的基础上，调节不同的参数，汇报不同参数对实验结果的影响，对结果进行分析，对实验结果影响大的参数和背后的原因。
2. 使用肿瘤数据进行**肿瘤分类**的实验，汇报训练集以及测试集的准确率。调节参数，整理不同的参数与实验结果的关系。对结果进行分析，对实验结果影响比较大的参数和原因。

2. 实验一

2.1 算法实现及参数调节说明

SVM 算法核心代码的说明：

处理 SVM 问题的过程与其他的机器学习实验过程类似，提出假设，配置环境，了解数据，数据预处理，划分训练集、测试集，利用模型训练测试，结果分析可视化。

对与算法核心部分，其实就是 SVM 支持向量机模型搭建与训练，用到了 libsvm 库中的 SVC 类，以及该类的训练 fit，测试 score 函数。核心 SVC 类拥有多个可调节参数；该实验关注以下三个参数：表示错误项的惩罚系数 C、选择模型所使用的核函数 kernel 和表示决策函数（样本到分离超平面的距离）的类型 decision_function_shape（有'ovo'，'ovr'两种）。

1．C 存在的意义在于维持 hard margin classification 和 soft margin classification 的平衡——使得间距大，但是间隔内的实例和分错的实例少甚至没有。惩罚系数小就会有更多的点在间隔（margin）内，或者分错，但是同时

可以是间距很大，这样的话泛化能力会更强。相反，惩罚系数大，间距很小但是不怎么会犯错，带来的副作用是可能会过拟合。

2. 我们使用的 SVC 本身运行速度较慢，可在训练集和测试集较小的情况下使用。

Class	Time complexity	Out-of-core support	Scaling required	Kernel trick
LinearSVC	$O(m \times n)$	No	Yes	No
SGDClassifier	$O(m \times n)$	Yes	Yes	No
SVC	$O(m^2 \times n)$ to $O(m^3 \times n)$	No	Yes	Yes

SVC 用到了一个技术 kernel trick，通过调节 kernel 参数改变核函数。核函数本身的目的在于避免计算训练集的 transformation ϕ，比如 2 度 poly 内核仅通过源向量 a 和 b 就可以计算ϕ(a)T ϕ(b)。（K(a, b) = (aT b)2）

$$\text{Linear:} \quad K(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}$$

$$\text{Polynomial:} \quad K(\mathbf{a}, \mathbf{b}) = \left(\gamma \mathbf{a}^T\mathbf{b} + r\right)^d$$

$$\text{Gaussian RBF:} \quad K(\mathbf{a}, \mathbf{b}) = \exp\left(-\gamma \| \mathbf{a} - \mathbf{b} \|^2\right)$$

$$\text{Sigmoid:} \quad K(\mathbf{a}, \mathbf{b}) = \tanh\left(\gamma \mathbf{a}^T\mathbf{b} + r\right)$$

针对其他的特定的训练集也会有其他不常用的 kernel。

3. 参数 decision_function_shape 有 ovo 和 ovr，ovo 需要训练更多的分类器，但是 ovo 中的每个切分出来的数据集都更小，因此分类器训练时间也将更短。时间开销 ovo 可能小于 ovr。

SVM 的 GridSearchCV 进行微调，也可以找到局部最优解：

[sklearn.model_selection.GridSearchCV — scikit-learn 1.1.2 documentation]

```
from sklearn.model_selection import GridSearchCV

parameters = {'kernel':('poly', 'rbf', 'sigmoid'), 'C':[0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]}
svc = svm.SVC()
clf = GridSearchCV(svc, parameters)
clf.fit(data_train, tag_train.ravel())
sorted(clf.cv_results_.keys())

print(clf.best_estimator_, clf.best_score_)
print(clf.cv_results_["mean_test_score"])
```

SVM 鸢尾花分类实验参数微调及实验结果：

Table 1: 调节 Parameter C，观测 linear，ovr 条件下分类准确度的变化

Parameters C	0.001	0.005	0.01	0.05	0.5	1	Kernel: linear, decision_function_shape: ovr
Training Prediction	0.676	0.695	0.886	0.962	0.981	0.981	
Test Accuracy	0.556	0.6	0.8	0.978	0.978	0.978	

Table 2: 更换 Parameter Kernel，观测 C=0.5，ovr 条件下分类准确度的变化

Parameters Kernel	linear	poly	rbf	sigmoid	precomputed	C: 0.5, decision_function_shape: ovr
Training Prediction	0.981	0.98	0.962	0.352	/ (matrix must be a square matrix)	
Test Accuracy	0.978	1	0.978	0.289		

Table 3: 其他变量一定，观测 ovo，ovr 条件下分类准确度的变化

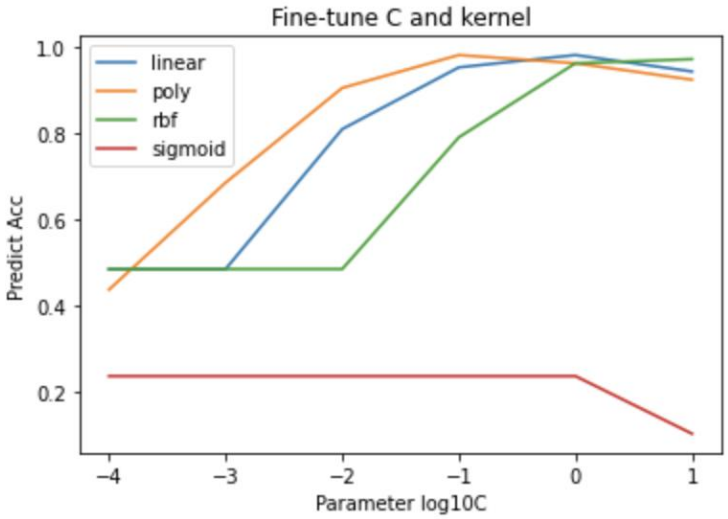
Parameters decision_function_shape	ovr	ovo	Kernel: linear, C: 0.5
Training Prediction	0.981	0.981	
Test Accuracy	0.978	0.978	

2.2 结果分析

SVM 鸢尾花分类实验最后的结果表明：

1. 在这个数据量和特征向量的案例中，表示决策函数类型的参数对实验结果的影响并不大。

2. 下图比较了不同的 kernel trick 和对应的不同的参数 C 下预测测试的准确度。C 过小，训练测试的准确度都会很差，但是对于 linear、poly 来说 C 过大也会带来结果的下降，模型的过拟合。Sigmoid 的表现一直不是很好。



3. 实验二

3.1 算法实现说明

在处理肿瘤数据集的过程中，需要考虑不同的分类名称和不同数量的特征向量。

除了与实验一相似的实验部分，我也进行了 SVC 管道化，运用 Pipeline 搭建模型，结果受 degree 参数影响较大。

```
Poly_svm_clf = Pipeline([
    ("poly_features", PolynomialFeatures(degree=3)),
    ("scaler", StandardScaler()),
    ("svm_clf", SVC(C=100, kernel='poly', decision_function_shape='ovr'))
])
Poly_svm_clf.fit(data_train, tag_train.ravel())


train_pred = Poly_svm_clf.predict(data_train)
test_pred = Poly_svm_clf.predict(data_test)

from sklearn.metrics import accuracy_score
print(accuracy_score(tag_train, train_pred))
print(accuracy_score(tag_test, test_pred))
```

SVM 肿瘤分类实验参数微调及实验结果：

Table 4 更换 Parameter Kernel，观测 C=0.5，ovr 条件下分类准确度的变化

Parameters Kernel	linear	poly	rbf	sigmoid	precomputed	C: 0.5, decision_function_shape : ovr
Training Prediction	0.392	0.392	0.480	0.424	/ (matrix must be a square matrix)	
Test Accuracy	0.352	0.278	0333	0.278		

- 经过实验发现 Linear 方法对时间的消耗很大，在之后的参数微调中，使用 rbf 作为 Kernel 进行后续实验。
- **在 3.2 最后的结果图中**，linear 仅使用了 4 个特征向量，因为经比较发现每多出一个特征向量，所需时间急剧增长，但是在分类效果上看相差并不是很大；因此后续的 Linear 测试结果仅基于前四种特征向量得出数据并节省试验时间。

Table 5 调节 Parameter C，观测 rbf，ovr 条件下分类准确度的变化

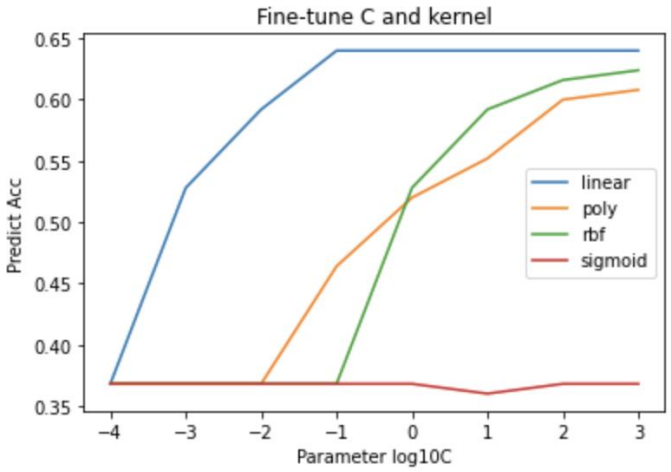
Parameters C	0.001	0.01	0.1	0.5	1	10	Kernel: linear, decision_function_shape: ovr
Training Prediction	0.368	0.368	0.368	0.480	0.567	0.616	
Test Accuracy	0.259	0.259	0.259	0333	0.574	0.556	

Table 6 其他变量一定，观测 ovo，ovr 条件下分类准确度的变化

Parameters decision_function_shape	ovr	ovo	Kernel: linear, C: 0.5
Training Prediction	0.480	0.480	
Test Accuracy	0333	0333	

3.2 结果分析

肿瘤 SVM 分类实验的结果与鸢尾花实验有一定的相似性：



对于 decision_function_shape 和 kernel，参数 C 的分析与实验一一致的部分不再赘述。其中的不同点在于，可以直观地看到 acc 普遍没有达到很高的结果。经过助教的解释，意识到可能所获取到的数据依然不足以支持 SVM 分类，可能还需要获取肿瘤图像等其他数据才能得到更好的结果。

4. 总结

在这门课程中的第一个实验中，我将课程中学习了解到的 SVM 算法深入实践。通过对各种参数的微调，和对比实验，我更好地理解了 SVM 中的 kernel trick 等概念，也对如何把实例分类、预测分类结果以及评估测试分类器效果等一系列实验过程有了更深刻的了解。