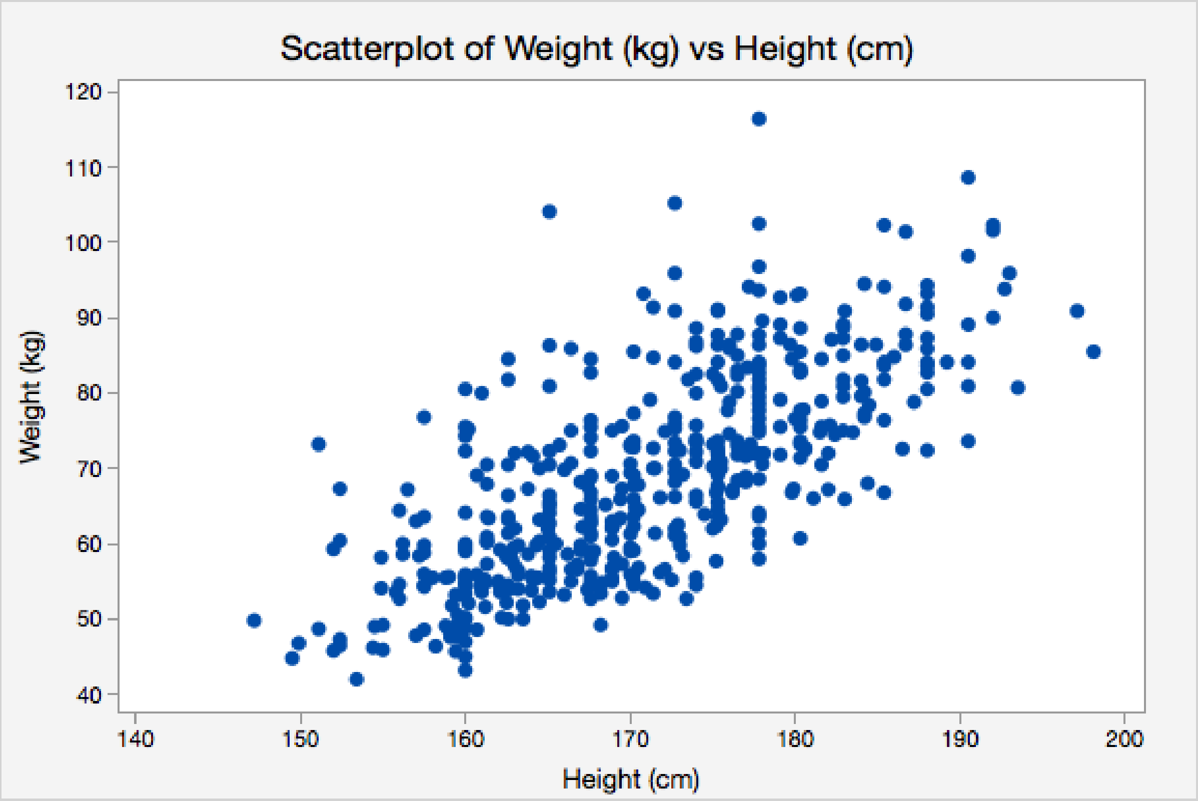
BRIEFING OF JOINT PROBABILITY

Shingchern D. You

Two random variables

- In many situations, we use more than one random variable(RV) to model outcomes
- For example, we want to model the height and weight of a person
- Let RV X represent height and RV Y represent weight, we have a scatterplot like the following

Scatterplot of X and Y



- Source: http://sungsoo.github.io/2014/01/11/scatter-plots.html

Correlation

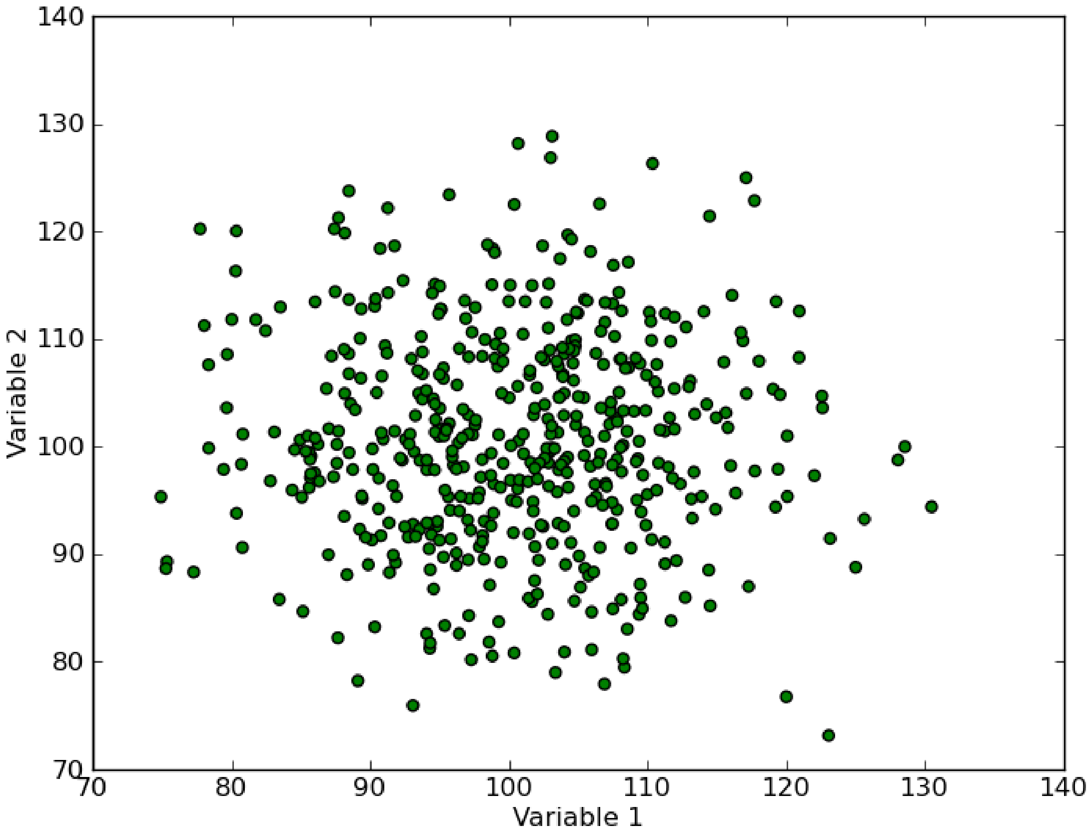
- It is easy to understand that a tall person usually is heavier
- This situation is known as “correlation” between two random variables

Joint probability

- To find the joint probability for RV X and Y, we need joint pdf (probability density function) $f(x, y)$
- If X and Y are independent, then $f(x, y) = f(x)f(y)$ and $E[X, Y] = E[X]E[Y]$
- If $E[X, Y] = E[X]E[Y]$, we say X and Y are uncorrelated, but may not be independent
- For jointly Gaussian, uncorrelated = independent

What if X and Y uncorrelated

- We will see no "trend" on the scatterplot (source: http://sungsoo.github.io/2014/01/11/scatter-plots.html)



Jointly Gaussian

- The following is the pdf for jointly Gaussian for **x** (Exercise: Find out the definition of jointly Gaussian)

$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$$

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

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix

Jointly Gaussian

- Let $x = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}$ be vector of real-valued RV, we have $\mu_i = E[X_i], s_{i,j} = E\left[X_i X_j\right] - \mu_i \mu_j$
- So, $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{\boldsymbol{n}} \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,d} \\ \vdots & \ddots & \vdots \\ s_{d,1} & \cdots & s_{d,d} \end{bmatrix}$
- It is easy to see $s_{i,j} = s_{j,i}$

Jointly Gaussian

- In really, we don't have these parameters
- Sample mean and sample covariance are ML estimates of true mean and true covariance in Gaussian distribution

Sample mean and sample covariance

- A simple computational illustration
- Three RV X, Y, and Z (i.e., $d = 3$)
- We have $x_1, x_2, \ldots, x_4$ from X
  We have $y_1, y_2, \ldots, y_4$ from Y
  We have $z_1, z_2, \ldots, z_4$ from Z
- $\mu_1 = \frac{x_1+x_2+x_3+x_4}{4}, \mu_2 = \frac{y_1+y_2+y_3+y_4}{4}$, etc.

Sample mean and sample covariance

- $s_{1,1} = (x_1 \cdot x_1 + x_2 \cdot x_2 + x_3 \cdot x_3 + x_4 \cdot x_4)/4 - \mu_1 \cdot \mu_1$
- $s_{1,3} = (x_1 \cdot z_1 + x_2 \cdot z_2 + x_3 \cdot z_3 + x_4 \cdot z_4)/4 - \mu_1 \cdot \mu_3$
- Etc.

Simple application

- For iris dataset, it has four dimensions. Data from each dimension are assumed from one RV
- We can then use ML to calculate sample mean and sample covariance for each class in training dataset
- Assign a data point (x_0, y_0, z_0, w_0) to class C_0 if the value of $f(x_0, y_0, z_0, w_0 | C_0)$ is largest (assuming equal class probability). Same as using discriminant function in textbook
- The f function is jointly Gaussian seen before

Regularizing covariance matrix

- Sometimes it is not easy to find inverse of covariance
- Known as ill-conditioned matrix
- You can check the condition number to know if your covariance matrix is ill-conditioned or not

Regularizing covariance matrix

- What can we do
- Method 1: Assume all RV are independent (like Naïve Bayesian). Thus, covariance matrix becomes a diagonal matrix (always invertible)
- Method 2: (MAP, Tikhonov Regularization) Add another matrix to covariance

$$\Sigma = \Sigma + \lambda I$$

where λ is a small positive number (source: http://freemind.pluskid.org/machine-learning/regularized-gaussian-covariance-estimation/)