Statistical Methods for Bioinformatics

Workout

Exercise 5.4.1, derive equation 5.6

- We want to minimize the variance for the following function

$$argmin_{\alpha} var(\alpha X + (1-\alpha)Y)$$

- The variance of a sum of two variables can be rewritten according to the Variance Sum Law ($\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$)

$$= \alpha^2\sigma^2_X + 2\alpha(1-\alpha)\sigma_{XY} + (1-\alpha)^2\sigma^2_Y$$

$$= \alpha^2(\sigma^2_X + \sigma^2_Y - 2\sigma_{XY}) - 2\alpha(\sigma^2_Y + \sigma_{XY}) + \sigma^2_Y$$

- Find derivative in α and the value for which it equals 0

$$2\alpha(\sigma^2_X + \sigma^2_Y - 2\sigma_{XY}) - 2(\sigma^2_Y + \sigma_{XY}) = 0$$

- Rewriting gives

$$\alpha = \frac{\sigma^2_Y - \sigma_{XY}}{\sigma^2_X + \sigma^2_Y - 2\sigma_{XY}}$$

Exercise 5.4.4

- Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

Use bootstrap to make a population of training sets. For every training set use the learning method to generate a predictive model. Use the models to predict Y for the value of interest of X. The standard deviation of the prediction can then be estimated from the variability in the predicted values.