Variance, Covariance, and Correlation

Terminology review

- Samples and Features
- Tidy Data Matrix

Mean

The sample mean of a feature is

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance

Definition: The (sample) variance of the feature measurements $x_1, \ldots, x_n$ is

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \mu_X^2$$

Covariance

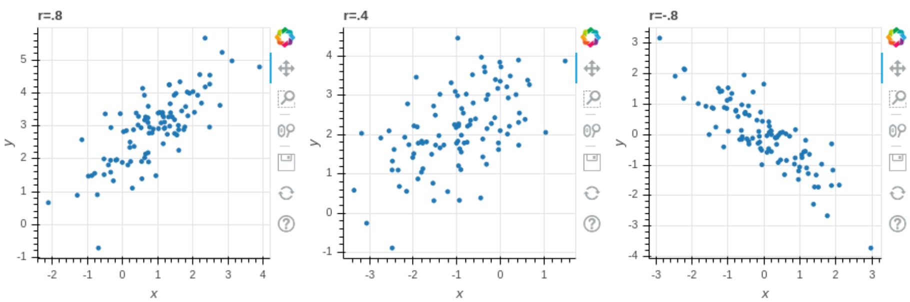
Definition: If $X = (x_1, \dots, x_N)$ and $Y = (y_1, \dots, y_N)$ are two feature vectors then the (sample) covariance is

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)$$

Correlation

Definition: Given feature vectors X and Y, the (sample) correlation coefficient r_{XY} is

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_{XX}\sigma_{YY}}$$



The covariance matrix

Definition: Let X be an $N \times k$ data matrix, and let X_0 be its centered version. The (sample) covariance matrix is the $k \times k$ symmetric matrix

$$D_0 = \frac{1}{N} X_0^{\intercal} X_0.$$

Visualizing the covariance matrix