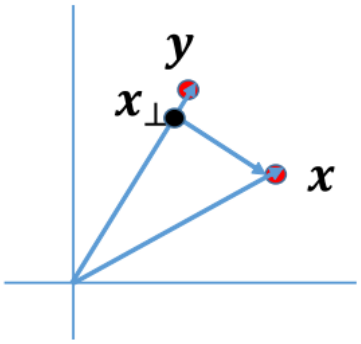
2.1

解:

(a) $x_\perp = \frac{x^T y}{\|y\|^2} y = \frac{2\sqrt{3}}{4}(1, \sqrt{3})^T = \left(\frac{\sqrt{3}}{2}, \frac{3}{2}\right)^T$.

(b) $x - x_\perp = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)^T$, 并且 $y^T(x - x_\perp) = \frac{\sqrt{3}}{2} - \frac{\sqrt{3}}{2} = 0$.

(c)



(d) 记 $c = \frac{x^T y}{\|y\|^2}$, 可得 $x_\perp = cy$。此外, 令 $z = \lambda y$, 我们可以计算 $\|x-z\|^2 - \|x-x_\perp\|^2$ 该式等价于 $\|x-\lambda y\|^2 - \|x-cy\|^2 = (\lambda^2 - c^2)\|y\|^2 - 2(\lambda - c)x^T y$。由于 $y \perp (x - x_\perp)$, 即意味着 $y^T(x - x_\perp) = 0$ 或 $y^T x = y^T x_\perp = cy^T y = c\|y\|^2$。因此, $\|x-z\|^2 - \|x-x_\perp\|^2 = (\lambda^2 - c^2)\|y\|^2 - 2(\lambda - c)c\|y\|^2 = (\lambda - c)^2\|y\|^2 = \|\lambda y - x_\perp\|^2 \geq 0$。由此可得 $\|x - x_\perp\|^2 \leq \|x - \lambda y\|^2$ 或等价地, $\|x - x_\perp\| \leq \|x - \lambda y\|$。

2.2

解:

(a) x > 0.

(b) x $= 6$.

2.3

解:

(a) $p(\boldsymbol{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$.

(b) $\ln p(\boldsymbol{x}) = -\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$.

(c) "The Matrix Cookbook" 中的公式 (86) 假设 W 为对称矩阵, 并且有 $\frac{\partial}{\partial s}(x - s)^T W(x - s) = -2W(x - s)$。由于 Σ^{-1} 是对称的 (Σ 是对称的), 我们可以利用该式 (用 Σ^{-1} 替换 W, μ 替换 s), 于是 $\frac{\partial \ln p(x)}{\partial \mu} = \Sigma^{-1}(x - \mu)$。

(d) 首先, 请注意 $|\Sigma^{-1}|$ 等价于 $\det(\Sigma^{-1})$。"The Matrix Cookbook" 中的

公式 (57) 表明 $\frac{\partial \ln \det(X)}{\partial X} = (X^{-1})^T = (X^T)^{-1}$，因此可得 $\frac{\partial}{\partial \Sigma^{-1}} \ln |\Sigma^{-1}| = \Sigma$（将 X 替换为 Σ^{-1}）。其次，“The Matrix Cookbook” 中的公式 (72) 表明 $\frac{\partial a^T X a}{\partial X} = aa^T$。将 X 替换为 Σ^{-1}，a 替换为 $(x-\mu)$，我们可以得到 $\frac{\partial}{\partial \Sigma^{-1}}(x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)(x-\mu)^T$。将上述两条结论相结合，有 $\frac{\partial \ln p(x)}{\partial \Sigma^{-1}} = \frac{1}{2}\Sigma - \frac{1}{2}(x-\mu)(x-\mu)^T$。

4.1

解：

$TP = TPR \times P = 20$，$FP = FPR \times N = 30$，$FN = P - TP = 80$，$TN = N - FP = 70$。因此，查准率为 $\frac{TP}{TP+FP} = \frac{20}{50} = 0.4$，查全率等于 $TPR = 0.2$。F1 值为 $\frac{2TP}{2TP+FP+FN} = \frac{40}{40+30+80} = \frac{4}{15}$（或 0.2667）。$ACC = \frac{TP+TN}{P+N} = \frac{20+70}{100+100} = 0.45$，错误率为 $1 - 0.45 = 0.55$。

4.5

解：

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	1.0000	0.2000	0.2000	0.2000
2	2	0.9	0.5000	0.2000	0	0
3	1	0.8	0.6667	0.4000	0.1167	0.1333
4	1	0.7	0.7500	0.6000	0.1417	0.1500
5	2	0.6	0.6000	0.6000	0	0
6	1	0.5	0.6667	0.8000	0.1267	0.1333
7	2	0.4	0.5714	0.8000	0	0
8	2	0.3	0.5000	0.8000	0	0
9	1	0.2	0.5556	1.0000	0.1056	0.1111
10	2	0.1	0.5000	1.0000	0	0
					0.6906	0.7278

(a) 所需的结果已填入表 1 “AUC-PR” 一列。

(b) 所需的结果已填入表 1 “AP” 一列。AP 和 AUC-PR 的值非常接近。

(c) 新的 AUC-PR 值为 0.6794，新的 AP 值为 0.7167。
(d) Matlab/Octave 示例代码为：

```
labels = [1 2 1 1 2 1 2 2 1 2];
scores = [1.0 0.9 0.8 0.7 0.5 0.5 0.4 0.3 0.2 0.1];
posclass = 1;
[~,I] = sort(scores,'descend');
labels = labels(I);
tp = [0 cumsum(labels==posclass)];
fp = [0 cumsum(labels~=posclass)];
prec = tp ./ (tp+fp);
prec(isnan(prec)) = 1;
recall = tp / sum(labels==1);
recall_delta = recall(2:end)-recall(1:end-1);
AUC_PR = 0.5*sum(recall_delta.*(prec(2:end)+prec(1:end-1)));
AP = sum(recall_delta.*prec(2:end));
```

4.7
解：
(a) $p(x) = p(x|y=1)\Pr(y=1) + p(x|y=2)\Pr(y=2) = 0.5N(-1, 0.25) + 0.5N(1, 0.25)$。
(b) 使用该代价矩阵，我们有 $\mathbb{E}_{(x,y)}[c_{y,f(x)}] = \int_{f(x)\neq y} p(x,y)dxdy = \Pr(y=1)\int_{f(x)=2} p(x|y=1)dx + \Pr(y=2)\int_{f(x)=1} p(x|y=2)dx$。对于任意 x，如果我们令 $f(x)=1$，它将使得 $\Pr(y=2)p(x|y=2)$ 成为误差的积分；如果我们令 $f(x)=2$，则会带来误差的是 $\Pr(y=1)p(x|y=1)$。因此，为了最小化代价，如果 $\Pr(y=2)p(x|y=2) > \Pr(y=1)p(x|y=1)$，我们必须预测 $f(x)=2$；否则，我们应该令 $f(x)=1$。换作一种更加简洁的形式，我们必须令 $f(x) = \arg\max_y \Pr(y)p(x|y)$。
因为 $\Pr(y)p(x|y) = p(x,y) = p(y|x)p(x)$，对于不同的 y，$p(x)$ 具有相同的值，最优化的规则可以描述成以下贝叶斯决策规则：$f(x) = \arg\max_y p(y|x)$。在多分类任务中，$\mathbb{E}_{(x,y)}[c_{y,f(x)}] = \sum_{i=1}^{C} \Pr(y=i)\int_{f(x)\neq i} p(x|y=i)dx$。易证最优化规则仍然是 $f(x) = \arg\max_y p(y|x)$。请注意对于多维特征向量该规则仍然成立，即 $f(x) = \arg\max_y p(y|x)$。
(c) 贝叶斯决策规则告诉我们，如果 $p(x|y=1)\Pr(y=1) > p(x|y=2)\Pr(y=2)$，则 $f(x)=1$，否则 $f(x)=2$。代入分布，该条件变为了

$\frac{1}{\sqrt{2\pi}0.5}\exp\left(-\frac{(x-(-1))^2}{2\times 0.25}\right) > \frac{1}{\sqrt{2\pi}0.5}\exp\left(-\frac{(x-1)^2}{2\times 0.25}\right)$，或 $(x+1)^2 < (x-1)^2$。因此，如果 $x<0$，我们应该令 $f(x)=1$，否则的话，$f(x)=2$。该例中的贝叶斯风险或贝叶斯错误为 $0.5\int_{x>0}\frac{1}{\sqrt{2\pi}0.5}\exp\left(-\frac{(x+1)^2}{2\times 0.25}\right)dx + 0.5\int_{x<0}\frac{1}{\sqrt{2\pi}0.5}\exp\left(-\frac{(x-1)^2}{2\times 0.25}\right)dx$。令 $\Phi(\cdot)$ 为 $N(0,1)$ 的累积分布函数，该式的两项均等于 $1-\Phi(2)$。因此，本例中的贝叶斯风险为 $2(1-\Phi(2))=0.0455$，或 4.55%。

(d) 现在 $\mathbb{E}_{(x,y)}[c_{y,f(x)}] = c_{12}\Pr(y=1)\int_{f(x)=2}p(x|y=1)dx + \Pr(y=2)\int_{f(x)=1}p(x|y=2)dx$。因此，如果 $\Pr(y=2)p(x|y=2) < c_{12}\Pr(y=1)p(x|y=1)$，则令 $f(x)=1$，否则 $f(x)=2$。该条件变成了

$$10\times\frac{1}{\sqrt{2\pi}0.5}\exp\left(-\frac{(x-(-1))^2}{2\times 0.25}\right) > \frac{1}{\sqrt{2\pi}0.5}\exp\left(-\frac{(x-1)^2}{2\times 0.25}\right)$$

或 $\ln(10)-2(x+1)^2 > -2(x-1)^2$，或 $x < \frac{\ln(10)}{8}$。分类边界移向了类别 2（从 0 到 $\frac{\ln(10)}{8}$）。这也不奇怪，因为现在类别 1 分类错误的代价更大。

6.1

解：

(a) 将行秩表示为 r $(0\le r\le m)$ 并令 $x_1, x_2, \ldots, x_r$ 表示 A 的行空间的一组基。那么，可以证明向量集 $Ax_1, Ax_2, \ldots, Ax_r$ 是线性独立的（但我们将推迟其证明过程）。因为 Ax_i 是 A 中各列的线性组合（由 x_i 中的各项进行加权），它们都位于 A 的列空间中。Ax_i 之间的独立性表明列秩至少为 r。也就是说，列秩大于或等于行秩。

类似地，将完全相同的过程应用于 A^T，我们可以得出结论，A 的行秩大于或等于列秩。结合这两个结论，唯一的可能就是列秩等于行秩。

独立性证明：如果 Ax_i 不是线性独立的，那么必然存在实数 c_i，使得 $0=\sum_{i=1}^r c_i Ax_i = A(\sum_{i=1}^r c_i x_i)$，并且至少有一个 c_i 不为零。请注意 $\sum_{i=1}^r c_i x_i$ 位于 A 的行空间中。如果 $\sum_{i=1}^r c_i x_i \neq 0$，我们就会发现一个矛盾的现象——$\sum_{i=1}^r c_i x_i$ 垂直于 A 中的每一个行向量，因此垂直于 A 的行空间中的任何向量（包括其自身）——我们必须有 $\sum_{i=1}^r c_i x_i = 0$。因为向量 x_i 构成基，$\sum_{i=1}^r c_i x_i = 0$ 就意味着 $c_1=c_2=\cdots=c_r=0$，这与我们的假设相矛盾。因此，r 个向量 Ax_i 的集合是线性独立的。

(b) 很明显，独立列或行的个数不会大于 n 或 m。因此，$\text{rank(A)} \le \text{min(m,n)}$。

(c) 令 $x_i, 0\le i\le \text{rank}(X)$ $(y_j, 0\le j\le \text{rank}(Y))$ 表示 $X(Y)$ 行空间的基。那么，$X+Y$ 中的任意行都可以写成 x_i 和 y_j 的线性组合。因此 $\text{rank}(X+Y)\le \text{rank}(X)+\text{rank}(Y)$。

(d) 假设 X 和 Y 的大小分别为 $m \times n$ 和 $n \times p$。令 $u_1, u_2, \ldots, u_r$ 表示 X 列空间的基。因为 XY 中第 i 列是 $\sum_{j=1}^{n} y_{ji} x_{:j}$，即这是 X 中各列的线性组合，而它又是向量 u 的线性组合。因此，XY 中的每一列都是向量 u 的线性组合。我们可以得出结论 $\text{rank}(XY) \leq r = \text{rank}(X)$。类似地，$\text{rank}(XY) = \text{rank}(Y^T X^T) \leq \text{rank}(Y^T) = \text{rank}(Y)$。因此，$\text{rank}(XY) \leq \min(\text{rank}(X), \text{rank}(Y))$。

(e) 假设 $m \geq n$ (因此 $\text{rank}(X) = n$)。我们需要证明 $\text{rank}(X^T X) = n$。一方面，$\text{rank}(X^T X) \leq \text{rank}(X) = n$。另一方面，我们来证明 $\text{rank}(X^T X) < n$ 会带来矛盾。$\text{rank}(X^T X) < n$ 意味着存在 $c \in \mathbb{R}^n$ $(c \neq 0)$ 以及 $X^T X c = 0$。在两边左乘 c^T 得到 $c^T X^T X c = 0$。这意味着 $\|Xc\|^2 = 0$，或 $Xc = 0$。但是，因为 $c \neq 0$，$Xc = 0$ 表明 Y 的 n 个列是线性相关的，这与假设 $\text{rank}(Y) = n$ 相矛盾。我们可以得出结论，$\text{rank}(X) = \text{rank}(X^T X)$ (并且 $\text{rank}(X^T X) = \text{rank}(X X^T)$ 总成立)。当 $m < n$ 时，证明过程也是类似的。

(f) $\text{rank}(x x^T) = 1$。$x x^T$ 中的每一列与 x 的区别仅在于标量乘子。令 $x \in \mathbb{R}^d$。谱分解显示 $X = \sum_{i=1}^{r} \lambda_i \xi_i \xi_i^T = E \Lambda E^T$，拥有 r 个非零特征值 λ_i，其对应的 r 个特征向量构成了矩阵 E。换句话说，$\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$，并且 $E \in \mathbb{R}^{d \times r}$。令 $Y = E \Lambda^{\frac{1}{2}}$，这明显是满秩的。根据前一个问题，我们可以得到 $X = Y Y^T$。因此，$\text{rank}(X) = \text{rank}(Y) = r$。

6.5

解:

(a) 由于 S_W 是 N 个秩为 1 的矩阵之和, 其秩 $\leq N < D$。因此, 这个 $D \times D$ 的矩阵不可能是满秩的。

(b) 矩阵 S_B 是 C 个秩为 1 的矩阵 (向量外积) 之和。这 C 个向量 $(m_i - m)$ 不是线性独立的——它们的加权和是 0 (参考公式 6.48 与公式 6.44)。因此，S_B 的秩最多为 $C - 1$。也就是说，最多只有 $C - 1$ 个广义特征向量。

(c) 如果 $N > D$，甚至是 $N \gg D$，假设样本 x_i $(1 \leq i \leq N)$ 构成一个满秩的数据矩阵是合理的 (即，可以找到 D 个独立样本)。因此，在这种情况下假设 $\text{rank}(S_W) = D$ 是合理的。

(d) $X^T S_B X = Q^T G^{-1} S_B G^{-T} Q = Q^T C Q$，这是对角的。

$X^T S_W X = Q^T G^{-1} S_W G^{-T} Q = Q^T (G^{-1} S_W G^{-T}) Q = Q^T I Q = Q^T Q = I$。

请注意，因为 G 是可逆的，从而 X 是可逆的，并且 Q 是正交的。

令 $\Lambda = X^T S_B X$。那么，$X^T S_B X = \Lambda = X^T S_W X \Lambda$。因为 X 是可逆的，我

们有 $S_B X = S_W X\Lambda = \Lambda S_W X$。因此，广义特征向量都在 X 的列中，并且广义特征值存在于 $\Lambda = X^T S_B X$ 的对角线上。
因为在一般情况下，G 是三角阵，且 $\det(G) \neq 1$，因此一般而言，广义特征向量既不是单位范数也不是正交的。
6.6
解：
请使用 OpenCV 来完成实验。关于 eigenface 重构实验, 不同的人可能具有不同的视觉灵敏度。应该在 80 到 250 不等。