

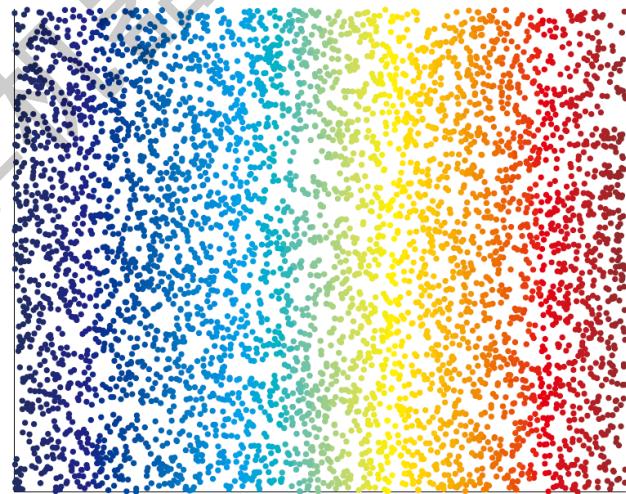
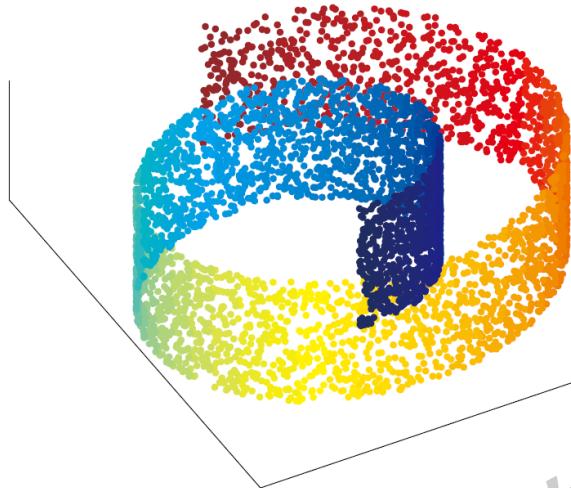
机器学习导论 (2021 春季学期)

习题课

主讲教师：周志华

主成分分析 (Principal Component Analysis, PCA)

正交属性空间中的样本点，如何使用一个超平面对所有样本进行恰当的表达？

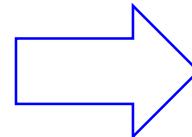


主成分分析 (Principal Component Analysis, PCA)

正交属性空间中的样本点，如何使用一个超平面对所有样本进行恰当的表达？

若存在这样的超平面，那么它大概应具有这样的性质：

- 最大可分性：样本点在这个超平面上的投影能尽可能分开
- 最近重构性：样本点到这个超平面的距离都足够近



主成分分析的两种等价推导

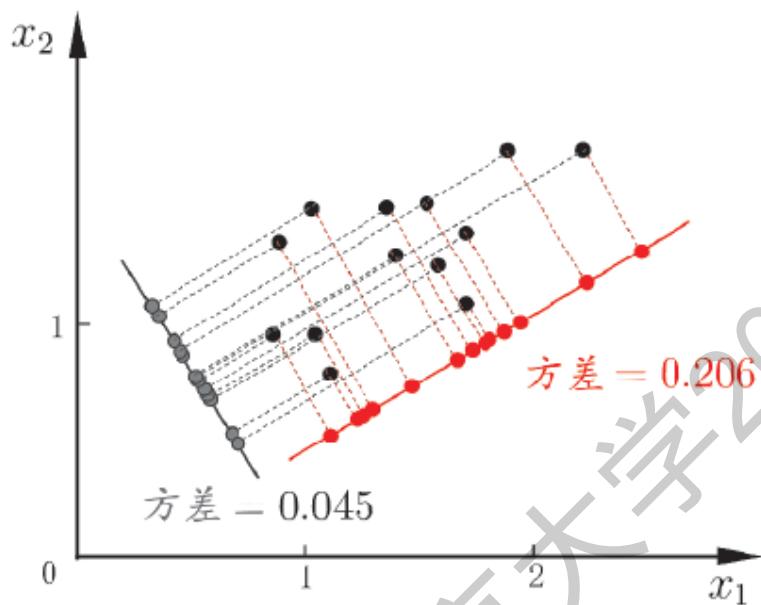
对样本进行中心化： $\sum_i x_i = \mathbf{0}$

PCA - 最大可分性

样本点 \mathbf{x}_i 在新空间中超平面上的投影是 $\mathbf{W}^T \mathbf{x}_i$ ，若所有样本点的投影能尽可能分开，则应该使得投影后样本点的方差最大化

投影后样本点的方差是 $\sum_i \mathbf{W}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{W}$

于是： $\max_{\mathbf{W}} \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W})$
s.t. $\mathbf{W}^T \mathbf{W} = \mathbf{I}$.



等价于：

$$\min_{\mathbf{W}} - \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W})$$

s.t. $\mathbf{W}^T \mathbf{W} = \mathbf{I}$.

PCA 求解

$$\begin{aligned} \max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

使用拉格朗日乘子法可得

$$\mathbf{X} \mathbf{X}^T \mathbf{W} = \lambda \mathbf{W}.$$

只需对协方差矩阵 $\mathbf{X} \mathbf{X}^T$ 进行特征值分解，并将求得的特征值排序： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ ，再取前 d' 个特征值对应的特征向量构成 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'})$ ，这就是主成分分析的解

关键变量：子空间方差

PCA - 最近重构性

对样本进行中心化: $\sum_i \mathbf{x}_i = \mathbf{0}$

假定投影变换后得到的新坐标系为 $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$, 其中 \mathbf{w}_i 是标准正交基向量

$$\|\mathbf{w}_i\|_2 = 1, \mathbf{w}_i^T \mathbf{w}_j = 0 (i \neq j).$$

若丢弃新坐标系中的部分坐标, 即将维度降低到 $d' < d$, 则样本点在低维坐标系中的投影是 $\mathbf{z}_i = (z_{i1}; z_{i2}; \dots; z_{id'})$ $z_{ij} = \mathbf{w}_j^T \mathbf{x}_i$

若基于 \mathbf{z}_i 来重构 \mathbf{x}_i , 则会得到 $\hat{\mathbf{x}}_i = \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j$.

PCA - 最近重构性 (续)

原样本点 \mathbf{x}_i 与基于投影重构的样本点 $\hat{\mathbf{x}}_i$ 之间的距离为

$$\begin{aligned} \sum_{i=1}^m \left\| \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j - \mathbf{x}_i \right\|_2^2 &= \sum_{i=1}^m \mathbf{z}_i^T \mathbf{z}_i - 2 \sum_{i=1}^m \mathbf{z}_i^T \mathbf{W}^T \mathbf{x}_i + \text{const} \\ &\propto -\text{tr} \left(\mathbf{W}^T \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{W} \right). \end{aligned}$$

\mathbf{w}_j 是正交基, $\sum_i \mathbf{x}_i \mathbf{x}_i^T$ 是协方差矩阵, 于是由最近重构性, 有:

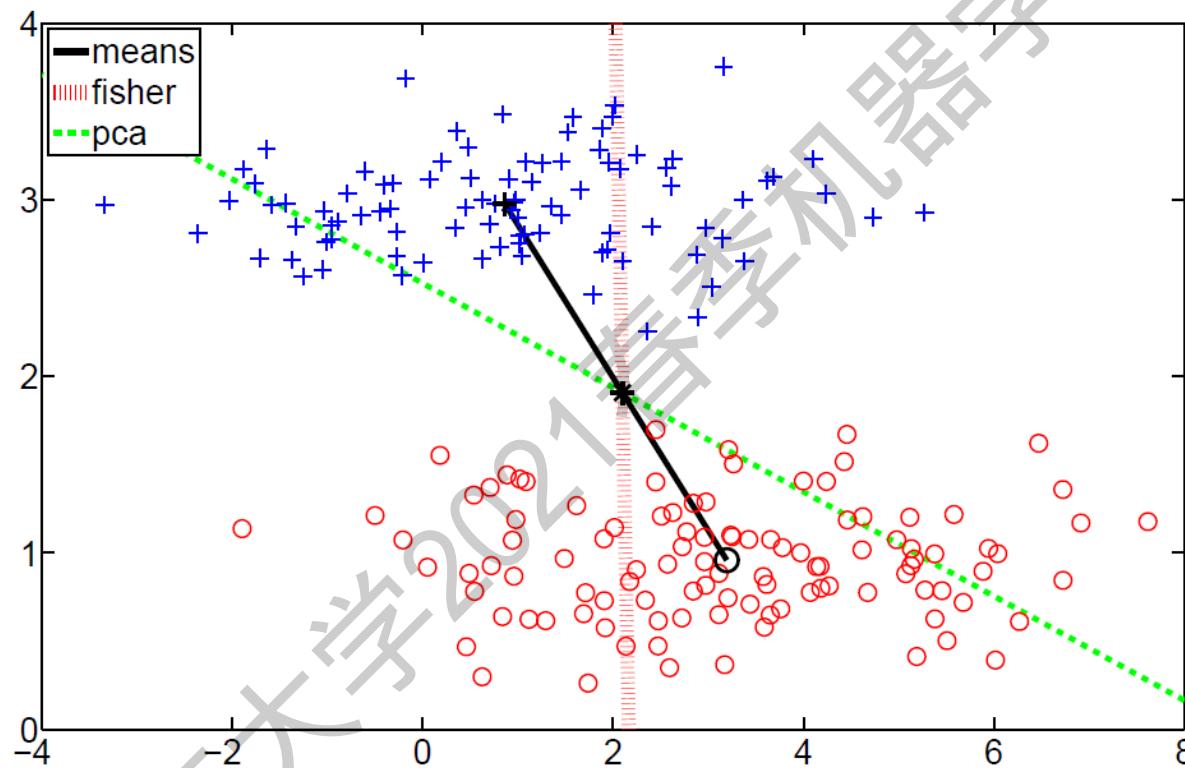
$$\begin{aligned} \min_{\mathbf{W}} \quad & -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

关键变量：重构误差

这就是主成分分析的优化目标

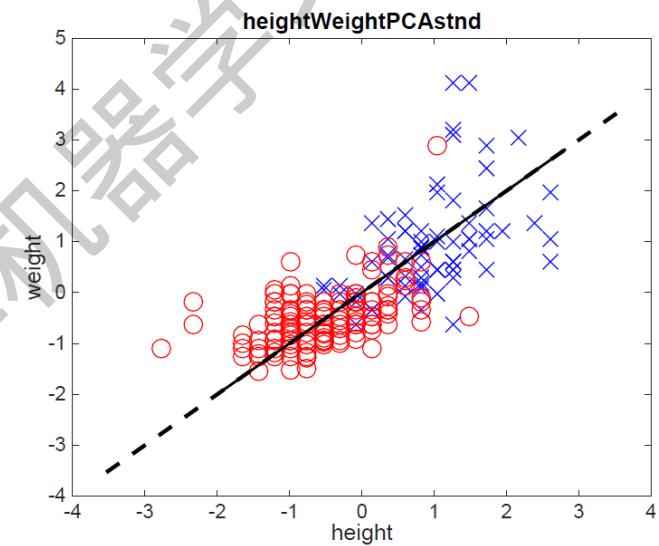
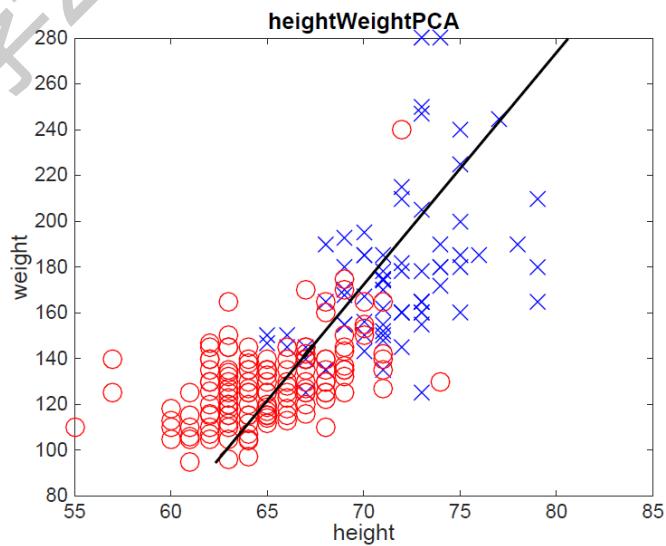
PCA - FDA

PCA是无监督学习方法，而FDA是监督学习方法，考虑了标记的作用。



PCA 应用

协方差矩阵易受到特征尺度影响



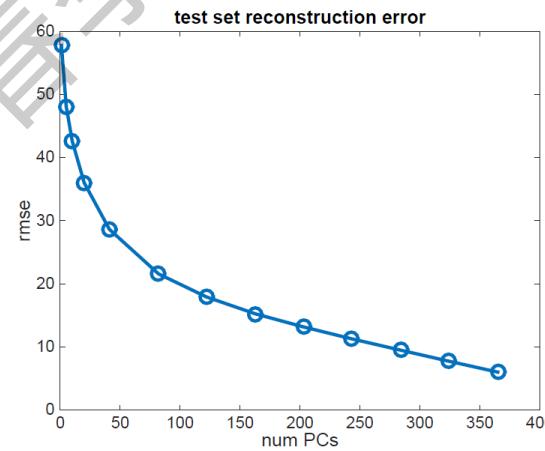
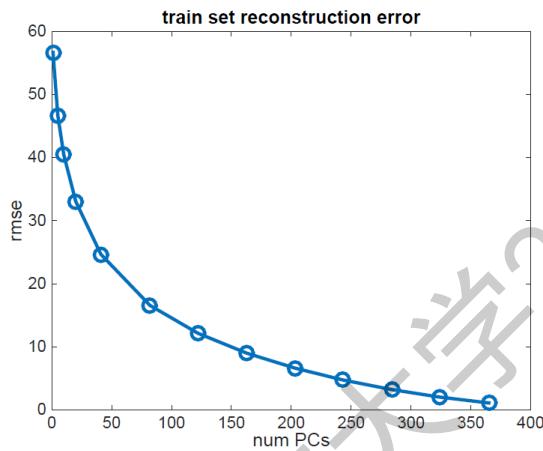
通过对数据进行标准化，使所有特征在同一尺度上

PCA 应用

d' 的设置：

- 用户指定
- 通过重构误差判断？

$$\sum_{i=1}^m \|x_i - \hat{x}_i\|_2^2$$



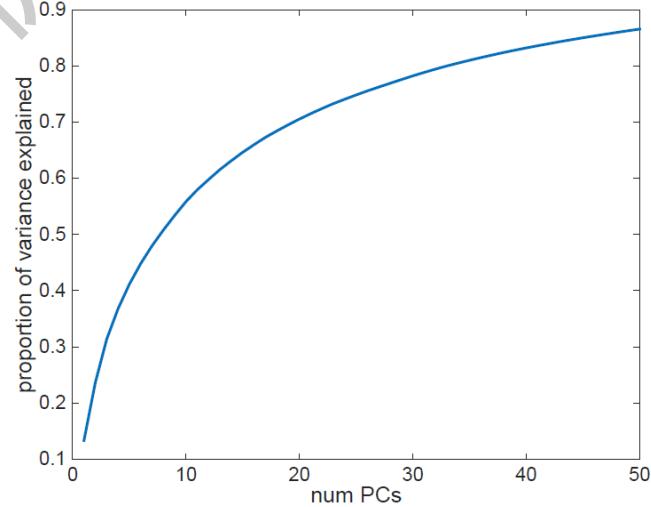
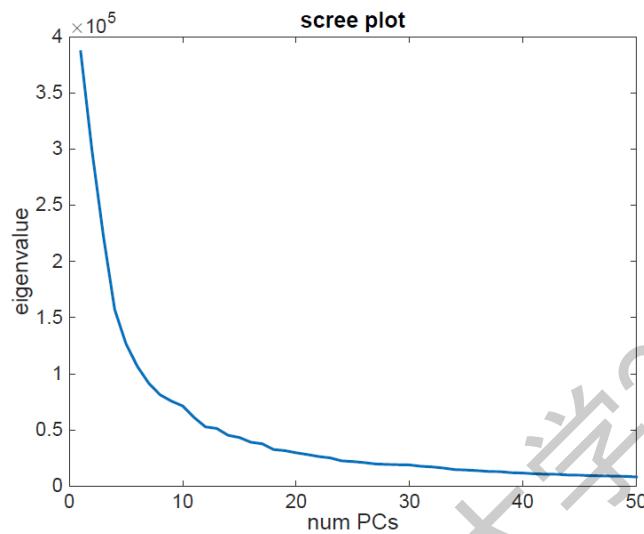
模型越复杂，重构误差越低

PCA 应用

d' 的设置：

- 用户指定
- 在低维空间中对 k 近邻或其他分类器进行交叉验证
- 设置重构阈值，例如 $t = 95\%$ ，然后选取最小的 d' 使得

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^d \lambda_i} \geq t.$$

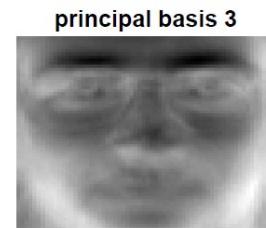


PCA 应用 (续)

PCA 是最常用的降维方法，在不同领域有不同的称谓

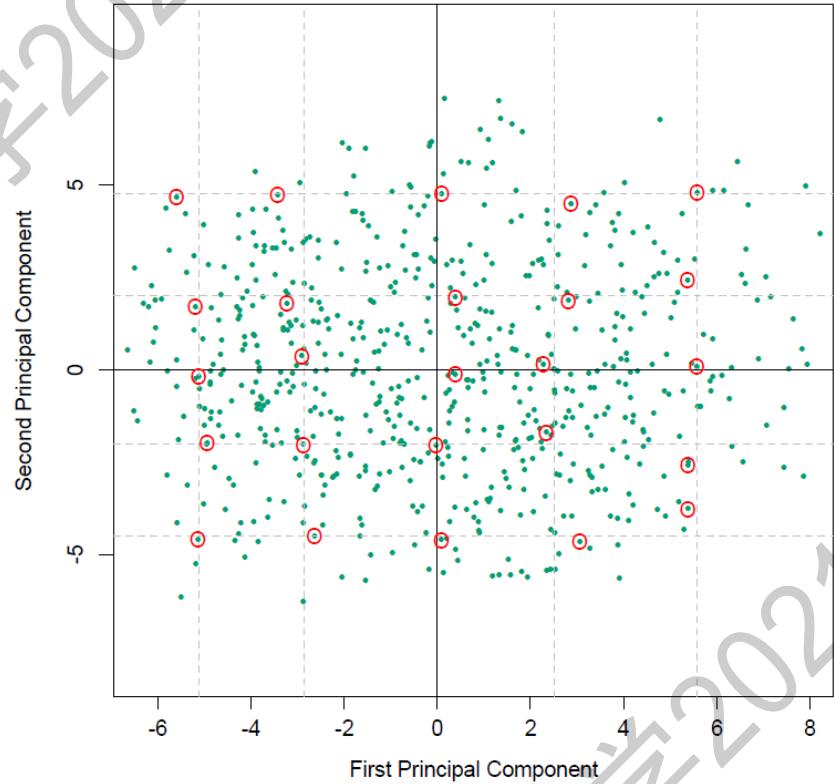
例如在人脸识别中该技术被称为“特征脸”(eigenface)

因为若将前 d' 个特征值对应的特征向量还原为图像，则得到



降维体现在哪里？

PCA 应用 (续)



$$\begin{matrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \mathbf{w}_4 \\ \mathbf{w}_5 \end{matrix}$$

$$\hat{f}(\lambda) = \boxed{\mathbf{w}_1} + \lambda_1 \cdot \boxed{\mathbf{w}_2} + \lambda_2 \cdot \boxed{\mathbf{w}_3}.$$

PCA 应用 (续)

PCA 可有效降低模型的自由度，可用于实现高维空间中的数据分析和可视化。



在前 d' 主成分方向进行随机扰动，从而实现通过较少的参数控制高维空间中图像风格的变化

PCA 的拓展1 - 自编码

从减小重构误差角度理解PCA

$$\min_{W^T W = I} \sum_{i=1}^m \|x_i - Wz_n\|_2^2 = \sum_{i=1}^m \|x_i - \textcolor{blue}{W}\textcolor{red}{W}^\top x_i\|_2^2$$

对W的形式进行推广

$$\min_{f, g} \sum_{i=1}^m \|x_i - \textcolor{blue}{g} \circ \textcolor{red}{f}(x_i)\|_2^2 = \sum_{i=1}^m \|x_i - \textcolor{blue}{decoder} \circ \textcolor{red}{encoder}(x_i)\|_2^2$$

PCA 的拓展2 – Robust PCA

- PCA的解是基于协方差的特征向量，记为 $\Sigma \propto X^T X = U_\Sigma \Lambda_\Sigma U_\Sigma^T$
- X 的奇异值分解记为 $X \approx U_X S_X V_X^T$

PCA的优化目标可改写为

$$\min_{rank(\hat{X})=d'} \|X - \hat{X}\|_2^2$$

Robust PCA

$$\min_{\hat{X}} \|X - \hat{X}\|_0 + rank(\hat{X})$$

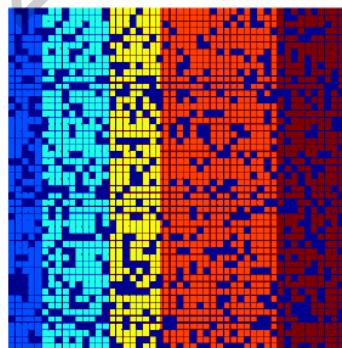


$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$

PCA 的拓展2 – Robust PCA

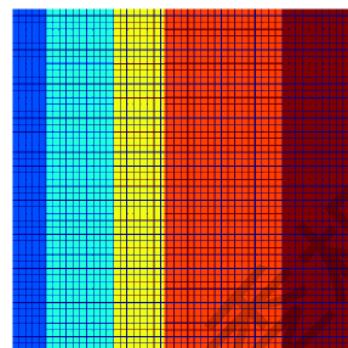
□ Robust PCA

$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$



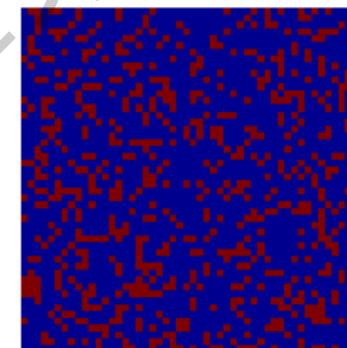
输入数据

=

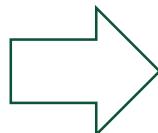
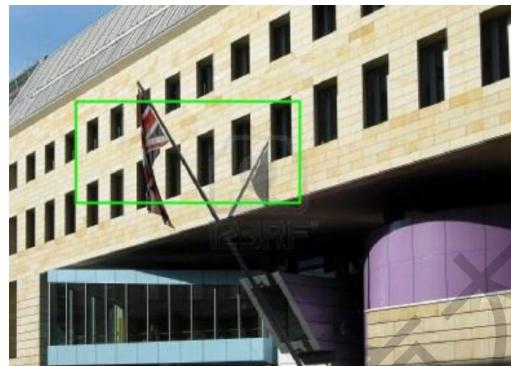


低秩

+



稀疏



PCA 的拓展2 – Robust PCA

□ Robust PCA

$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$



$y \in \mathbb{R}^m$
Test image

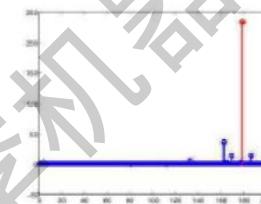
=

$$\begin{bmatrix} \text{faces} \\ \text{faces} \\ \text{faces} \\ \text{faces} \end{bmatrix}$$

$$A = [A_1 \mid A_2 \mid \dots \mid A_k]$$

Combined training dictionary

\times



$x \in \mathbb{R}^n$
coefficients

+



$e \in \mathbb{R}^m$
corruption,
occlusion

PCA 的拓展3 – 函数推广

PCA构建一组正交基，对数据进行重构

$$\begin{aligned} \sum_{i=1}^m \left\| \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j - \mathbf{x}_i \right\|_2^2 &= \sum_{i=1}^m \mathbf{z}_i^T \mathbf{z}_i - 2 \sum_{i=1}^m \mathbf{z}_i^T \mathbf{W}^T \mathbf{x}_i + \text{const} \\ &\propto -\text{tr} \left(\mathbf{W}^T \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{W} \right). \end{aligned}$$

推广到函数空间，对任意（周期）函数 $x(t)$ 进行重构

□ 构建函数正交基

□ $\{\cos n\omega t, \sin n\omega t\}, n = 0, 1, 2, \dots, \infty$

□ 优化重构系数

$$x(t) = \color{red}{a_0} + \sum_{n=1}^{\infty} \color{red}{a_n} \cos n\omega t + \sum_{n=1}^{\infty} \color{red}{b_n} \sin n\omega t$$

$$\begin{aligned} \color{red}{a_n} &= \frac{\langle x, \cos n\omega t \rangle}{\langle \cos n\omega t, \cos n\omega t \rangle} \\ &= \frac{2}{T} \int_{t_0}^{t_0+T} x(t) \cos n\omega t \, dt \end{aligned}$$

PCA 的拓展3 – 函数推广

傅里叶级数：推广到函数空间，对任意（周期）函数 $x(t)$ 进行重构

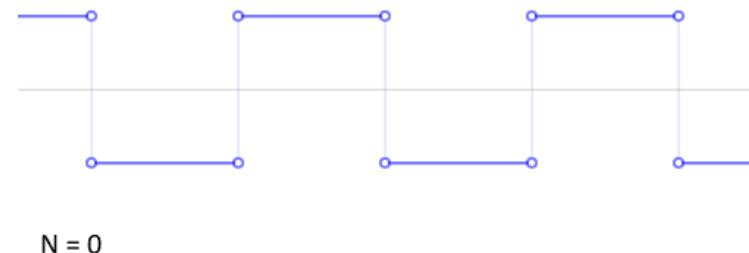
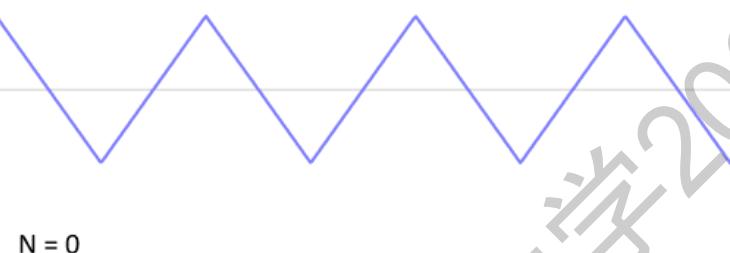
□ 构建函数正交基

□ $\{\cos n\omega t, \sin n\omega t\}, n = 0, 1, 2, \dots, \infty$

□ 优化重构系数

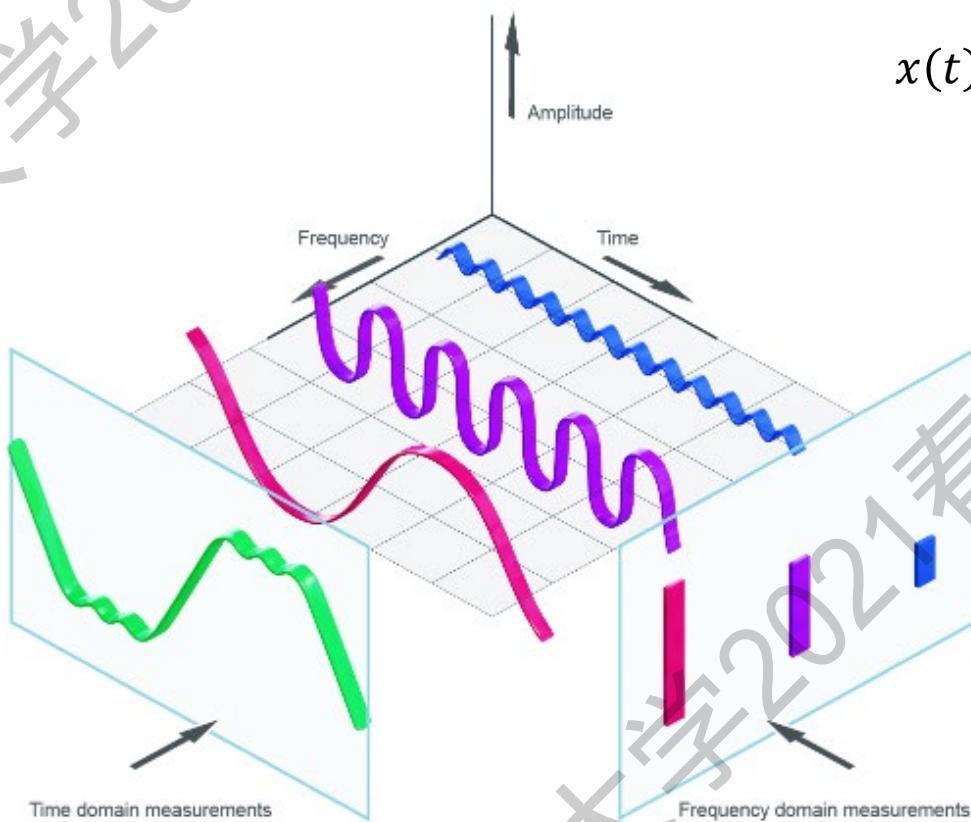
$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t$$

$$\begin{aligned} a_n &= \frac{\langle x, \cos n\omega t \rangle}{\langle \cos n\omega t, \cos n\omega t \rangle} \\ &= \frac{2}{T} \int_{t_0}^{t_0+T} x(t) \cos n\omega t \, dt \end{aligned}$$



PCA 的拓展3 – 函数推广

傅里叶级数：推广到函数空间，对任意（周期）函数 $x(t)$ 进行重构



$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t$$

习题1-1 线性回归

给定数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, 其中 $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id}) \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, 当我们采用线性回归模型求解时, 实际上是在求解下述优化问题:

$$\hat{\mathbf{w}}_{\text{LS}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2, \quad (1)$$

其中 $\mathbf{y} = [y_1, \dots, y_m]^\top \in \mathbb{R}^m$, $\mathbf{X} = [\mathbf{x}_1^\top; \mathbf{x}_2^\top; \dots; \mathbf{x}_m^\top] \in \mathbb{R}^{m \times d}$, $\mathbf{1}$ 为全 1 向量, 其维度可由其他元素推导而得。在实际问题中, 我们常常不会直接利用线性回归对数据进行拟合, 这是因为当样本特征很多, 而样本数相对较少时, 直接线性回归很容易陷入过拟合。为缓解过拟合问题, 常对公式(1)引入正则化项, 通常形式如下:

$$\hat{\mathbf{w}}_{\text{reg}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2 + \lambda \Omega(\mathbf{w}), \quad (2)$$

其中, $\lambda > 0$ 为正则化参数, $\Omega(\mathbf{w})$ 是正则化项, 根据模型偏好选择不同的 Ω 。

习题1-1 线性回归

给定数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, 其中 $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id}) \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, 当我们采用线性回归模型求解时, 实际上是在求解下述优化问题:

$$\hat{\mathbf{w}}_{\text{LS}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2, \quad (1)$$

$$\hat{\mathbf{w}}_{\text{reg}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2 + \lambda \Omega(\mathbf{w}), \quad (2)$$

□ 闭式解

$$\mathbf{w}_{\text{LS}}^* = \left(\mathbf{X}^\top \mathbf{X} - \frac{1}{m} \mathbf{X}^\top \mathbf{1} \mathbf{1}^\top \mathbf{X} \right)^{-1} \left(\mathbf{X}^\top - \frac{1}{m} \mathbf{X}^\top \mathbf{1} \mathbf{1}^\top \right) \mathbf{y}$$

$$\mathbf{b}_{\text{LS}}^* = \frac{1}{m} (\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X} \mathbf{w}_{\text{LS}}^*)$$

$$\hat{\mathbf{w}}_{\text{Ridge}}^* = \left(\mathbf{X}^\top \mathbf{X} - \frac{1}{m} \mathbf{X}^\top \mathbf{1} \mathbf{1}^\top \mathbf{X} + 2\lambda \mathbf{I}_d \right)^{-1} \left(\mathbf{X}^\top - \frac{1}{m} \mathbf{X}^\top \mathbf{1} \mathbf{1}^\top \right) \mathbf{y}$$

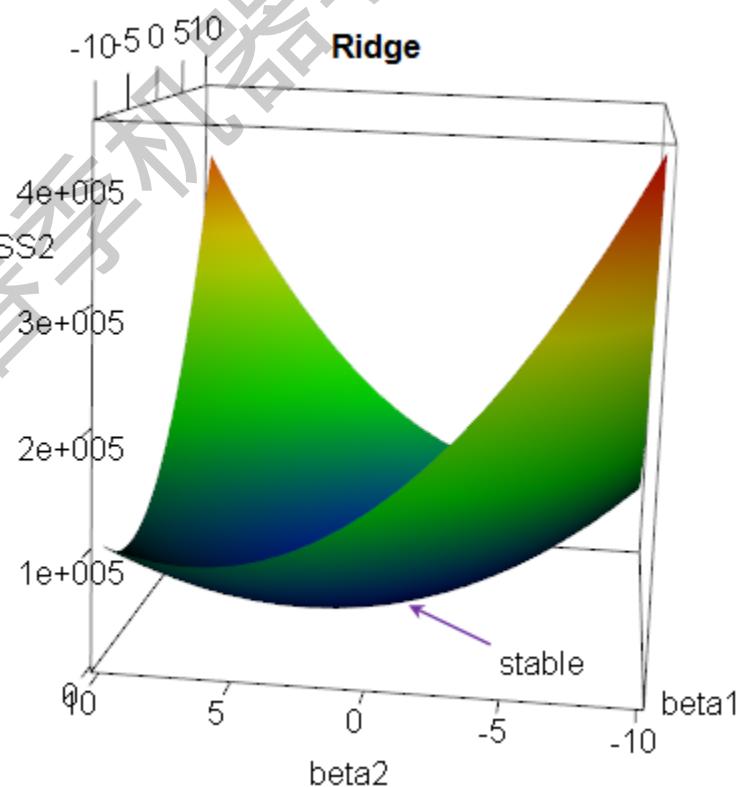
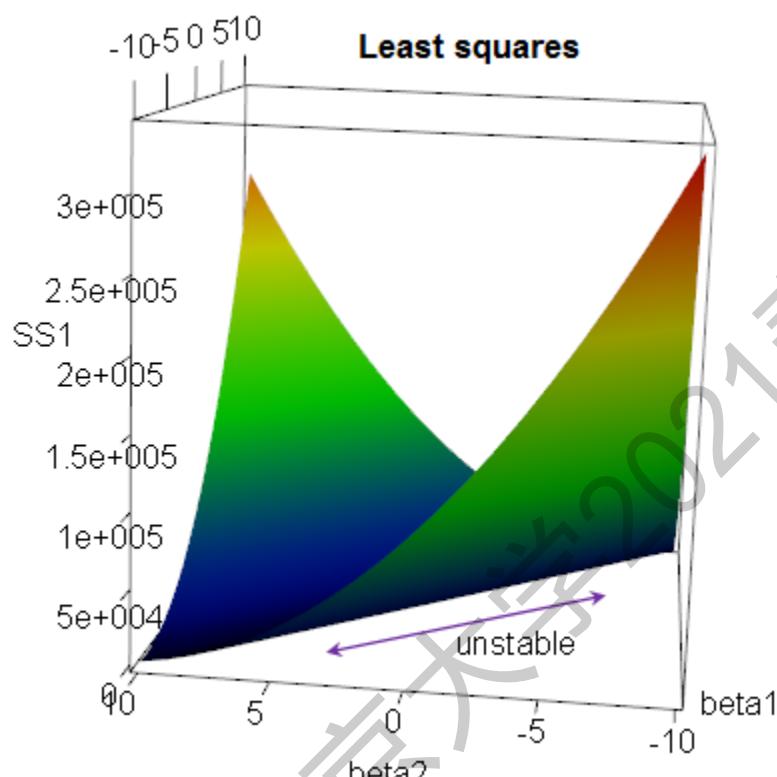
$$\mathbf{b}_{\text{Ridge}}^* = \frac{1}{m} (\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X} \hat{\mathbf{w}}_{\text{Ridge}}^*)$$

习题1-1 线性回归

□ 闭式解

$$\hat{\mathbf{w}}_{\text{Ridge}}^* = \left(\mathbf{X}^\top \mathbf{X} - \frac{1}{m} \mathbf{X}^\top \mathbf{1} \mathbf{1}^\top \mathbf{X} + 2\lambda \mathbf{I}_d \right)^{-1} \left(\mathbf{X}^\top - \frac{1}{m} \mathbf{X}^\top \mathbf{1} \mathbf{1}^\top \right) \mathbf{y}$$

$$b_{\text{Ridge}}^* = \frac{1}{m} (\mathbf{1}^\top \mathbf{y} - \mathbf{1}^\top \mathbf{X} \hat{\mathbf{w}}_{\text{Ridge}}^*)$$



习题1-1 线性回归

给定数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, 其中 $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id}) \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, 当我们采用线性回归模型求解时, 实际上是在求解下述优化问题:

$$\hat{\mathbf{w}}_{\text{LS}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2, \quad (1)$$

$$\hat{\mathbf{w}}_{\text{reg}}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{X}\mathbf{w} + \mathbf{1}\mathbf{b}^\top - \mathbf{y}\|_2^2 + \lambda \Omega(\mathbf{w}), \quad (2)$$

- 分类问题

- 直接分类
- One-hot 编码
- 其他编码方式

- 高维处理

$$(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1} \mathbf{X} = \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda\mathbf{I})^{-1}$$

习题3-2 支持向量机

在本题中，我们复习支持向量机（SVM）的推导过程。考虑 N 维空间中的二分类问题，即 $\mathbb{X} = \mathbb{R}^N$, $\mathbb{Y} = \{-1, +1\}$ 。现在，我们从某个数据分布 \mathcal{D} 中采样得到了一个包含 m 个样本的数据集 $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 。令 $h : \mathbb{X} \rightarrow \mathbb{Y}$ 表示某个线性分类器，即 $h \in \mathcal{H} = \{\mathbf{x} \rightarrow \text{sign}(\mathbf{w}^\top \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$ 。

- (1) [3pts] 对于一个样本 (\mathbf{x}, y) ，请用包含 $\mathbf{x}, y, \mathbf{w}, b$ 的不等式表达“该样本被分类正确”；
- (2) [3pts] 我们知道，线性分类器 h 对应的方程 $\mathbf{w}^\top \mathbf{x} + b = 0$ 确定了 N 维空间中的一个超平面。令 $\rho_h(\mathbf{x})$ 表示点 \mathbf{x} 到由 h 确定的超平面的欧式距离，试求 $\rho_h(\mathbf{x})$ ；
- (3) [4pts] 定义分类器 h 的间隔 $\rho_h = \min_{i \in [m]} \rho_h(\mathbf{x}_i)$ 。现在，我们希望在 \mathcal{H} 中寻找“能将所有样本分类正确且间隔最大”的分类器。试写出该优化问题。我们将该问题称为问题 \mathcal{P}^1 ；

习题3-2 支持向量机

在本题中，我们复习支持向量机（SVM）的推导过程。考虑 N 维空间中的二分类问题，即 $\mathbb{X} = \mathbb{R}^N$, $\mathbb{Y} = \{-1, +1\}$ 。现在，我们从某个数据分布 \mathcal{D} 中采样得到了一个包含 m 个样本的数据集 $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 。令 $h : \mathbb{X} \rightarrow \mathbb{Y}$ 表示某个线性分类器，即 $h \in \mathcal{H} = \{\mathbf{x} \rightarrow \text{sign}(\mathbf{w}^\top \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$ 。

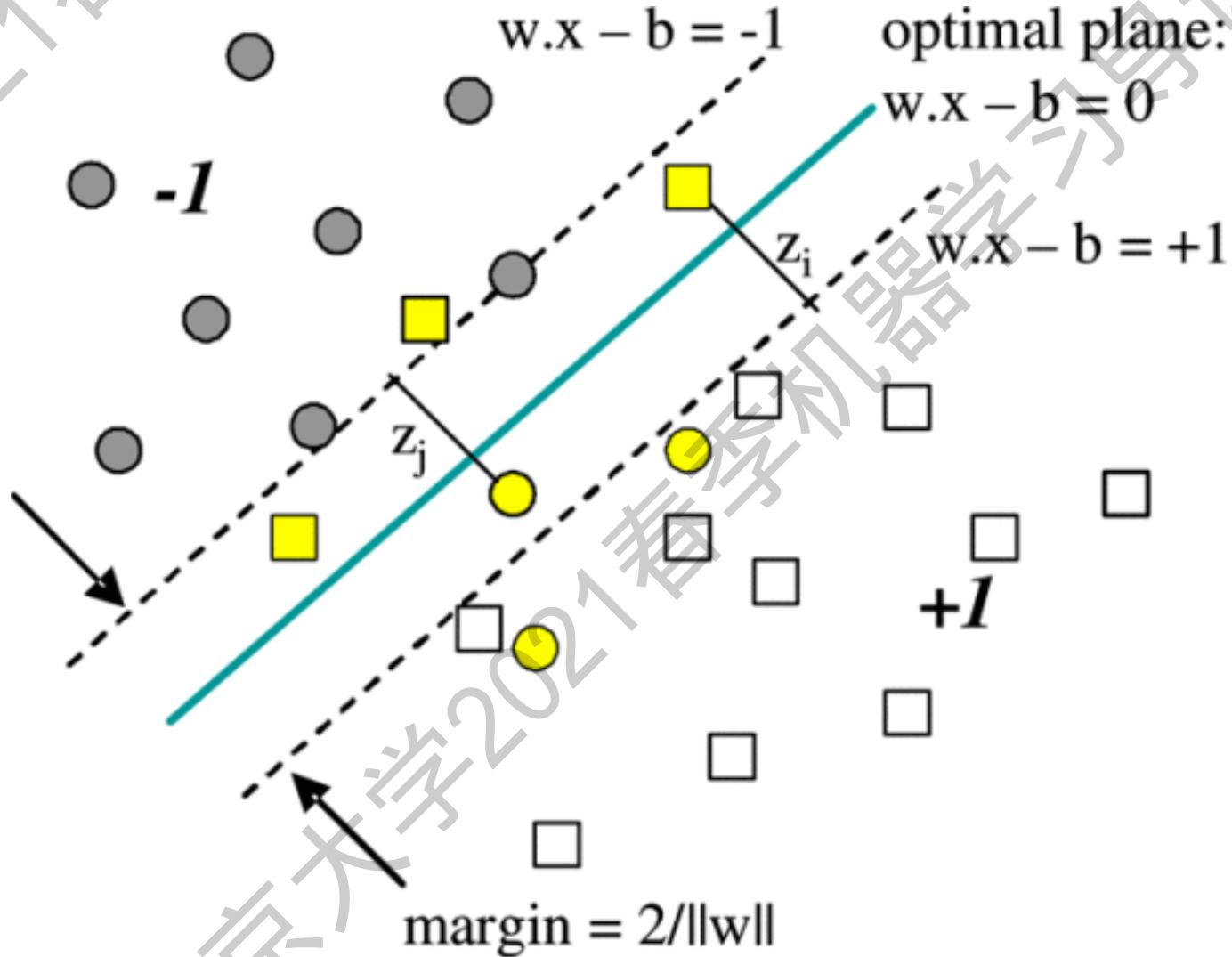
- (10) [10pts] 设 $\{\alpha_i^*\}_{i=1}^m$ 是对偶问题 \mathcal{P}^4 的最优解，设 (\mathbf{w}^*, b^*) 是原问题 \mathcal{P}^3 的最优解。请使用 $\{\alpha_i^*\}_{i=1}^m$ 和 $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 表达 \mathbf{w}^* 和 b^* ，给出过程；
- (11) [10pts] 利用上一小问中的结果，经过一些代数运算，我们发现：可以只用 $\{\alpha_i^*\}_{i=1}^m$ 简洁地表达 $\|\mathbf{w}^*\|^2$ 。请写出这个表达，给出推导过程；
- (12) [5pts] 我们注意到，在问题 \mathcal{P}^2 中，分类器的间隔由式子 $\frac{1}{\|\mathbf{w}\|}$ 表达。再结合上一小问的结果，你可以得到何种启发？

(10) $\mathbf{w}^* = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$; 对于任意一个支持向量 \mathbf{x}_i ，有 $b = y_i - \sum_{j=1}^m \alpha_j y_j \mathbf{x}_j^\top \mathbf{x}_i$ 。

(11) $\|\mathbf{w}^*\|^2 = \|\boldsymbol{\alpha}^*\|_1$

(12) 最大化间隔相当于最小化 $\boldsymbol{\alpha}$ 的 ℓ_1 范数，可以解释支持向量的稀疏性。

LS-SVM和SVM



习题1-2 多分类对率回归

教材的章节3.3介绍了对数几率回归解决二分类问题的具体做法。假定现在的任务不再是二分类问题，而是多分类问题，其中 $y \in \{1, 2, \dots, K\}$ 。请将对数几率回归算法拓展到该多分类问题。

- (1) [15pts] 给出该对率回归模型的“对数似然”(log-likelihood);
- (2) [10pts] 请仿照课本公式3.30，计算该“对数似然”的梯度；
- (3) [Bonus 5pts] 对于样本类别分布不平衡的问题，基于以上的推导会出现怎样的问题，应该进行怎样的应对？谈谈你的看法。

提示1：假设该多分类问题满足如下 $K - 1$ 个对数几率，

$$\begin{aligned}\ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^T \mathbf{x} + b_1 \\ \ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^T \mathbf{x} + b_2 \\ &\dots \\ \ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^T \mathbf{x} + b_{K-1}\end{aligned}$$

提示2：定义指示函数 $\mathbb{I}(\cdot)$,

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{若 } y \text{ 等于 } j \\ 0 & \text{若 } y \text{ 不等于 } j \end{cases}$$

习题1-2 多分类对率回归

教材的章节3.3介绍了对数几率回归解决二分类问题的具体做法。假定现在的任务不再是二分类问题，而是多分类问题，其中 $y \in \{1, 2, \dots, K\}$ 。请将对数几率回归算法拓展到该多分类问题。

- (1) [15pts] 给出该对率回归模型的“对数似然”(log-likelihood);
- (2) [10pts] 请仿照课本公式3.30，计算该“对数似然”的梯度；

□ 归一化

$$p(y = i | \mathbf{x}) = \frac{e^{\beta_i^T \hat{\mathbf{x}}}}{1 + \sum_{k=1}^{K-1} e^{\beta_k^T \hat{\mathbf{x}}}}, i = 1, 2, \dots, K-1$$

$$p(y = K | \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k^T \hat{\mathbf{x}}}}$$

$$\ell(\beta) = \sum_{i=1}^m \sum_{k=1}^K \mathbb{I}(y_i = k) \ln(p(y_i = k | x_i))$$

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta_j} &= \sum_{i=1}^m \sum_{k=1}^K \left(\mathbb{I}(y_i = j) \ln \left(e^{\beta_i^T \hat{\mathbf{x}}} \right) - \mathbb{I}(y_i = j) \ln \left(1 + \sum_{k=1}^K e^{\beta_k^T \hat{\mathbf{x}}} \right) \right) \\ &= \sum_{i=1}^m \sum_{k=1}^K \left(\mathbb{I}(y_i = j) \hat{\mathbf{x}}_i - \mathbb{I}(y_i = j) \ln \left(1 + \sum_{k=1}^K e^{\beta_k^T \hat{\mathbf{x}}} \right) \right) \\ &= \sum_{i=1}^m \left(\sum_{k=1}^K \mathbb{I}(y_i = j) \hat{\mathbf{x}}_i - p(y_i = j | x_i) \hat{\mathbf{x}}_i \right) \\ &= \sum_{i=1}^m \hat{\mathbf{x}}_i (\mathbb{I}(y_i = j) - p(y_i = j | \hat{\mathbf{x}}_i))\end{aligned}$$

习题2-1 线性判别分析

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的 K 分类问题, $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ 为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \dots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k , 类概率 $\Pr(y = k) = \pi_k$. 若 $\mathbf{X} \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$

- (1) [6pts] (贝叶斯最优分类器) 从贝叶斯决策论的角度出发, 对样本 \mathbf{x} 做出的最优预测应为 $\arg \max_y \Pr(y | \mathbf{x})$. 因此, 只需考察 $\ln \Pr(y = k | \mathbf{x})$ 的大小, 即可得到贝叶斯最优分类器, 这也正是推导LDA的一种思路. 请证明: 在题给假设下, $\arg \max_y \Pr(y | \mathbf{x}) = \arg \max_k \delta_k(\mathbf{x})$, 其中 $\delta_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln \pi_k$ 为LDA在分类时的判别式.

$$\begin{aligned} \arg \max_y \Pr(y | \mathbf{x}) &= \arg \max_y \ln \Pr(y | \mathbf{x}) \\ &= \arg \max_y \ln \Pr(y) \Pr(\mathbf{x} | y) &= \arg \max_k \delta_k(\mathbf{x}). \\ &= \arg \max_y \ln \pi_y + \ln \Pr(\mathbf{x} | y) \end{aligned}$$

习题2-1 线性判别分析

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的 K 分类问题, $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ 为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \dots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k , 类概率 $\Pr(y = k) = \pi_k$. 若 $\mathbf{X} \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$

- (2) [6pts] 假设 $K = 2$, 记 $\hat{\pi}_k = \frac{m_k}{m}$, $\hat{\boldsymbol{\mu}}_k = \frac{1}{m_k} \sum_{y_i=k} \mathbf{x}_i$, $\hat{\boldsymbol{\Sigma}} = \frac{1}{m-K} \sum_{k=1}^K \sum_{y_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$. LDA 使用这些经验量替代真实参数, 计算判别式 $\delta_k(\mathbf{x})$ 并按照第(1)问中的准则做出预测. 请证明: 在 $\mathbf{x}^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) > \frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) - \ln(m_2/m_1)$ 时 LDA 将样本预测为第 2 类.

- (2) 要预测为第 2 类, 只需 $\delta_2(\mathbf{x}) > \delta_1(\mathbf{x})$. 令 $\hat{\delta}_2(\mathbf{x}) > \hat{\delta}_1(\mathbf{x})$, 化简整理可得

$$\mathbf{x}^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) > \frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) - \ln(m_2/m_1).$$

习题2-1 线性判别分析

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的 K 分类问题, $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ 为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \dots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k , 类概率 $\Pr(y = k) = \pi_k$. 若 $\mathbf{X} \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$

- (3) [16pts] (线性回归) 考虑第(2)问中的二分类问题, 并将第 1 类样本的标记 y 设为 $-\frac{m}{m_1}$, 将第 2 类样本的标记 y 设为 $\frac{m}{m_2}$. 仿照线性回归, 得到下列优化问题:

$$\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^m (y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta})^2. \quad (2)$$

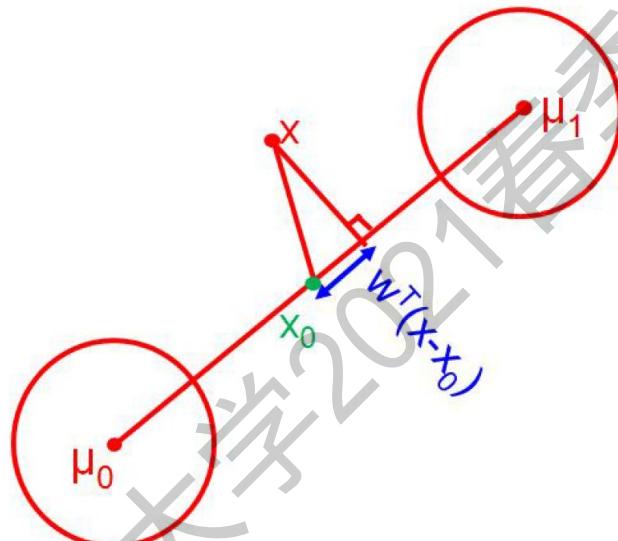
请证明: 上述优化问题的最优解满足 $\boldsymbol{\beta}^* \propto \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, 即通过线性回归解得的 \mathbf{x} 系数与第(2)问中LDA的判别规则表达式中的 \mathbf{x} 系数同向.

习题2-1 线性判别分析

- (3) [16pts] (线性回归) 考虑第(2)问中的二分类问题，并将第1类样本的标记 y 设为 $-\frac{m}{m_1}$ ，将第2类样本的标记 y 设为 $\frac{m}{m_2}$ 。仿照线性回归，得到下列优化问题：

$$\min_{\beta, \beta_0} \sum_{i=1}^m (y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta})^2. \quad (2)$$

请证明：上述优化问题的最优解满足 $\boldsymbol{\beta}^* \propto \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$ ，即通过线性回归解得的 \mathbf{x} 系数与第(2)问中LDA的判别规则表达式中的 \mathbf{x} 系数同向。



Geometry of LDA in the 2 class case where $\Sigma_1 = \Sigma_2 = \mathbf{I}$.

习题2-1 线性判别分析

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的 K 分类问题, $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ 为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \dots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k , 类概率 $\Pr(y = k) = \pi_k$. 若 $\mathbf{X} \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$

请回答下列问题:

- (4) [6pts] (对率回归) 通过课本的介绍可知对率回归假设对数几率为特征 \mathbf{x} 的线性函数, 而由第(1)问可知, 在LDA 中, 对数几率 $\ln \frac{\Pr(y=k|\mathbf{x})}{\Pr(y=l|\mathbf{x})}$ 也可以写成 $\beta_0 + \mathbf{x}^T \boldsymbol{\beta}$ 的形式, 从这一角度来看, 这两种模型似乎是相同的? 哪种模型做出的假设更强? 请说明理由.

习题2-1 线性判别分析

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的 K 分类问题, $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ 为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \dots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k , 类概率 $\Pr(y = k) = \pi_k$. 若 $\mathbf{X} \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$

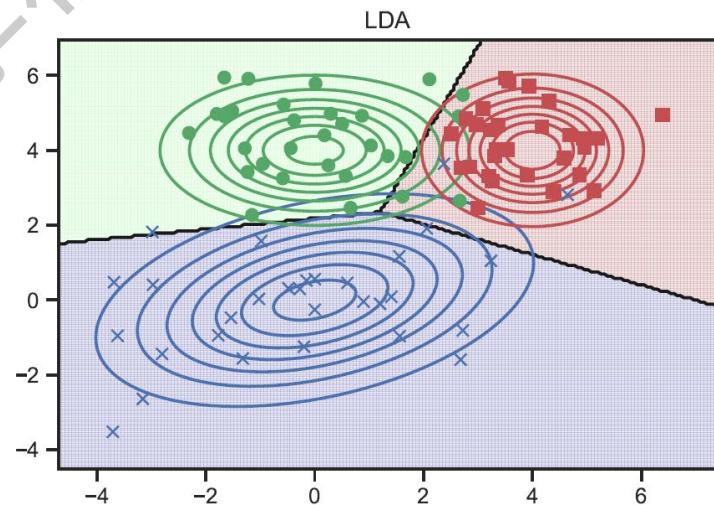
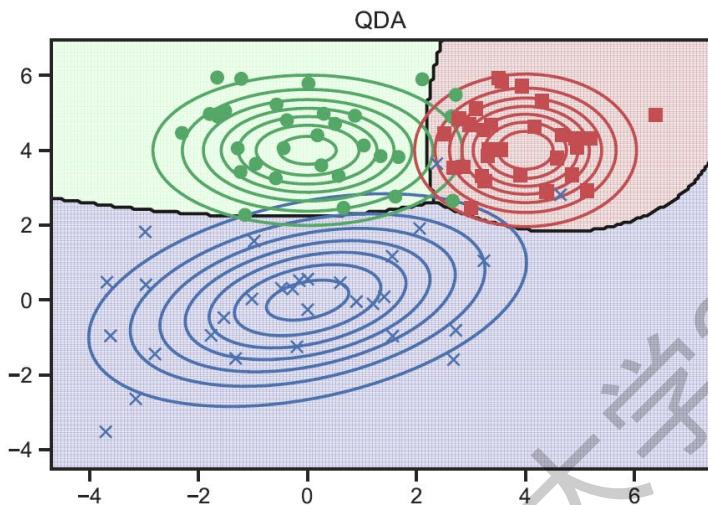
请回答下列问题:

- (5) [6pts] (二次判别分析) 假设各类样本仍服从正态分布, 但第 k 类样本从 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 中独立同分布采样得到, 即不假设各类的协方差矩阵相同. 请按照第①问中的思路, 给出分类应采用的判别式 $\delta_k(\mathbf{x})$, 使得 $\arg \max_y \Pr(y | \mathbf{x}) = \arg \max_k \delta_k(\mathbf{x})$. 此时判别式是一个关于 \mathbf{x} 的二次函数, 这一做法被称为二次判别分析 (Quadratic Discriminant Analysis, QDA).

习题2-1 线性判别分析

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的 K 分类问题, $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ 为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \dots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k , 类概率 $\Pr(y = k) = \pi_k$. 若 $X \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1)$$



习题2-2 广义瑞利商

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}. \quad (5)$$

(1) [15pts] 请证明: 瑞利商满足

$$\lambda_{\min}(\mathbf{A}) \leq \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \leq \lambda_{\max}(\mathbf{A}), \quad (6)$$

其中 \mathbf{A} 为实对称矩阵, $\lambda(\mathbf{A})$ 为 \mathbf{A} 的特征值.

Solution. (1) 设 $\mathbf{A}\boldsymbol{\xi}_i = \lambda_i \boldsymbol{\xi}_i$, $i = 1, 2, \dots, n$, $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_n\}$ 构成一组单位正交基, 从而存在一组 $\{\alpha_i\}_{i=1}^n$ 使得

$$\mathbf{w} = \sum_{i=1}^n \alpha_i \boldsymbol{\xi}_i,$$

代入瑞利商的定义, 可得

$$\frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \frac{(\sum_i \alpha_i \boldsymbol{\xi}_i^T) \mathbf{A} (\sum_i \alpha_i \boldsymbol{\xi}_i)}{(\sum_i \alpha_i \boldsymbol{\xi}_i^T) (\sum_i \alpha_i \boldsymbol{\xi}_i)} = \frac{\sum_i \alpha_i^2 \lambda_i}{\sum_i \alpha_i^2}.$$

可见瑞利商是 \mathbf{A} 的特征值的加权平均, 所以

$$\lambda_{\min}(\mathbf{A}) \leq \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \leq \lambda_{\max}(\mathbf{A}).$$

习题2-2 广义瑞利商

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}. \quad (5)$$

- (2) [15pts] 请证明: 如果 \mathbf{A} 为实对称矩阵, \mathbf{B} 为正定矩阵, 那么广义瑞利商满足

$$\lambda_{\min}(\mathbf{B}^{-1} \mathbf{A}) \leq \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \leq \lambda_{\max}(\mathbf{B}^{-1} \mathbf{A}). \quad (7)$$

- (2) 对 \mathbf{B} 进行正交对角化, 可得 $\mathbf{B} = \mathbf{P}^T \boldsymbol{\Lambda} \mathbf{P}$, 其中 \mathbf{P} 为特征向量构成的正交矩阵, $\boldsymbol{\Lambda}$ 为对角线元素为相应特征值 (正数) 的对角矩阵. 于是,

$$\frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} = \frac{\tilde{\mathbf{w}}^T \mathbf{C} \tilde{\mathbf{w}}}{\tilde{\mathbf{w}}^T \tilde{\mathbf{w}}},$$

其中 $\tilde{\mathbf{w}} = \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{P} \mathbf{w}$, $\mathbf{C} = \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{P} \mathbf{A} \mathbf{P}^T \boldsymbol{\Lambda}^{-\frac{1}{2}}$. 由前一问知

$$\lambda_{\min}(\mathbf{C}) \leq \frac{\tilde{\mathbf{w}}^T \mathbf{C} \tilde{\mathbf{w}}}{\tilde{\mathbf{w}}^T \tilde{\mathbf{w}}} \leq \lambda_{\max}(\mathbf{C}).$$

因此 \mathbf{C} 与 $\mathbf{B}^{-1} \mathbf{A}$ 有相同的特征值, 所以

$$\lambda_{\min}(\mathbf{B}^{-1} \mathbf{A}) \leq \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \leq \lambda_{\max}(\mathbf{B}^{-1} \mathbf{A}).$$