שאלות לדוגמה למבחן:

1. נתונות ארבע תצפיות חד מימדיות שערכיהן 1,2,2.5,3. מריצים עליהן אלגוריתם *k-means* עם *k=2*.
   a. מהן החלוקות הסופיות האפשריות?
   b. מהי החלוקה האופטימאלית?

2. נתונות תצפיות חד מימדיות $X_1, \dots, X_n$ שהן עירוב של שני משתנים נורמליים עם תוחלת 0 ושונות לא ידועה.
   a. מהי פונקציית $\log likelihood$ של הנתונים. (לכתוב תוך שימוש בפרמטרים π, σ_1, σ_2)
   b. תארו אלגוריתם מסוג *EM* שאומד את הפרמטרים π, σ_1, σ_2. הסבירו את תשובתכם.

3. נתון מודל רגרסיה רב ממדית $\mathrm{Y}_i = \beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j + \varepsilon_i \quad i = 1, \dots, n$
   יהיו $\hat{\beta}_0^L, \hat{\beta}_1^L, \dots, \hat{\beta}_p^L$ אומדי *LASSO* עבור פרמטר λ מסויים.
   הראו שמתקיים:
   $$\hat{\beta}_0^{\mathrm{L}} = \overline{\mathrm{Y}} - \sum_{\mathrm{j}=1}^{\mathrm{p}} \hat{\beta}_{\mathrm{j}}^{\mathrm{L}} \bar{X}_j$$
   כאשר $\overline{\mathrm{Y}} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} Y_i$ ו- $\overline{\mathrm{X}}_{\mathrm{j}} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} X_{i,j}$.

4. נתון מודל רגרסיה פשוטה: $\mathrm{Y}_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$. שוקלים שני מודלים: מודל *A* עם חותך בלבד (כלומר $\mathrm{Y}_i = \beta_0 + \varepsilon_i$) ומודל *B* המודל המלא (עם *X* ועם חותך). הראו שלפי הקריטריון של C_p נבחר במודל *A* אם ורק אם $\hat{\beta}_1^2 < \frac{2\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

הדרכה:

a. הראו שנבחר במודל *A* אם ורק אם $\mathrm{TSS} + 2\hat{\sigma}^2 < RSS + 4\hat{\sigma}^2$ כאשר

$$\mathrm{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \ , \mathrm{RSS} = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X\_i\right)^2$$

b. העזרו בתוצאה מהכיתה כדי להראות ש $\mathrm{TSS} - \mathrm{RSS} = \mathrm{TSS}\, r^2$ כאשר

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

c. הראו ש $TSS\ r^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2\, \hat{\beta}_1^2$.