Solutions to example questions:

1. The k-means algorithm stops when the partition to clusters stays the same. This happens when the partition is

(i) $C_1 = (1),\ C_2 = (2, 2.5, 3)$; the centers are 1, 2.5. The loss is $0 + 2 \cdot \frac{1}{2^2} = \frac{1}{2}$.

(ii) $C_1 = (1, 2),\ C_2 = (2.5, 3)$; the centers are 1.5, 2.75. The loss is $2 \cdot \frac{1}{2^2} + 2 \cdot \frac{1}{4^2} = \frac{5}{8}$.

For example the partition $C_1 = (1, 2, 2.5),\ C_2 = (3)$ is not possible since the centers are $\frac{11}{6}, 3$ and 2.5 is closer to 3 than to $\frac{11}{6}$. The optimal partition is (i): $C_1 = (1),\ C_2 = (2, 2.5, 3)$.

2. a. Let Δ be a random variables that assumes the values 1 and 2; Δ represents the cluster from which the observation is drawn. We have that the conditional density is $f(x|\Delta = 1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-x^2}{2\sigma_1^2}\right)$ and $f(x|\Delta = 2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{-x^2}{2\sigma_2^2}\right)$. Also, $P(\Delta = 1) = \pi$. Therefore, the density of X is

$$f(x) = f(x|\Delta = 1)P(\Delta = 1) + f(x|\Delta = 2)P(\Delta = 2) = \pi \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-x^2}{2\sigma_1^2}\right) + (1-\pi)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{-x^2}{2\sigma_2^2}\right),$$

and the log likelihood is

$$\ell(X_1, \ldots, X_n; \pi, \sigma_1^2, \sigma_2^2) = \sum_{i=1}^n \log\left\{\pi \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-X_i^2}{2\sigma_1^2}\right) + (1-\pi)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{-X_i^2}{2\sigma_2^2}\right)\right\}.$$

b. Start with initial estimate $\hat{\pi}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$. The EM algorithm iterates between the E- and M- steps.

The E step is to compute estimate of Δ_i (the cluster of observation i) when the value of parameters is $\hat{\pi}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$. Using Bayes rule, the estimate is

$$\begin{aligned}\hat{P}(\Delta_i = 1) = P_{\hat{\pi},\hat{\sigma}_1^2,\hat{\sigma}_2^2}(\Delta_i = 1|X_i) &= \frac{P_{\hat{\pi},\hat{\sigma}_1^2,\hat{\sigma}_2^2}(X_i|\Delta_i = 1)P_{\hat{\pi}}(\Delta_i = 1)}{P_{\hat{\pi},\hat{\sigma}_1^2,\hat{\sigma}_2^2}(X_i|\Delta_i = 1)P_{\hat{\pi}}(\Delta_i = 1) + P_{\hat{\pi},\hat{\sigma}_1^2,\hat{\sigma}_2^2}(X_i|\Delta_i = 2)P_{\hat{\pi}}(\Delta_i = 2)} \\ &= \frac{\pi \frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}} \exp\left(\frac{-X_i^2}{2\hat{\sigma}_1^2}\right)}{\pi \frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}} \exp\left(\frac{-X_i^2}{2\hat{\sigma}_1^2}\right) + (1-\pi)\frac{1}{\sqrt{2\pi\hat{\sigma}_2^2}} \exp\left(\frac{-X_i^2}{2\hat{\sigma}_2^2}\right)}.\end{aligned}$$

The M step is based on computing maximum likelihood estimates based on $\hat{P}(\Delta_1 = 1), \ldots, \hat{P}(\Delta_n = 1)$. If $Y_1, \ldots, Y_m$ are i.i.d $N(0, \sigma^2)$, then the MLE is $\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^m Y_i^2$. Hence, if we knew $\Delta_1, \ldots, \Delta_n$ then the estimates would be

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n I(\Delta_i = 1)Y_i^2}{\sum_{i=1}^n I(\Delta_i = 1)},\ \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n I(\Delta_i = 2)Y_i^2}{\sum_{i=1}^n I(\Delta_i = 2)}.$$

Since the Δ's are unknown the estimates of Δ are plugged-in and the new estimates of $\pi, \sigma_1^2, \sigma_2^2$ are

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n \hat{P}(\Delta_i = 1)Y_i^2}{\sum_{i=1}^n \hat{P}(\Delta_i = 1)},\ \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n \{1 - \hat{P}(\Delta_i = 1)\}Y_i^2}{\sum_{i=1}^n \{1 - \hat{P}(\Delta_i = 1)\}},\ \hat{\pi} = \frac{\sum_{i=1}^n \hat{P}(\Delta_i = 1)}{n}.$$

3. The Lasso estimates $\hat{\beta}_0, \dots, \hat{\beta}_p$ are the minimizers of the function

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Since β_0 does not appear in the penalty term the minimizer is obtained when the derivative of the first part is zero, i.e., when

$$2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j X_{ij}) = 0.$$

dividing the last equation by 2n yields the desired result.

4. a. According to model A, $Y_i = \beta_0 + \varepsilon_i$ and the LSE is $\hat{\beta}_0 = \bar{Y}$. By a theorem from class an unbiased estimate of the testing error is

$$\widehat{MSE}_{te}(A) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2\hat{\sigma}^2 \cdot 1/n = TSS/n + 2\hat{\sigma}^2/n.$$

Similarity,

$$\widehat{MSE}_{te}(B) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 + 2\hat{\sigma}^2 \cdot 2/n = RSS/n + 4\hat{\sigma}^2/n.$$

Model A is selected iff

$$\widehat{MSE}_{te}(A) < \widehat{MSE}_{te}(B) \iff TSS + 2\hat{\sigma}^2 < RSS + 4\hat{\sigma}^2.$$

b. We showed in class that $R^2 = \frac{TSS-RSS}{TSS} = r^2$. Therefore, $TSS - RSS = r^2 TSS$.

c. We have that

$$r^2 TSS = \frac{\left[\sum_{i=1}^{n}(X_i - \bar{X})Y_i\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \bar{X})Y_i\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n}(X_i - \bar{X})^2\hat{\beta}_1^2.$$

Therefore, Model A is selected iff

$$TSS + 2\hat{\sigma}^2 < RSS + 4\hat{\sigma}^2 \iff TSS - RSS < 2\hat{\sigma}^2 \iff \hat{\beta}_1^2 < \frac{2\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$