Ex3 record and report

王敏行 id：2018012386 wangmx18@mails.tsinghua.edu.cn

SVM 主要参数定义：

- Kernel：核函（参考了 10 SVM - 核函数 - 简书 (jianshu.com)）；
  - Linear：线性$K(x,z) =< x,z >$
  - Rbf：径向基，也称高斯核函数$K(x,z) = \exp(-\gamma||x-z||^2)$。
  - Poly：多项式$K(x,z) = (\gamma < x,z > +r)^d$
  - Sigmoid：$K(x,z) = \tanh(\gamma < x,z > +r)$
- C：正则项惩罚的强度反比于 C。
- Gamma：在 rbf、poly 和 sigmoid 的核函数中，x（输入）前会有一个系数 gamma
- Degree：仅对 poly 有效，多项式的阶数（表达式中的指数 d）

训练过程：

SVM 的训练，即在限制条件下进行|w|的最小值。对于线性不可分的数据集，需转化成用 KKT 条件下的朗格朗日乘子法求解。即先求出权重向量的值，再利用权重向量求出方向向量$w = \sum \alpha_i y_i x_i$，再利用支持向量求出 b。

序列最小优（SMO）算法就是一种利用迭代完成参数优化的算法。这个算法的每一步都是在权重向量α_i满足非负的情况下，尝试满足$\sum \alpha_i y_i = 0$。通过参考 EM 算法，SMO 每次优化两个权重向量并将其他的权重向量视作常数，进行局部优化。具体可以参考 11 SVM - SMO - 序列最小优化算法 - 简书 (jianshu.com)。

实验结果：

本实验中，利用 gridsearch 对 SVM 的核函数、C、gamma 进行优化。一共优化了 10 组参数（12 组参数中剔除等价的四组），结果显示除了‘linear’核函数之外，使用其它核函数的模型都有严重的过拟合现象，结果详见 notebook 的输出。最佳的参数组合为 `C=1, gamma=0.001, kernel='linear'`。在最佳参数下，用 trainset1 作为训练集，以 testset1 作为测试集，准确率分别为 `train acc:0.79840 validation acc:0.78031`。

显然 SVM 的性能在大量数据的表现并不佳（用了 70 min 才 grid search 了 12 个模型），根据 sklearn.SVM 的建议，选取 linearSVC（功能与 linear 内核的 SVC 相同，但是不同于 SVC 用 libsvm 编译，linearSVC 用 liblinear 编译，故在大样本上有更好的表现）。

对比运行时间，SVC 用了 17 min，linearSVC 用时 1.3 sec。二者采取完全相同的训练集和测试集，准确率也一样。

附：grid search 结果

{'C': 0.001, 'gamma': 0.001, 'kernel': 'rbf'} train accuracy:0.51625, test accuracy:0.51625

{'C': 0.001, 'gamma': 0.001, 'kernel': 'linear'} train accuracy:0.78263, test accuracy:0.76800

{'C': 0.001, 'gamma': 0.001, 'kernel': 'poly'} train accuracy:0.99850, test accuracy:0.69925

{'C': 0.001, 'gamma': 1, 'kernel': 'rbf'} train accuracy:0.51625, test accuracy:0.51625

{'C': 0.001, 'gamma': 1, 'kernel': 'linear'} train accuracy:0.78263, test accuracy:0.76800

{'C': 0.001, 'gamma': 1, 'kernel': 'poly'} train accuracy:0.99912, test accuracy:0.69300

{'C': 1, 'gamma': 0.001, 'kernel': 'rbf'} train accuracy:1.00000, test accuracy:0.52100

{'C': 1, 'gamma': 0.001, 'kernel': 'linear'} train accuracy:0.80331, test accuracy:0.77500

{'C': 1, 'gamma': 0.001, 'kernel': 'poly'} train accuracy:0.99962, test accuracy:0.68900

{'C': 1, 'gamma': 1, 'kernel': 'rbf'} train accuracy:1.00000, test accuracy:0.51625

{'C': 1, 'gamma': 1, 'kernel': 'linear'} train accuracy:0.80331, test accuracy:0.77500

{'C': 1, 'gamma': 1, 'kernel': 'poly'} train accuracy:0.99912, test accuracy:0.69300