データサイエンス特論　授業課題　第二回分（統計、多次元処理）

以下のクイズに対する Python プログラム（または Jupyter notebook）を作成せよ。ただし、適当なパッケージ（例 NumPy, SciPy, scikit-learn, pandas）を使ってよいとする。

(1) 第一回で紹介した、IRIS のデータ（全 150 データ）において、4 次元の属性（SL = Sepal Length, SW =Sepal Width, PL = Petal Length, PW = Petal Width）を 4 次元データと解釈し、平均値と共分散行列を求めよ（以下の式参照）。なお式では、x = SL, y = SW, z = PL, w = PW としている。平均は 4 次元ベクトル、共分散行列は 4 x 4 行列となる。μ は平均、σ は標準偏差、ρ は相関係数を表す。

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \\ \mu_w \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y & \rho_{xz}\sigma_x\sigma_z & \rho_{xw}\sigma_x\sigma_w \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 & \rho_{yz}\sigma_y\sigma_z & \rho_{yw}\sigma_y\sigma_w \\ \rho_{xz}\sigma_x\sigma_z & \rho_{yz}\sigma_y\sigma_z & \sigma_z^2 & \rho_{zw}\sigma_z\sigma_w \\ \rho_{xw}\sigma_x\sigma_w & \rho_{yw}\sigma_y\sigma_w & \rho_{zw}\sigma_z\sigma_w & \sigma_w^2 \end{bmatrix}$$

(2) 次に、IRIS データからランダムに 2 つのデータを選ぶという操作を 3 回行い、毎回、2 つのデータのユークリッド距離とマハラノビス距離をそれぞれ求めよ。なお、ランダムに選ばれたデータに関しては、そのインデックス（0 から 149）、ならびにオリジナルの 4 つの値(SL, SW, PL, PW)とその種類（Setosa, Versicolor, Virginica）をプリントすること。距離は小数点以下 4 桁までとする。なお、資料 2 にマハラノビス距離の定義はあるが、任意の 2 つのデータ x, y が与えられた場合は、そのページの式を $D^2 = (\mathbf{x}-\mathbf{y})^T \mathrm{S}^{-1}(\mathbf{x}-\mathbf{y})$ と解釈すること。乱数の初期値は時刻に同期させるとか、実行するたびに結果を変えること。

【実行例】

```
試行 1
1 つ目のデータ：index = 20, setosa:[5.4 3.4 1.7 0.2]
2 つ目のデータ：index = 33, setosa:[5.5 4.2 1.4 0.2]
Euclid(x,y) = 0.8602
Mahalanobis(x,y) = 2.0827
試行 2
1 つ目のデータ：index = 115, virginica:[6.4 3.2 5.3 2.3]
2 つ目のデータ：index = 16, setosa:[5.4 3.9 1.3 0.4]
Euclid(x,y) = 4.5935
Mahalanobis(x,y) = 2.9022
試行 3
1 つ目のデータ：index = 114, virginica:[5.8 2.8 5.1 2.4]
2 つ目のデータ：index = 54, versicolor:[6.5 2.8 4.6 1.5]
Euclid(x,y) = 1.2450
Mahalanobis(x,y) = 3.8496
```