数学基础部分

1.1 令 $\boldsymbol{x}=(\sqrt{3},1)^T,\boldsymbol{y}=(1,\sqrt{3})^T$ 为两个向量，$\boldsymbol{x}_\perp$ 表示 $\boldsymbol{x}$ 在 $\boldsymbol{y}$ 上的投影.

(a) $\boldsymbol{x}_\perp$ 的值是多少?

(b) 证明 $\boldsymbol{y}\perp(\boldsymbol{x}-\boldsymbol{x}_\perp)$。

(c) 画草图来说明上述向量之间的关系。

(d) 证明对于任意 $\lambda\in\mathbb{R}$ 有 $||\boldsymbol{x}-\boldsymbol{x}_\perp||\le||\boldsymbol{x}-\lambda\boldsymbol{y}||$。(提示：这些向量之间的几何关系表明 $||\boldsymbol{x}-\boldsymbol{x}_\perp||^2+||\boldsymbol{x}_\perp-\lambda\boldsymbol{y}||^2=||\boldsymbol{x}-\lambda\boldsymbol{y}||^2$)

1.2 令 X 为一个 5×5 的实对称矩阵，其特征值为 $1,1,3,4$ 和 x。

(a) 求关于 x 的一个充要条件，使得 X 为正定矩阵。

(b) 若 $\det(X)=72,x$ 的值是多少?

1.3 令 $\boldsymbol{x}$ 为一个 d 维随机向量，并且 $\boldsymbol{x}\sim N(\boldsymbol{\mu},\sum)$。

(a) 令 $p(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 的概率密度函数。求 $p(\boldsymbol{x})$ 的表达式。

(b) 求 $\ln p(\boldsymbol{x})$ 的表达式。

(c) 如果你有《The Matrix Cookbook》一书，其中哪条公式将有助于你求解 $\frac{\partial\ln p(\boldsymbol{x})}{\partial\boldsymbol{\mu}}$？其值为多少?

(d) 相似地，如果我们将 $\sum^{-1}$(而不是 $\sum$) 看作变量，哪条公式 (或哪些公式) 将有助于你求解 $\frac{\partial\ln p(\boldsymbol{x})}{\partial\sum^{-1}}$? 其值为多少?

评估

2.1 在一个二分类问题中，已知 $P=N=100$ (即在测试集中有 100 个正例和 100 个负例)。如果 $FPR=0.3,TPR=0.2$，那么查准率、查全率、F_1 值各是多少？它的准确率和错误率是多少?

2.2 (AUC-PR 和 AP) 我们尚未讨论过 AUC-PR 度量的计算细节。对于二分类任务而言，我们假设每个样本 $\boldsymbol{x}$ 都有一个得分 $f(\boldsymbol{x})$，并按照这些得分对测试样本进行降序排序，然后，对于每个样本，我们将分类阈值设置为当前样本的得分（即只有当前样本以及之前的样本会被分为正类)。在该阀值处可以计算得到一对查准率和查全率。PR 曲线是通过连接相邻的点绘制得到的。AUC-PR 就是 PR 曲线下的面积。

令 (r_i,p_i) 表示第 i 对查全率和查准率 $(i=1,2,...)$。在计算面积时，r_i 和 r_{i-1} 之间的面积是使用梯形插值 $(r_i-r_{i-1})\frac{p_i+p_{i-1}}{2}$ 计算得到的，其中 r_i-r_{i-1} 表示在 x 轴上的长度，p_i 和 p_{i-1} 是两根垂直线段在 y 轴上的长度。针对所有 i 的值求和，我们得到了 AUC-PR 值。请注意，我们假设第一对 $(r_0,p_0)=(0,1)$, 这是对应于阈值 $+\infty$ 的一个伪匹配对。

(a) 对于表 1所示的 10 个测试样本 (下标从 1 到 10), 当阈值设为当前样本的 $f(x_i)$ 值时，计算查准率 (p_i) 和查全率 (r_i) 的值。令类别 1 为正类，补全表 1中的其他值。将梯形近似值 $(r_i-r_{i-1})\frac{p_i+p_{i-1}}{2}$ 填入第 i 行的 “AUC-PR” 一列; 将其总和填值入最后一行。

(b) 平均精度（Average Precision, AP）是另外一种能将 PR 曲线概括为数字的方法。与 AUC-PR 类似，AP 使用矩形来近似 r_i 和 r_{i-1} 之间的面积，为 $(r_i-r_{i-1})p_i$。将此近似值填入第 i 行的 “AP” 一列中；并将其总和填入最后一行。AUC-PR 和 AP 都是对 PR 曲线的总结，因此它们的值应该彼此相似。是吗?

表 1: AUC-PR 和 AP 的计算

下标	类别标记	得分	查准率	查全率	AUC-PR	AP
0			1.0000	0.0000	-	-
1	1	1.0				
2	2	0.9				
3	1	0.8				
4	1	0.7				
5	2	0.6				
6	1	0.5				
7	2	0.4				
8	2	0.3				
9	1	0.2				
10	2	0.1				
					(?)	(?)

(c) AUC-PR 和 AP 都对标记的顺序很敏感。如果交换一下第 9 行和第 10 行的类别标记，那么新的 AUC-PR 和 AP 是多少?

(d) 基于类别标记、得分和正类，编程计算 AUC-PR 和 AP 的值。使用表 1中的测试样本集来验证你程序的正确性。

2.3 (贝叶斯决策理论) 考虑一个二分类任务，其中标记 $y \in \{1, 2\}$。如果一个样本 $x \in \mathbb{R}$ 属于类别 1，那么它是由类条件分布（class conditional distribution）$p(x|y=1) = N(-1, 0.25)$ 生成的，第 2 类的样本从类条件分布 $p(x|y=2) = N(1, 0.25)$ 中采样得到。假设 $\Pr(y=1) = \Pr(y=2) = 0.5$。

(a) p.d.f. $p(x)$ 是多少?

(b) 让我们使用代价矩阵 $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ 如果我们选择 $f(x) = \arg\max_y p(y|x)$ 作为 x 的预测，证明对于任意 x，代价 $\mathbb{E}_{(x,y)}[c_{y,f(x)}]$ 是最小化的，因此它是最优解。如果 $y \in \{1, 2, \ldots, C\}$ $(C > 2)$ (即在多分类问题中)，此规则是否仍是最优?

(c) 使用代价矩阵 $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ 和贝叶斯决策理论，对于该任务而言，哪个分类策略是最优的?

(d) 如果代价矩阵为 $\begin{bmatrix} 0 & 10 \\ 1 & 0 \end{bmatrix}$ (即当真实标记为 1 但预测标记为 2 时，代价增加到了 10)。新的决策规则是什么?

FLD

3.1 (矩阵的秩) 令 A 表示一个大小为 $m \times n$ 的实矩阵。它的秩（rank）被记为 $\text{rank}(A)$，其定义是 A 中线性无关（linearly independent）的行的最大个数，也称为行秩（row rank）。类似地，列秩（column rank）是 A 中线性无关的列的最大个数。

(a) 证明行秩等于列秩。因此，

$$\text{rank}(A) = \text{rank}(A^T).$$

(b) 证明

$$\text{rank}(A) \le \min(m, n).$$

(c) 令 X 与 Y 表示两个相同大小的矩阵。证明

$$\text{rank}(X+Y) \leq \text{rank}(X) + \text{rank}(Y).$$

(d) 证明只要矩阵乘法是良定义的，就有

$$\text{rank}(XY) \leq \min\big(\text{rank}(X), \text{rank}(Y)\big).$$

(e) 如果一个大小为 $m \times n$ 的矩阵 X 满足

$$\text{rank}(X) = \min(m, n),$$

则称该矩阵为**满秩的** (full rank)。证明如果 X 是满秩的，则有

$$\text{rank}(X) = \text{rank}(XX^T) = \text{rank}(X^TX).$$

(f) 令 $\boldsymbol{x}$ 表示一个向量，$\boldsymbol{x}\boldsymbol{x}^T$ 的秩是多少？证明实对称矩阵 X 的秩等于其非零特征值的数量。

3.2 (求解 FLD) 考虑一个 C ($C > 2$) 类的 FLD 问题，该问题中包含 N 个 D 维的样本。

(a) 如果 $N < D$，S_W 可逆吗？

(b) 证明最多可以获得 $C - 1$ 个广义特征向量。(提示：S_B 的秩是多少？)

(c) 请解释为什么在 $N > D$，特别是 $N \gg D$ 的情况下，S_W 很可能是可逆的？

(d) 当 S_W 可逆时，广义特征值问题 $S_B\boldsymbol{w} = \lambda S_W\boldsymbol{w}$ 等价于 $S_W^{-1}S_B\boldsymbol{w} = \lambda\boldsymbol{w}$。但在实际问题中，我们并不会对 $S_W^{-1}S_B\boldsymbol{w} = \lambda\boldsymbol{w}$ 进行求解。一种用于求解 FLD 的算法如下所示。

第一步，先假设 S_W 是正定的，利用 Cholesky 分解找到 G 使得 $S_W = GG^T$。

第二步，计算 $C = G^{-1}S_BG^{-T}$。

第三步，对角化 C，找到正交矩阵 Q，使得 Q^TCQ 为对角矩阵。

最后，令 $X = G^{-T}Q$。

证明 X^TS_BX 是对角矩阵，并且 X^TS_WX 是单位阵。证明广义特征向量位于 X 的列中，广义特征值位于 X^TS_BX 的对角线上。根据这种方式计算得到的广义特征向量有单位范数吗？是正交的吗？

3.3 (PCA+FLD 人脸识别) 在人脸识别中，PCA 与 FLD 都很有用。

(a) ORL 是人脸识别领域中较早的一个数据集。从 `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html` 下载 ORL 数据集。阅读下载页面中的说明并了解该数据集的格式。

(b) OpenCV 是一个开源计算机视觉库，它提供了很多关于各种计算机视觉应用的实用函数。从 `http://opencv.org/` 下载该库。从 `http://docs.opencv.org/` 上学习 OpenCV 的基础知识。

(c) 可在网址 `http://docs.opencv.org/2.4/modules/contrib/doc/facerec/facerec.tutorial.html` 找到一个关于人脸识别的 OpenCV 教程。尝试理解教程中的每一行代码，特别是那些关于 Eigenface (PCA) 和 Fisherface (FLD) 的代码。在 ORL 数据集上运行该实验，分析这些方法得到的识别结果之间的差异。

(d) 在 Eigenface 实验中，你可以使用 eigenfaces (特征脸，即特征向量) 来重构近似的人脸图像。修改 OpenCV 教程中的源代码，并使用不同数量的 eigenfaces 来观察可视化的结果。如果你希望从 eigenfaces 中重构的人脸看上去与原始输入的人脸图像之间难以区分，那么你需要多少张 eigenfaces？