

模式识别

特征归一化
Fisher线性判别分析
应用：人脸识别

吴建鑫

南京大学计算机系 & 人工智能学院，2020

目标

- ✓ 掌握并能应用常见的特征归一化方法
- ✓ 能应用FLD，并能掌握其推导过程
- ✓ 能将PCA和FLD应用到人脸识别当中去
- ✓ 提高目标
 - 进一步能将本章方法应用到实际研究问题中去（研究生、部分本科生）
 - 对线性判别在不同条件下的变化，有兴趣的可以进一步阅读

特征归一化

Feature normalization

1. 每维度归一

✓ per-dimension normalization

- 虚拟的例子（判别性别）

- 假设用两个特征：身高和体重

- 如果1. 身高单位毫米，体重单位吨，那么？

- 如果2. 身高单位公里，体重单位克，那么？

- 很多时候，不同的维度需要统一到同样的取值范围！

✓ 训练集： $\mathbf{x}_1, \dots, \mathbf{x}_n$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$

- 对每一维 j ，其数据为 $x_{1j}, x_{2j}, \dots, x_{nj}$

- 取其最小值 $x_{min,j}$ 和最大值 $x_{max,j}$

- 对这一维的任何数据 $x_{ij} \leftarrow \frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}}$

稀疏数据

✓ 新数据的范围是？各维度统一了吗？

- $[0 \ 1]$
- 若某一维 $x_{max,j} = x_{min,j}$?
- 也可以统一到 $[-1 \ 1]$

$$x_{ij} \leftarrow 2 \times \left(\frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}} - 0.5 \right)$$

✓ 稀疏数据 sparse data: 数据中很多维度值为0

- 如果所有数据 ≥ 0 ，在两种归一化中，原来是0的会变成什么？

2. ℓ_2 或 ℓ_1 归一化

✓ 若各维度取值范围的不同是有意义的，但是不同数据点之间的**大小**（如向量长度norm）应保持一致

- 对每个数据 $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$

$$x_{ij} \leftarrow \frac{x_{ij}}{\|\mathbf{x}_i\|_{\ell_2}} \quad \|\mathbf{x}_i\|_{\ell_2} = \sqrt{\mathbf{x}_i^T \mathbf{x}_i}$$

✓ ℓ_1 归一化

- 适用于非负的特征，即 $x_{ij} \geq 0$ 总成立
- 若数据 \mathbf{x}_i 是**直方图**(histogram)时，经常是最佳的

$$x_{ij} \leftarrow \frac{x_{ij}}{\|\mathbf{x}_i\|_{\ell_1}} \quad \|\mathbf{x}_i\|_{\ell_1} = \sum_{j=1}^d |x_{ij}|$$

3. zero-norm, unit variance

- ✓ 有时候有理由相信每一个维度是服从高斯分布的
 - 希望每一个维度归一化到 $N(0,1)$
- ✓ 对每一维 j , 其数据为 $x_{1j}, x_{2j}, \dots, x_{nj}$
 - 计算其均值 $\hat{\mu}_j$ 和方差 $\hat{\sigma}_j^2$
 - 对每一个特征值

$$x_{ij} \leftarrow \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

归一化测试数据

- ✓ 怎样归一化测试数据？
 - 从测试集寻找最大值、最小值、均值？
- ✓ 除了在测试的时候，永远不要使用测试数据！
 - 测试集和训练集应该使用相同的归一化方法
 - 还记得吗？训练和测试集应该从相同的 $p(\mathbf{x})$ 取样
 - 同样的归一化会保持这个限定！
 - 这个原则同样适用于交叉验证！
- ✓ 那么，怎样做？
 - 保存从训练集上取得的归一化参数(parameter)
 - 使用同样的公式和保存的参数来归一化测试集

归一化小结

- ✓ 归一化的方法应该是根据数据的特点来选择的
 - 在做任何机器学习之前，先搞清你的数据的特点
 - 稀疏？
 - 每一维有没有含义？
 - 每一维里面值的分布情况？Gauss？
 - 看你的数据！Do visualization！
- ✓ 归一化可能对准确度有极大的影响！
 - 在有些例子里，正确的归一化能大幅度提高accuracy
- ✓ 不同的归一化方法可以混合使用

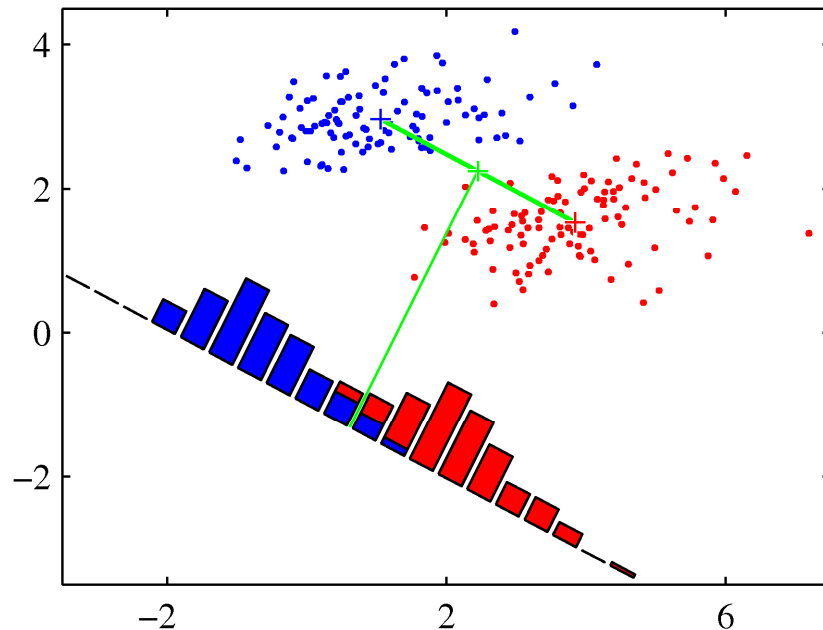
Fisher线性判别分析

Fisher's Linear Discriminant analysis (FLD, 有时候LDA)

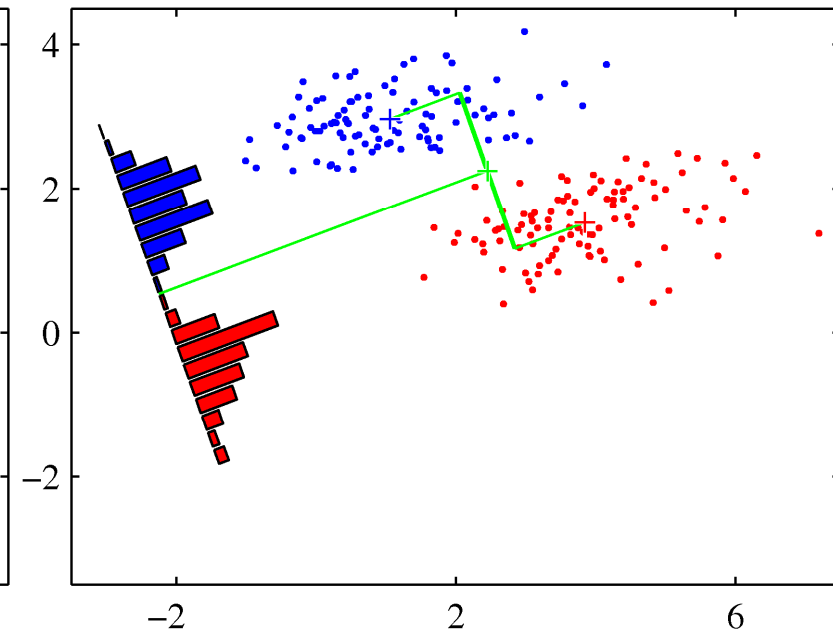
为什么需要FLD ?

- ✓ 理论上可以证明, PCA在数据是单个高斯分布是最佳
 - PCA有利于表示数据, 但和分类无关
- ✓ 分类问题中, 不同类别的分布 $p(\mathbf{x}|\mathbf{y} = i)$ 不能相同
- ✓ 如何提取特征(extract feature), 最有利于分类?
 - FLD是某些限制条件下最佳的线性特征提取方法
optimal linear feature extraction method under certain assumptions

Idea: FLD的动机 (motivation)



Bad linear feature (projection)



Good linear feature (projection)

Image courtesy of Christopher M. Bishop, author of PRML
<http://research.microsoft.com/en-us/um/people/cmbishop/prml/webfigs.htm>

用数学形式表示formalize

- ✓ 两个类别 $y_i \in \{1,2\}$, 数据 \mathbf{x}_i , 两类各有 N_1, N_2 个点
- ✓ 希望寻找一个投影方向 projection direction,
 $\mathbf{u} = \mathbf{w}^T \mathbf{x}$, 使得两个类别的数据在投影以后容易被
分开 separate

- ✓ 两个类各自的均值为

- $\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{y_i=1} \mathbf{x}_i,$

- $\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{y_i=2} \mathbf{x}_i$

- 投影以后的均值为 $m_1 = \mathbf{w}^T \boldsymbol{\mu}_1,$ $m_2 = \mathbf{w}^T \boldsymbol{\mu}_2$

Objective: Fisher's Criterion

- ✓ 怎样描述“分开”的程度(separation)?
- ✓ Maximize $(m_2 - m_1)^2$? 问题?
 - 这个值可以无限大。怎么解决?
 - 加限制条件 $\mathbf{w}^T \mathbf{w} = 1$
 - 看前面的图，这个值不是越大越好。怎么解决?
- ✓ Fisher准则
 - 在要求 $|m_2 - m_1|$ 尽量大的同时，要求两类在投影以后尽量集中，或者不分散。怎么度量分散程度?

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

分散程度的度量

✓ 对一维数据，自然的度量是方差或散度 ($k = 1, 2$)

$$s_k^2 = \sum_{y_i=k} (u_i - m_k)^2$$

- 称为类内散度 within class scatter

✓ $s_1^2 + s_2^2$: 总的类内散度

- total within-class scatter

✓ $s_k^2 = \sum_{y_i=k} (u_i - m_k)^2 = \sum_{y_i=k} (\mathbf{w}^T (\mathbf{x}_i - \boldsymbol{\mu}_k))^2 =$
 $\mathbf{w}^T \sum_{y_i=k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{w}$

✓ $(m_2 - m_1)^2 = \mathbf{w}^T (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \mathbf{w}$

散布矩阵

✓ $S_k = \sum_{y_i=k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$ 是什么？

✓ 类内散布矩阵 within-class scatter matrix

$$S_W = S_1 + S_2$$

✓ 类间散布矩阵 between-class scatter matrix

$$S_B = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T$$

✓ Fisher准则的矩阵形式(为什么要有矩阵形式？)

- $\max J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}, \text{ s.t. } \mathbf{w}^T \mathbf{w} = 1$

- 这种形式称为广义瑞利商 generalized Rayleigh quotient

Optimization: 如何求解？

- ✓ (Simplification/transformation) 练习：用拉格朗日乘子法，证明（记得查表）最优时必须满足

$$S_B \mathbf{w} = \lambda S_W \mathbf{w}$$

- ✓ 该问题称为广义特征值generalized eigenvalue问题
 - 得到“ S_B 和 S_W ”的广义特征值和广义特征向量
 - Generalized eigenvalue (eigenvector) of S_B and S_W
- ✓ 但是我们不用去解这个问题
 - $S_B \mathbf{w} = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T \mathbf{w} \propto (\mu_2 - \mu_1)$
 - $(\mu_2 - \mu_1) = \lambda S_W \mathbf{w}!$

FLD的步骤

1. 计算 μ_2, μ_1
2. 计算 S_W
3. 计算 $\mathbf{w} = S_W^{-1}(\mu_2 - \mu_1)$
4. 归一化:

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

如果不可逆怎么办？

- ✓ 如果数据很少或者维度很高， S_W 很可能不可逆
 - 广义逆矩阵generalized inverse matrix
- ✓ S_W 是实对称的，而且至少是半正定的
 - $S_W = E\Lambda E^T$, $\lambda_{ii} \geq 0$
- ✓ Moore - Penrose伪逆pseudoinverse
 - 若 $\lambda_{ii} > 0$ ，定义 $\lambda_{ii}^+ = 1/\lambda_{ii}$ ，否则定义 $\lambda_{ii}^+ = 0$
 - Λ 的M-P伪逆为： $\Lambda^+ = \text{diag}(\lambda_{11}^+, \lambda_{22}^+, \dots, \lambda_{dd}^+)$
 - S_W 的伪逆为

$$S_W^+ = E\Lambda^+ E^T$$

如果大于2类怎么办？

✓ C类问题

- μ_i, N_i, m_i, S_i 和2类问题中一样定义
- $S_W = \sum_{i=1}^C S_i$, 很容易从2类问题推广
- 定义 $N = \sum_{i=1}^C N_i$
- 定义总均值 $\mu = \frac{1}{N} \sum_{i=1}^C N_i \mu_i = \frac{1}{N} \sum_x x$

✓ S_B 没有定义，无法直接从2类问题推广

- 总散布矩阵 total scatter matrix, $S_T = \sum_x (x - \mu)(x - \mu)^T$
- $S_T = S_W + \sum_{i=1}^C N_i (\mu_i - \mu)(\mu_i - \mu)^T = S_W + S_B$
- 定义多类的 $S_B = \sum_{i=1}^C N_i (\mu_i - \mu)(\mu_i - \mu)^T$
- 练习：证明，当 $C = 2$ 时，有 $S_T = S_W + S_B$

更多的投影方向

$$\max J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

- ✓ 求解广义特征值问题

$$S_B \mathbf{w}_i = \lambda_i S_W \mathbf{w}_i$$

- ✓ 最多能得到 $C - 1$ 个有效的投影方向
 - 为什么?

- ✓ 利用Matlab来获得解

应用：人脸识别

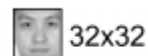
Application: face recognition

人脸

- ✓ 为什么人脸数据特别适合PCA和FLD?
- ✓ 用什么分类器?
- ✓ ORL人脸数据集:
<http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>
- ✓ OpenCV人脸识别tutorial
http://docs.opencv.org/modules/contrib/doc/facerec/facerec_tutorial.html
- ✓ 准备作业：首先需要在windows/linux/mac下安装OpenCV

张量Tensor: 深度学习的基石

- ✓ 标量(scalar, 纯量): $x \in \mathbb{R}$
- ✓ 向量(vector): $\mathbf{x} \in \mathbb{R}^d$
- ✓ 矩阵(matrix): $X \in \mathbb{R}^d \times \mathbb{R}^d$
- ✓ 进一步推广?
 - 如果 $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$
 - 称为张量tensor, 上例是3阶张量
 - 标量、向量、矩阵分别是0、1、2阶张量
- ✓ 张量的操作, 最基本的是向量化vectorize
 - 将矩阵的各行堆积stack起来



32x32

Reshaping/Vectorization

1024x1

人脸识别有哪些可以做的简单实验？

✓ ??

进一步的阅读

✓ 不同条件或要求下的线性特征抽取

- 如<http://cs.nju.edu.cn/wujx/paper/icml2005.pdf>

✓ 张量和多线性特征抽取

- Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data
<http://www.crcpress.com/product/isbn/9781439857243>
- 人脸图像向量化的图来自该书

✓ 关于特征值和特征向量

- Golub & van Loan, Matrix Computation, 3rd ed.
<http://www.cs.cornell.edu/courses/cs621/Books/GVL/index.htm>