Principles of Data Science Project 1
Dimension Reduction

Hongzhou Liu	Xuanrui Hong	Qilin Chen
517030910214	517030910227	517030910155
deanlhz@sjtu.edu.cn	aaa@bbb.ccc	aaa@bbb.ccc

To Xuanrui Hong:
请先完成 Method 部分
截止日期：周六晚 24 点
提交方式：`git`

- 可以使用 ppt 上的图片、公式，使用公式必须要完整体现这个方法的内涵
- 可以参考维基百科，但需要做一定的改写, i.e. paraphrase
- 参考文献获取途径：前往 `sci-kit learn` 官方网站，搜索对应的方法，如 PCA，在相应文档页面查找是否有参考文献。或者找到左上角 User Guide，寻找相应内容，如 2.Unsupervised Learning-2.5.Decomposing signals in components (matrix factorization problems)
- 可以参考往届报告，行文风格、参考文献等，但不必每个方法处处都引用

To Qilin Chen:
请完成 Experiement 中 Feature Selection 之外的部分及 Conclusion 部分（周日晚 Xuanrui Hong 会完成 Feature Selection 部分的分析，并 `push` 至仓库中，请使用 `git pull` 获取）
截止日期：周二晚 24 点
提交方式：`git`

- 所有实验结果均在 Prj1/Result 下，包括不同方法、不同维度的降维结果用线性 SVM 以及核 SVM（RBF 核）的分类精度（文本、点线图），以及 2 维降维结果的可视化（训练集、测试集）
- Baseline 已经给出，请以 baseline 为基准比较各方法结果
- 请恰当使用 `figure`、`subfigure` 等方式美观地排版图片
- 请正确地设计表格，以表现不同降维方法、不同分类器（线性 SVM 以及核 SVM）产生的结果变化，并突出最佳参数设置
- AutoEncoder 部分请查看代码（Prj1/Projection/AutoEncoder.py）获取神经网络结构、超参数等数据，并在文中体现
- 关于参考文献：前往 `sci-kit learn` 官方网站，搜索对应的方法，如 PCA，在相应文档页面查找是否有参考文献。或者找到左上角 User Guide，寻找相应内容，如 2.Unsupervised Learning-2.5.Decomposing signals in components (matrix factorization problems)。除此之外还可以寻找其他参考文献。
- 可以参考往届报告，行文风格、参考文献等
- 请修改上方 `author` 部分你的邮箱
- 由于是多人同时完成不同的部分，每次修改前请使用 `git pull` 保持与远程仓库的同步，修改完毕后及时使用 `git push`

Abstract

TODO: Hongzhou Liu

1 Method

1.1 Feature Selection

1.1.1 Select-k-best

SelectKBest is one of the methods of univariate feature selection, which works by selecting the best features based on univariate statistical tests. It removes all but the k highest scoring features.

1.1.2 Variance Threshold

VarianceThreshold is a simple approach to feature selection. It removes all features whose variance doesn' t meet some threshold. The principle is that features with small variance often contain less data information.

1.1.3 Tree-based Selection

Tree-based feature selection combines SelectFromModel and ExtraTreesClassifier. SelectFromModel is a meta-transformer that can be used along with any estimator that has a $coef\_$ or $feature\_importances\_$ attribute after fitting. The features whose $coef\_$ or $feature\_importances\_$ values are below the provided threshold parameterare are considered unimportant and removed. ExtraTreesClassifier can be used to compute feature importances, which happens to cooperate with SelectFromModel to discard irrelevant features.

1.2 Feature Projection

1.2.1 PCA

Principal component analysis is one of the most widely used data dimensionality reduction algorithms. It performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.[1] Formally, the optimization goal is

$$\max_{v} \frac{1}{n} \sum_{i=1}^{n} (v^T x_i)^2 = \frac{1}{n} v^T X X^T v \tag{1}$$

where v is the new axis.

$$s.t. \quad v^T v = 1 \tag{2}$$

Using lagrange Multiplier we can get

$$X X^T v = \lambda v \tag{3}$$

We can see that v is the eigenvector of XX^T, and λ is the corresponding eigenvalue. Therefore, v can be calculated by performing eigenvalue decomposition to the co-variance matrix XX^T. Then we can get the data after dimensionality reduction.

1.2.2 Kernel PCA

In general, principal components analysis is suitable for linear dimensionality reduction of data. Kernel PCA can achieve nonlinear dimensionality reduction of data and is used to process linear inseparable data sets.

The general idea of KPCA is: for the matrix in the input space, we first use a non-linear mapping to map all samples in a high-dimensional or even infinite-dimensional space, and then perform PCA dimensionality reduction in this high-dimensional space.

1.2.3 LDA

TODO: Xuanrui Hong

1.3 Feature Learning

1.3.1 t-SNE

TODO: Xuanrui Hong

1.3.2 LLE

TODO: Xuanrui Hong

1.3.3 AutoEncoder

TODO: Xuanrui Hong

2 Experiment

2.1 Baseline

TODO: Hongzhou Liu
Training Set : Test Set = 6:4
Linear SVM: Best C = 0.002, best accuracy = 0.932815 (baseline)
Kernel SVM with RBF kernel: Best C = 5.0, best accuracy = 0.935160 (baseline)

2.2 Feature Selection

2.2.1 Select-k-best

TODO: Xuanrui Hong

2.2.2 Variance Threshold

TODO: Xuanrui Hong

2.2.3 Tree-based Selection

TODO: Xuanrui Hong

2.3 Feature Projection

2.3.1 PCA

TODO: Qilin Chen

2.3.2 LDA

TODO: Qilin Chen

2.4 Feature Learning

2.4.1 t-SNE

TODO: Qilin Chen

2.4.2 LLE

TODO: Qilin Chen

2.4.3 AutoEncoder

TODO: Qilin Chen

3 Conclusion

TODO: Qilin Chen

Acknowledgement

References

[1] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.