实验1. 度量学习（Metric Learning）

2017 年 10 月 25 日

综述

在机器学习领域中，如何选择合适的距离度量准则一直都是一个重要而困难的问题。因为度量函数的选择非常依赖于学习任务本身，并且度量函数的好坏会直接影响到学习算法的性能。为了解决这一问题，我们可以尝试通过学习得到合适的度量函数。距离度量学习(Distance Metric Learning)的目标是学习得到合适的度量函数，使得在该度量下更容易找出样本之间潜在的联系，进而提高那些基于相似度的学习器的性能。本实验的目的是掌握距离度量学习的基本思路方法并应用到真实场景中去。本实验的要求语言为Python3。在实验开始前请认真阅读《机器学习》教程中相关章节，同时也可参考http://www.cs.cmu.edu/~liuy/dist_overview.pdf。

任务 1(50%)

自己在已有代码框架下实现一个 myDML.py 模块。请在该模块中实现方法：train(traindata) 和 distance(inst_a, inst_b)，要求直接 import myDML 后可调用 myDML.train(traindata) 完成训练过程，完成训练后调用 myDML.distance(inst_a, inst_b) 即可返回任意两样本间的距离。本次任务只需要改动 myDML.py 文件，请勿修改 test_myDML.py 文件。实现你的 myDML.py 模块后在Python文件目录下执行 python3 test_myDML.py 即可查看结果。

提交如下文件：

- Python文件：myDML.py。要求在Python文件目录下，执行 python3 test_myDML.py 即可进行基本功能测试（测试数据为moons dataset，数据分布可见 myDML/dataset/moons/moons_train(moons_test).png）。
- ReadMe文件：ReadMe.pdf。ReadMe 文件需要对 myDML.py 的实现进行说明。在该文件中至少需要说明度量函数的学习目标、使用的优化算法等情况（并注明是否参考开源实现，如果有，请注明自己的实现和对方的不同之处）。

任务 2 (50%)

你现在已经有了一个自己的距离度量学习库，可以用它做一些事情了。在本任务中，你将对数据集Letter Recognition Data Set进行预测。可登陆https://archive.ics.uci.

edu/ml/datasets/Letter+Recognition查看数据集。数据集中各个属性的具体说明见链接内Data Set Description。你无需下载数据集，因为本次任务中已经为你完成了30次数据划分。每次使用1820个样本进行训练，使用780个样本进行测试。这里使用 test error 作为本次任务的评价指标(值越小越好)，同时使用欧几里得距离+KNN作为本次任务的baseline。

实验设置说明：

- 本次任务首先将数据集随机选择70%的样本用做训练数据(注意保持各类样本比例不变)。训练完成后，在剩下的30%数据上进行测试。训练中可使用 cross-validation 进行调参。需要注意的是调用 train(dataset) 时有严格的时间限制(10分钟)，请尽量优化你的代码并且留有一定冗余时间。
- 度量函数训练完成后，测试中使用你的度量函数 + KNN (K = 1, 3, 5) 对测试数据进行预测。对预测结果计算 test error。实验需要重复30次，然后需要汇报 test error 的均值和标准差。测试时间可能较长，请耐心等待。

提交如下文件：

- 运行结果文件：evaluation.txt。该文件内容即在Python文件目录下运行 python letter_recognition.py 返回的输出内容。

  evaluation.txt中内容如下，请不要增加其他内容。

baseline+knn(k=1):	mean ± std
myMetric+knn(k=1):	mean ± std
baseline+knn(k=3):	mean ± std
myMetric+knn(k=3):	mean ± std
baseline+knn(k=5):	mean ± std
myMetric+knn(k=5):	mean ± std

加分鼓励任务 (10%)

一般而言，距离度量学习可以分为 global DML(例如Relevant Components Analysis)和 local DML(例如Local Fisher Discriminant Analysis)，本次任务中请实现另一种与你在以上任务中类型不同的度量学习方法，并且在任务2的数据集上进行测试评价并比较他们的区别。实现技术细节和比较情况也请在 ReadMe.pdf 中说明。