

Part 1. 手写数字分类

设计函数显示图片

我实现了一个函数 showDigits 可以输入 digits_data 的任意一行, 显示出该数字的图像。分别显示了黑白图像和灰度图像。

1. 实现一个朴素贝叶斯分类器

首先考虑将数据集二值化, 即对于灰度值大于 127 的点视为 1, 否则视为 0; 统计训练集中数字 1 的比例, 这是处理 Parameter estimation 即已知数字估计参数的后验概率的一个最简单的思考方式。

通过上述方式计算出 $P(X_i | C = 3)$ 和 $P(X_i | C = 8)$. 完成参数的训练和估计。

对于测试集, 需要计算 $P(C = d | X_1, X_2, \dots, X_{256}) = \frac{P(C) * \prod_{i=1}^{256} P(X_i | C)}{P(X_1, X_2, \dots, X_{256})}$, 由于分母是常数, 所以只需要考虑分子。为保证连乘不会使结果 0 溢出, 我使用了取对数的方法, 并且为了使求对数时不会出现 $\ln(0)$ 的情况, 我统一加上了一个小数字 0.0000001。

根据公式分别计算出测试集的后验概率 $P(C = 3 | X_1, X_2, \dots, X_{256})$ $P(C = 8 | X_1, X_2, \dots, X_{256})$ 。

因为 $P(C = 3 | X_1, X_2, \dots, X_{256}) + P(C = 8 | X_1, X_2, \dots, X_{256}) = 1$, 所以我定义 threshold 为 $P(C = 3 | X_1, X_2, \dots, X_{256}) < \text{threshold}$ 。程序中即 $(\text{prior_d3} + \log(1/\text{threshold} - 1)) < \text{prior_d8}$

2. 实现一个最小二乘分类器

有两种方法计算参数 w 的值:

- 转化为求二次优化问题的解

$$\begin{aligned} & \min_w (Xw - y)^2 + \lambda * w^T w \\ & = \min_w (Xw - y)^T (Xw - y) + \lambda * w^T w \\ & = \min_w \left(\frac{1}{2} w^T 2 * (X^T X + \lambda I) w - 2y^T Xw + y^T y \right) \end{aligned}$$

用 quadprog 函数求解。

- 求关于 w 的偏导数, 得到 w 的解

解为 $(X^T * X + \lambda I)^{-1} * X^T * y$

为使结果更加准确，对于线性回归我添加了一个参数 b，即 $W^T x + b$ ，参数 b 在矩阵 w 的第一行第一列。

3. 实现一个支持向量机分类器

首先一个很实用的计算距离的函数 pdist 和 pdist2。其中 squareform(pdist(f, 'euclidean')) 和 pdist2(f, f, 'euclidean') 效果是一样的。

那么线性基函数 $X^T X = (\text{traindata} * \text{traindata}')$

Gaussian RBF=

$\exp(-\text{squareform}(\text{pdist}(\text{traindata}, 'euclidean')) * \text{pdist}(\text{traindata}, 'euclidean')) / (\sigma^2)$

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}}$$

用解凸二次规划问题的方法解该问题中的参数 α_i 。

`alpha=quadprog(H, -ones(n,1), [], [], y_train', 0, 0*ones(n,1), C*ones(n,1));`

根据公式计算出参数 b：

$$b = y_i - \sum_{j=1}^n \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{for any } i \text{ that } \alpha_i \neq 0$$

再估计测试集：

$$y^* = \text{sign} \left(\sum_{i \in SV} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}') + b \right)$$

注 svm 函数参数 sigma 是由数据集的 d 相关的，即输入时是 [0.01d, 0.1d, d, 10d, 100d]，此时输入参数时必须先计算数据集的 d，这样才能得到正确的结果。

5-fold 验证

five-fold.m 是 5-fold 验证的统一的程序。

| 5-fold 编号 | 1 | 2 | 3 | 4 | 5 |
|-----------|----------|----------|----------|----------|----------|
| 数据集 | The rest | The rest | The rest | The rest | The rest |
| 测试集 | 1-200 | 201-400 | 401-600 | 601-800 | 801-1000 |

贝叶斯

横坐标是 threshold，纵坐标是 5-fold 编号，表格内为正确率

| | 0.5 | 0.6 | 0.7 | 0.75 | 0.8 | 0.85 | 0.9 |
|---|--------|--------|--------|--------|--------|--------|--------|
| 1 | 0.9400 | 0.9400 | 0.9400 | 0.9400 | 0.9400 | 0.9450 | 0.9450 |
| 2 | 0.9650 | 0.9650 | 0.9600 | 0.9600 | 0.9600 | 0.9600 | 0.9600 |
| 3 | 0.9400 | 0.9400 | 0.9400 | 0.9450 | 0.9450 | 0.9550 | 0.9500 |
| 4 | 0.9650 | 0.9650 | 0.9650 | 0.9650 | 0.9600 | 0.9550 | 0.9550 |

| | | | | | | | |
|---------|--------|--------|--------|--------|--------|--------|--------|
| 5 | 0.9400 | 0.9450 | 0.9350 | 0.9400 | 0.9450 | 0.9450 | 0.9400 |
| Average | 0.95 | 0.951 | 0.948 | 0.95 | 0.95 | 0.952 | 0.95 |

由于是二分类的 bayes 分类器，而且直观上理解数字 3 和数字 8 是均匀的，按照这个理解归一化后可以认为概率高的即视为该数字。

在实测过程中，根据以上结果，可以看到 threshold=0.5 是，相对而言识别正确率都是较高的。比较符合直观的印象。

根据多分类 bayes classifier 公式：

$$\hat{y} = \operatorname{argmax}_{k \in \{1, \dots, K\}} p(C_k) \prod_{i=1}^n p(x_i | C_k).$$

即认为识别为概率最大的 C，map 到二分类的情况下也是认为 threshold=0.5。总而言之无论从数据实测角度还是从公式分析角度，在这个实例下 threshold=0.5 是最佳的。

线性回归

横坐标是 lambda，纵坐标是 5-fold 编号，表格内为正确率

| | 1e-4 | 0.01 | 0.1 | 0.5 | 1 | 5 | 10 | 100 | 1000 | 5000 | 10000 |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.975 | 0.975 | 0.975 | 0.975 | 0.985 |
| 2 | 0.965 | 0.965 | 0.965 | 0.965 | 0.965 | 0.975 | 0.975 | 0.98 | 0.98 | 0.98 | 0.98 |
| 3 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.965 | 0.965 | 0.965 | 0.965 |
| 4 | 0.96 | 0.96 | 0.96 | 0.965 | 0.965 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.975 |
| 5 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.97 | 0.97 | 0.975 | 0.985 |
| Average | 0.972 | 0.972 | 0.972 | 0.973 | 0.973 | 0.976 | 0.975 | 0.972 | 0.972 | 0.973 | 0.978 |

规范化项的提出就是为了避免过拟合，lambda 越大则表示惩罚力度越强，惩罚因子的确定和 w 的表现有一定的关系。在本实验中，根据实测的数据，lambda 改变对于正确率的影响影响其实不算很大，波动为 1-5 个判错，综合表现而言 lambda=10000 时效果最好，得到了很高的正确率。

SVM

横坐标是 sigma，纵坐标是 5-fold 编号，表格内为正确率

C=1

| C=1 | 0.01d | 0.1d | d | 10d | 100d |
|---------|-------|-------|-------|-------|-------|
| 1 | 0.455 | 0.455 | 0.455 | 0.455 | 0.455 |
| 2 | 0.505 | 0.505 | 0.505 | 0.505 | 0.505 |
| 3 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 |
| 4 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| 5 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 |
| Average | 0.482 | 0.482 | 0.482 | 0.482 | 0.482 |

C=10

| C=10 | 0.01d | 0.1d | d | 10d | 100d |
|------|-------|-------|-------|-------|-------|
| 1 | 0.94 | 0.455 | 0.455 | 0.455 | 0.455 |
| 2 | 0.965 | 0.505 | 0.505 | 0.505 | 0.505 |

| | | | | | |
|---------|-------|-------|-------|-------|-------|
| 3 | 0.955 | 0.44 | 0.44 | 0.44 | 0.44 |
| 4 | 0.975 | 0.5 | 0.5 | 0.5 | 0.5 |
| 5 | 0.955 | 0.51 | 0.51 | 0.51 | 0.51 |
| Average | 0.958 | 0.482 | 0.482 | 0.482 | 0.482 |

C=100

| C=100 | 0.01d | 0.1d | d | 10d | 100d |
|---------|-------|-------|-------|-------|-------|
| 1 | 0.98 | 0.455 | 0.455 | 0.455 | 0.455 |
| 2 | 0.985 | 0.505 | 0.505 | 0.505 | 0.505 |
| 3 | 0.98 | 0.44 | 0.44 | 0.44 | 0.44 |
| 4 | 0.82 | 0.5 | 0.5 | 0.5 | 0.5 |
| 5 | 0.805 | 0.51 | 0.51 | 0.51 | 0.51 |
| Average | 0.914 | 0.482 | 0.482 | 0.482 | 0.482 |

C=1000

| C=1000 | 0.01d | 0.1d | d | 10d | 100d |
|---------|-------|-------|-------|-------|-------|
| 1 | 0.965 | 0.94 | 0.455 | 0.455 | 0.455 |
| 2 | 0.99 | 0.965 | 0.505 | 0.505 | 0.505 |
| 3 | 0.97 | 0.955 | 0.44 | 0.44 | 0.44 |
| 4 | 0.975 | 0.975 | 0.5 | 0.5 | 0.5 |
| 5 | 0.985 | 0.955 | 0.51 | 0.51 | 0.51 |
| Average | 0.977 | 0.958 | 0.482 | 0.482 | 0.482 |

理论上分析，C 是分类成本，如 Dima 所说。C 越大，你得到的偏差越低，方差越高。低偏差是因为你惩罚了误分类的成本。C 越小，你得到的偏差越高，方差越低。高斯 RBF 核函数中，Sigma 越大，分离面越平滑；Sigma 越小，分离面越细致。这是因为 sigma 越小，核函数对 x 的衰减越快，这就放大了数据 x 之间的差别，即 $k(x)$ 对 x 值的变化很敏感，因此 SVM 的分离面变得细致；同样的道理，sigma 越大，核函数对 x 的衰减越慢，这使 $k(x)$ 对 x 的变化变得钝化（即不敏感），进而使 SVM 的分离面变得平滑。

从本实验实践上看，C 偏大并且 sigma 偏小表现地正确率更好，说明本实验比较偏 hard margin，并且 SVM 的 Gaussian 分割面也比较 hard，根据之前的 linear classifier 表现非常好来看，这样的参数结果也是比较符合预期的。

Part 2. 图片去噪

设计函数显示图片

我实现了一个函数 showPics 可以输入图片数据的任意一行，显示出该人脸图像。

可选的第二个参数是一个 bmp(bitmap)类型的字符串，用以存储该图片。

参考使用：
`showPics(ground_truth(1,:));`
`showPics(ground_truth(1,:), 'groud_1.bmp');`

实现算法

基本是按照 PPT 中的线性 PCA 算法设计。

$$S = \frac{1}{n}XX^T$$
$$X' = PP^TX$$

讨论

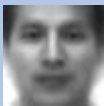
重构误差分析

取 threshold 0.9 0.95 0.999 0.999999

| Threshold | 0.9 | 0.95 | 0.999 | 0.999999 |
|-----------|--------|--------|--------|----------|
| 1 | 0.0253 | 0.0088 | 0.0079 | 0.0078 |
| 2 | 0.0145 | 0.0106 | 0.0060 | 0.0062 |
| 3 | 0.0186 | 0.0100 | 0.0056 | 0.0055 |
| 4 | 0.0213 | 0.0096 | 0.0045 | 0.0044 |
| 5 | 0.0189 | 0.0040 | 0.0026 | 0.0025 |
| 6 | 0.0089 | 0.0078 | 0.0034 | 0.0033 |

从重构误差来看，随着 threshold 的增加，重构误差都是降低的[1, 2, 3, 4, 5, 6]，而且重构误差在 0.9 到 0.95 间降低了很多，有 50% 的降低，而且 threshold=0.999 和 threshold=0.999999 时重构误差基本差不多。

重构视觉上的分析

| Threshold | 0.9 | 0.95 | 0.999 | 0.999999 | Ground truth |
|-----------|---|---|---|--|---|
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |



从表中可以看出, $\text{threshold}=0.9$ 时基本只能识别出人脸的基本特征, 所有结果看上去都像人脸但是和 ground truth 差距很远。

随着 threshold 增加降噪效果从视觉上看是越来越好, 在 $\text{threshold}=0.999$ 时, 视觉上看和 ground truth 已经很像了。

为什么 PCA 可以降噪

举例而言, 假设三维空间中有一系列点, 这些点分布在一个过原点的斜面上, 如果用自然坐标系 x,y,z 这三个轴来表示这组数据的话, 需要使用三个维度, 而事实上, 这些点的分布仅仅是在一个二维的平面上, 那么自然可以把 x,y,z 坐标系旋转一下, 使数据所在平面与 x,y 平面重合。如果把旋转后的坐标系记为 x',y',z' , 那么这组数据的表示只用 x' 和 y' 两个维度表示即可。这就是数据降维。

上面认为把数据降维后并没有丢弃任何东西, 因为这些数据在平面以外的第三个维度的分量都为 0。现在, 假设这些数据在 z' 轴有一个很小的抖动, 那么我们仍然用上述的二维表示这些数据, 理由是我们认为这两个轴的信息是数据的主成分, 而这些信息对于我们的分析已经足够了, z' 轴上的抖动很有可能是噪声, 也就是说本来这组数据是有相关性的, 噪声的引入, 导致了数据不完全相关, 但是, 这些数据在 z' 轴上的分布与原点构成的夹角非常小, 也就是说在 z' 轴上有很大的相关性, 综合这些考虑, 就可以认为数据在 x',y' 轴上的投影构成了数据的主成分。

所以 PCA 的思想是将 n 维特征映射到 k 维上 ($k < n$), 这 k 维是全新的正交特征。这 k 维特征称为主成分, 是重新构造出来的 k 维特征, 而不是简单地从 n 维特征中去除其余 $n-k$ 维特征。这样就剔除了和标签无关的特征, 再按照转换矩阵生成原图, 就达到了降噪的效果。

参考文档

https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Probabilistic_model bayes 分类的基本思想

http://www.lx.it.pt/~mtf/learning/Bayes_lecture_notes.pdf 手写数字识别的 method of Parameter estimation

https://en.wikipedia.org/wiki/Tikhonov_regularization L2 Regularization 的特性简介

<https://www.quora.com/What-are-C-and-gamma-with-regards-to-a-support-vector-machine> SVM 的参数理解

<http://zzy07053437.blog.163.com/blog/static/2075520872012102725946123/> 高斯 RBF 核函数中 Sigma 的取值和 SVM 分离面的关系

<http://blog.csdn.net/lujiandong1/article/details/46386201> SVM 的两个参数 C 和 gamma

<http://blog.csdn.net/zhongkelee/article/details/44064401> 主成分分析 (PCA) 原理详解

<http://yajunok.blog.163.com/blog/static/65657620089179480257/> matlab 矩阵运算

https://mqshen.gitbooks.io/prml/content/Chapter3/basis/geometry_least_square.html 最小二乘法几何解释

<http://blog.sciencenet.cn/blog-531885-589056.html> pdist pdist2 距离生成