STA 35C Statistical Data Science III

Final exam solution

Instructor: Dogyoon Song

Problem 1 (16 points total). Multiple-choice Questions

(a) Linear regression: R^2

- ■ R^2 measures the proportion of variability in Y explained by the model.
- □ R^2 can take on any values between -1 and 1.
- ■ If $R^2 = 1$, then the model's predicted values match the actual response perfectly.
- □ A large R^2 always guarantees excellent out-of-sample performance.

(b) Classification thresholds & errors

- □ In a discriminative approach (e.g. logistic regression), we explicitly model $p(X \mid Y)$ and $p(Y)$.
- □ A false negative occurs when the actual class is negative (Y=0) but we predict positive (Y=1).
- □ A generative method (e.g. LDA) often models $p(X \mid Y)$ and $p(X)$.
- ■ Lowering p^* generally increases the number of predicted positives.

(c) Best subset selection

- ■ Best subset selection fits all possible subsets of predictors of each size to find the best subset.
- □ It always returns the model with the lowest test error among all subsets.
- ■ It can suffer from high computational cost if the number of predictors is large.
- □ It cannot choose a different best model size than backward stepwise selection.

(d) Cross-validation

- □ k-fold cross-validation splits the data into k folds, using 1 fold for training and $k-1$ folds for testing.
- ■ Leave-one-out cross-validation is a special case of k-fold where each fold contains exactly 1 observation.
- ■ Cross-validation can be used to select tuning parameters or compare models' predictive performance.
- □ k-fold cross-validation always guarantees the absolute minimum test error.

(e) Regularization

- ☐ Ridge regression penalizes coefficients by their absolute values (ℓ_1-penalty).
- ■ Increasing the regularization parameter λ generally shrinks coefficients toward zero.
- ☐ Lasso always outperforms ordinary least squares in terms of test error.
- ☐ Ridge can set some coefficients exactly to zero, similar to Lasso.

(f) Regression splines: Degree-3 spline

- ■ A degree-3 spline is a piecewise cubic polynomial function.
- ■ It is continuous up to its second derivative at each knot.
- ☐ It must have a continuous third derivative at each knot.
- ☐ "Natural splines" are splines with additional degrees of freedom for a more natural fit.

(g) PCA (Principal Component Analysis)

- ■ The first and the second principal components are orthogonal to each other.
- ☐ The first principal component always aligns with the direction of smallest variance in X.
- ■ PCA can reduce the dimensionality of X by projecting onto a few components.
- ☐ PCA requires a labeled response Y to identify the principal axes.

(h) Clustering

- ☐ Clustering is a supervised learning task.
- ☐ In clustering, each observation has a numeric response Y used to form clusters.
- ■ k-means and hierarchical clustering are two common approaches.
- ☐ Clustering can only be done with at most two features per observation.

Problem 2 (24 points total). True/False with Justification

(a) "If the first principal component explains 80% of the variance in X, then it must be the single best predictor of the response Y among all principal components"

False.

Reason: The first principal component is chosen to capture the most of the variance in X along the direction, not necessarily to maximize predictive power for Y. High variance in X does not guarantee a strong correlation with Y.

(b) "When using cross-validation, having more folds (e.g., $k = 10$ instead of $k = 5$) always guarantees a lower test error."

False.

Reason: Using more folds (e.g., $k = 10$ instead of $k = 5$) typically stabilizes the estimated validation error (across multiple possible random splits of a *given* dataset) and decreases in the estimated test error for a given dataset, however, can increase the variance (variability across multiple datasets). Thus, it does not guarantee a lower actual test error.

(c) "Lowering the decision threshold p^* in a logistic model will generally increase both the false positive rate and the true positive rate."

True.

Reason: Lowering the threshold p^* tends to classify more observations as positive, which increases both the true positive rate and the false positive rate.

(d) "Adding more predictors to a linear model always reduces the test error."

False.

Reason: Adding more predictors can reduce bias but may increase variance; as such, overfitting may occur and raise test error rather than lowering it.

(e) "In k-means clustering, the final partition is guaranteed to be the global minimum of the objective function."

False

Reason: The k-means algorithm can converges to a local (rather than global) minimum depending on initialization, so it is not guaranteed to always find the global minimum.

(f) "When we want to keep the probability of making at least one false rejection below 5%, we typically control the False Discovery Rate (FDR) at 5%."

False.

Reason: The family-wise error rate (FWER) is the probability of at least one false rejection among all tests, whereas the false discovery rate (FDR) is the expected proportion of false rejections among those rejected. Controlling FDR at 5% does not ensure the probability of *any* false rejection is below 5%.

Problem 3 (20 points total). Statistical Learning

(a) (6 points) Define *training error* and *test error*, and explain their roles in fitting and evaluating a model.

- **Training error** measures how well the model fits the data on which it was trained, e.g. via mean squared error or classification error. Models are often fit by minimizing this error.
- **Test error** is the average loss on new, unseen data, reflecting the model's ability to generalize to observations not used in training. This is typically the error used to evaluate the performance of the fitted model on new, unseen test dataset.

(b) (7 points) In the right figure below, state which curve (upper or lower) represents *test error*. Then, explain the bias-variance tradeoff in one or two sentences, describing how (squared) bias, variance, and irreducible error might change as the model flexibility increases. Lastly, choose the model you would use (among three options) and briefly justify.

- The **upper** curve (initially decreasing, then increasing) represents the test error.
- *Bias–variance trade-off:* As model flexibility increases, the squared bias typically decreases, variance increases, and the irreducible error is constant; these combine to form the U-shaped test-error curve.
- *Model choice:* Select the model at the point minimizing the test-error curve (the mid-complexity choice among the three options), where bias and variance are best balanced.

(c) (7 points) State the purpose of cross-validation in one sentence, and describe how *k-fold cross-validation* computes validation error using a training dataset of size n.

- **Purpose:** Cross-validation estimates a model's out-of-sample prediction error using only the training data.
- **Procedure:**
  - Split the training data into k folds of roughly equal size.
  - For each fold $j = 1, \ldots, k$, fit the model on the other $k-1$ folds ($\approx \frac{k-1}{k} n$ points) and compute the loss on fold j ($\approx \frac{n}{k}$ points).
  - Average these k validation losses to obtain the cross-validated error estimate.

  Use: We compare these estimates across different models or tuning-parameter values and pick the one with the lowest estimated test error.

Problem 4 (40 points total + 3 bonus points). Regression

(a) (10 points total). Suppose you have a dataset $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and want to regress Y on X.

(i) **(5 points)** Describe how to obtain least squares estimates for the regression coefficients in the model:

$$\text{(Model 1)} \qquad Y \approx \beta_0 + \beta_1 X.$$

Either write down the objective function to be minimized, or explain visually (e.g. with a scatter plot) how the best-fit line and parameters are determined.

The least squares estimates $\hat{\beta}_0, \hat{\beta}_2$ are obtained by minimizing the **residual sum of squares** (or equivalently, mean squared error, up to normalization) of the model:

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Geometrically, this gives the best-fit line by making the sum of squared vertical distances from each (X_i, Y_i) to the line as small as possible in the ℓ_2 (mean of squares) sense.

(ii) **(5 points)** Suppose the fitted model yields $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 5$. Predict Y at $x_{\text{new}} = 2$, and explain why the actual y-value (y_{new}) may differ from the predicted value.

$$\hat{Y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} = -3 + 5 * 2 = 7.$$

The realized value y_{new} may differ from 7 because of the random error term ε in the data-generating process $Y = \beta_0 + \beta_1 X + \varepsilon$;. The error term ε accounts for omitted (=not included in the model) factors and pure randomness in generating Y.

(b) (15 points total + 3 bonus points). Now consider two more flexible models:

$$\text{(Model 2)} \qquad Y \approx \beta_0 + \beta_1 X + \beta_2 X^2,$$

$$\text{(Model 3)} \qquad Y \approx \begin{cases} \beta_{1,0} + \beta_{1,1} X, & X \le c, \\ \beta_{2,0} + \beta_{2,1} X, & X > c. \end{cases} \qquad (c : \text{hyperparameter})$$

(i) **(5 points)** Explain under what circumstances these might be preferred over a simple linear model (Model 1). Discuss their relative flexibility and possible impacts on bias, variance, or irreducible error.

- **Flexibility.** Adding X^2 term (Model 2) captures global curvature, and is useful when Y varies nonlinearly with X (curved). The piece-wise linear model with a breakpoint c (Model 3) captures a sudden change or "kink" in the slope at $X = c$, and will be preferred when there is an abrupt change in the trend before and after a certain threshold of X.
- **Bias–variance.** Compared with Model 1, both Model 2 and Model 3 can *reduce bias* if the true mean function is curved or has a kink, but at the expense of *increased variance*. If the sample size is sufficiently large to support the extra parameters, total test error may decrease. Irreducible error is unaffected.

(ii) **(5 points)** Suppose you fit Models 1, 2, and 3 by least squares to the same dataset depicted below. Sketch each regression function below. For Model 3, choose a breakpoint c appropriately.

- *Model 1:* A single straight line through the data.
- *Model 2:* A smooth, quadratic shape (parabola) matching the overall trend of the data points.
- *Model 3:* Two line segments with different slopes, possibly (but not necessarily) meeting at a breakpoint c.

(iii) **(5 points)** Define a *cubic spline* and a *natural cubic spline*, specifying the constraints they must satisfy and how they differ.

- **Cubic spline:** A function that is piece-wise cubic between pre-chosen knots, with continuity in the function, first derivative, and second derivative at every knot.
- **Natural cubic spline:** A cubic spline with additional constraints forcing the spline to be linear beyond the boundary knots (i.e. its second derivative is zero outside the boundary region). This prevents erratic extrapolation (e.g. oscillations) near the boundaries.

(iv*) (3 bonus points*) What is a *smoothing spline*? Briefly explain what the fitted curve look like in the two limits: (a) as $\lambda \to 0$ and (b) as $\lambda \to \infty$.

For any $\lambda > 0$, a smoothing spline refers to the function $\hat{f}_\lambda$ such that

$$\hat{f}_\lambda = \arg\min_{g \in \mathcal{C}^2} \Big\{ \sum_{i=1}^{n} \big(Y_i - g(X_i)\big)^2 \; + \; \lambda \int \big(g''(t)\big)^2 \, dt \Big\},$$

where C^2 means the set of all smooth functions (twice continuously differentiable, to be precise). Here, the parameter λ balances data fidelity and smoothness. For any λ, the resulting function is a natural cubic spline with cut points at $x_1, \ldots, x_n$.

- As $\lambda \to 0$, the penalty is negligible, and the spline $\hat{f}_\lambda$ interpolates all data (becoming very wiggly).
- As $\lambda \to \infty$, the smoothness penalty dominates and the spline $\hat{f}_\lambda$ converges to the *linear* least-squares line (second derivative forced to zero everywhere).

(c) (15 points total). Finally, you want to predict Y using three predictors: D (binary, e.g. treatment indicator), X_1, and X_2, all numeric except D being categorical (0/1). You fit three models:

$$\begin{aligned} &\text{Model A: } Y \approx \alpha_0 + \alpha_1 \, D, \\ &\text{Model B: } Y \approx \beta_0 + \beta_1 \, X_1, \\ &\text{Model C: } Y \approx \gamma_0 + \gamma_1 \, D + \gamma_2 \, X_1 + \gamma_3 \, X_2, \end{aligned}$$

and obtain the following R^2 values and coefficient estimates:

$$\begin{aligned} &\text{Model A: } R^2 = 0.4 \quad \alpha_1 = 2, \text{ with SE} = 0.75; \\ &\text{Model B: } R^2 = 0.6 \quad \beta_1 = 4, \text{ with SE} = 1.25; \\ &\text{Model C: } R^2 = 0.9 \quad \gamma_1 = -4, \; \gamma_2 = 1, \; \gamma_3 = 3, \text{ with SEs } (1.33, \; 2.5, \; 0.5), \quad \operatorname{corr}(X_1, X_2) = 0.8. \end{aligned}$$

(i) **(5 points)** Which model likely has the best *predictive* fit? Define R^2 and explain its meaning briefly.
Model C ($R^2 = 0.9$) explains 90% of the variability in Y and likely gives the best predictive fit. By definition,

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

measures the proportion of total variance in Y explained by the model (i.e., predictors).

(ii) **(5 points)** How do we interpret β_1 in Model B? How does this differ from interpreting γ_2 in Model C?

- In Model B, β_1 quantifies the expected change in Y for a one-unit increase in X_1, ignoring (=not accounting for) other variables.
- In Model C, γ_2 measures the expected change in Y for a one-unit increase in X_1, *holding D and X_2 fixed*, and thus, quantifies a partial effect.

(iii) **(5 points)** Based on the information above, discuss what these data suggest about Y's relationship to D, X_1, X_2. (*Hint:* Discuss the significance of correlation, sign of the effect, etc.)

- D **and** Y: The positive value $\alpha_1 = 2$ (SE 0.75) in Model A looks significant, but might have been confounded by X_2. In Model C, $\gamma_1 = -4$ (SE 1.33) indicates a roughly 4-unit decrease in Y for $D = 1$ versus $D = 0$, after adjusting for X_1 and X_2.
- X_1 **and** Y: $\gamma_2 = 1$ (SE 2.5) is not clearly significant, likely due to the strong correlation (0.8) between X_1 and X_2. Its effect was more apparent in Model B, which did not include X_2, however, the seemingly significant association $\beta_1 = 4$ (SE 1.25) in Model B is likely due to the indirect channel through X_2. In conclusion, the unique contribution of X_1 on Y is uncertain.
- X_2 **and** Y: The large R^2 and the highly significant $\gamma_3 = 3$ (SE 0.5) suggest X_2 is a strong positive predictor for Y after adjusting for others.
- Overall, Y responds most clearly to X_2 and to D (negative effect); direct evidence for an independent effect of X_1 on Y is weak once X_2 is included in the model.

Problem 5 (40 points total + 3 bonus points). Classification

(a) (10 points total). Consider a two-dimensional logistic regression model:

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \qquad \text{where} \quad p(X) = \Pr[Y = 1 \mid X].$$

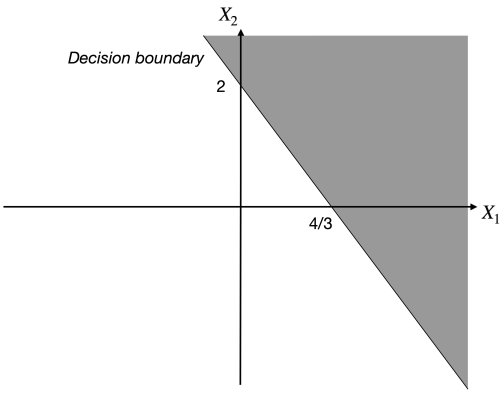
Suppose the estimated coefficients are $\hat{\beta}_0 = -4,\ \hat{\beta}_1 = 3,\ \hat{\beta}_2 = 2$.

(i) **(5 points)** For $x_{\text{test}} = (1, 1)$, compute $\hat{p}(x_{\text{test}})$ and decide whether $\hat{y}_{\text{test}} = 1$ or 0 using the decision rule "$\hat{y} = 1$ if and only if $\hat{p}(x) \geq p^* = 0.5$".

$$\hat{p}(x_{\text{test}}) = \frac{1}{1 + e^{-(-4+3\cdot 1+2\cdot 1)}} = \frac{1}{1+e^{-1}} \approx 0.731.$$

Since $0.731 \geq 0.5 = p^*$, predict $\hat{y}_{\text{test}} = 1$.

(ii) **(5 points)** In the figure below, draw the decision boundary for $p^* = 0.5$, label intercepts, and indicate the region where $\hat{Y} = 1$.



$$\begin{aligned} \hat{y} = 1 &\iff \hat{p}(x) \geq 0.5 \\ &\iff -4 + 3x_1 + 2x_2 \geq 0 \\ &\iff x_2 \geq -\frac{3}{2}x_1 + 2. \end{aligned}$$

Points lying above the line (including the line) in the Figure above are classified as $\hat{Y} = 1$, and points below as $\hat{Y} = 0$.

(b) (15 points total). The PDF of a Gaussian random variable with mean μ and variance σ^2 is

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\,\sigma^2}\Big).$$

(i) **(5 points)** Explain briefly how *generative modeling* approaches for classification use Bayes' rule to formulate the conditional probability $\Pr[Y|X]$, and discuss how it differ from *discriminative* approaches.
Generative models (e.g., LDA) estimate $\Pr(Y \mid X)$ by modeling the class-conditional density of X, $p(X \mid Y)$, and the marginal probability for each class, $p(Y)$, and then using Bayes' rule:

$$\Pr(Y = k \mid X) = \frac{\Pr(X \mid Y = k)\Pr(Y = k)}{\sum_{k'} \Pr(X \mid Y = k')\Pr(Y = k')}.$$

By contrast, discriminative methods (e.g., logistic regression) estimate $\Pr(Y \mid X)$ directly using a postulated functional form, without modeling the class-conditional density of X.

(ii) **(5 points)** Suppose that we have a dataset

$$\text{Class 0: } x = \{-3,\ -2,\ 0,\ 1\}, \qquad \text{Class 1: } x = \{4,\ 5,\ 6\}.$$

What class would LDA predict for $x = 2$?

- Estimated means: $\hat{\mu}_0 = -1,\ \hat{\mu}_1 = 5$.
- Pooled variance is common. Priors: $\pi_0 = 4/7,\ \pi_1 = 3/7$.
- $x = 2$ is exactly midway between the means; with a larger prior for class 0 the LDA discriminant favors $\boxed{\text{Class 0}}$.

(iii) **(5 points)** If you additionally collect two points, $x = \{3.5, 6.5\}$ from Class 1, does that change your class prediction with LDA at $x = 2$? Justify briefly.

New $\pi_1 = 5/9 > \pi_0$ and $\hat{\mu}_1$ stays at 5. The prior term now tilts the discriminant in favor of class 1, so $x = 2$ is assigned to $\boxed{\text{Class 1}}$. Thus the prediction changes because the class proportions (and hence the prior odds) have changed.

(c) (15 points total + 3 bonus points).

(i) **(5 points)** From 100 data points, you obtained the following confusion matrix (at $p^* = 0.5$):

	Pred = 1	Pred = 0
$Y = 1$	52	8
$Y = 0$	6	34

Compute the true positive rate (TPR) and false positive rate (FPR).

$$\text{TPR} = \frac{52}{52+8} \approx 0.867 \quad \text{and} \quad \text{FPR} = \frac{6}{6+34} = 0.15.$$

(ii) **(5 points)** If p^* increases from 0.5 to 0.9, do you expect more, fewer, or the same number of false positives/negatives? Provide a reason briefly.

Raising p^* to 0.9 produces fewer predicted positives, so false positives decrease while false negatives increase.

(iii) **(5 points)** In some scenarios (e.g. medical tests, fraud detection), p^* might be set to a value other than 0.5. Give an example scenario in one sentence where $p^* > 0.5$ might be preferable, then explain how that choice affects decisions and why it could lead to a more desirable outcome.

Example scenario with $p^* > 0.5$: Approving large bank loans cautiously.
For instance, in fraud detection or large bank loans, one might choose $p^* > 0.5$ to require stronger evidence before predicting "positive." This reduces false positives (e.g. granting risky loans) at the cost of missing some actual positives but may be more desirable given the higher cost of false positives.

(iv*) (3 bonus points*) How do we pick p^* to maximize TPR while keeping FPR $\leq 10\%$? Draw an example ROC curve below, and illustrate your decision on it, marking your chosen operating point with each axis and coordinate clearly labeled.

On the ROC curve, vary the threshold from 0 to 1 and plot (FPR, TPR). To constrain FPR ≤ 0.1, pick the point on the curve with FPR ≤ 0.1 and the highest possible TPR, which typically is located at where the ROC curve intersects the vertical line FPR $= 0.1$. That operating point corresponds to the desired threshold p^*.

Problem 6 (20 points total). Inference & Hypothesis Testing

(a) (10 points) You have an estimate $\hat{\theta} = 10$ of a parameter θ, which lacks a standard error formula available. You bootstrap and acquire 5 estimates:

$$6,\ 8,\ 8,\ 11,\ 12.$$

Using a *normal approximation*, construct a 95% confidence interval for θ, with $\hat{\theta}$ and the standard deviation estimated from the 5 bootstrap estimates. (*Hint*: $z_{0.975} \approx 1.96$.) Then state how to interpret the "95%" in this confidence interval—i.e., which probability is intended to be about 95%.

- *Bootstrap standard error*: Estimates $s = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2}$ with $\hat{\theta}^* = (6, 8, 8, 11, 12)$ gives $s \approx 2.449$.
- *95% confidence interval (normal approx.)*:

$$\hat{\theta} \pm z_{0.975}\, s = 10 \pm 1.96 * 2.449 = (5.20,\ 14.80).$$

- *Interpretation:* Repeating this sampling and bootstrap procedure many times would produce intervals containing the true θ in about 95 % of those repetitions.

(b) (10 points total). For a single null hypothesis H_0, the table on the left below shows the probabilities of each outcome ($p_1+p_2+p_3+p_4 = 1$). Now suppose we have m (e.g. 100) hypotheses tested simultaneously; let N_1, N_2, N_3, N_4 count each outcome, so $N_1 + N_2 + N_3 + N_4 = m$.

Single	H_0 is true	H_0 is not true
Reject H_0	p_1	p_2
Not reject H_0	p_3	p_4

Multiple	H_0 is true	H_0 is not true
Reject H_0	N_1	N_2
Not reject H_0	N_3	N_4

(i) **(5 points)** If we reject H_0 at level α (e.g. 0.05), express this requirement as an inequality involving p_1, p_2, p_3, p_4, and α.
Controlling the level at α requires

$$\Pr(\text{reject } H_0 \mid H_0 \text{ true}) = \frac{p_1}{p_1 + p_3} \leq \alpha.$$

(ii) **(5 points)** Define the *family-wise error rate (FWER)* and the *false discovery rate (FDR)* using these probabilities/counts. Explain their meanings in one sentence each.

- *Family–wise error rate (FWER):*

$$\text{FWER} = \Pr(N_1 \geq 1)$$

→ Probability of at least one false rejection among all m hypothesis tests.

- *False discovery rate (FDR):*

$$\text{FDR} = \mathrm{E}\left[\frac{N_1}{N_1 + N_2}\right]$$

→ Expected proportion of false rejections among all rejections.

Problem 7 (20 points total). Model Selection & PCA

(a) (7 points) Compare and contrast *principal component analysis (PCA)* against *best subset selection*. In your answer, (i) state the main goal of PCA vs. best subset selection, and (ii) outline how each procedure operates.

PCA	**Best subset selection**
Unsupervised learning	Supervised learning
Dimension reduction by finding orthogonal directions of maximal variance in X.	*Choose the predictor set* that minimizes prediction error for Y.
Finds the direction $\mathbf{u}$ along which the projection of X has maximal variance (with orthogonality constraints); keeps the first few components.	Fits every subset (or an efficient approximation) and picks the one with best fit (R^2_{adj}, cross-validation etc.).

(b) (6 points) Suppose you have a two-dimensional dataset of five points:

$$\mathcal{X} = \big\{(-4,\,2),\ (-1,\,3),\ (0,\,4),\ (2,\,6),\ (3,\,5)\big\}.$$

Compute the *directional variance* of this dataset along the vector $\mathbf{e}_1 = (1,\,0)$ and $\mathbf{e}_2 = (0,\,1)$. Then, decide whether $\mathbf{u}_1$ (the first principal component) is "closer" to $\mathbf{e}_1$ or $\mathbf{e}_2$, and briefly justify.

$$\mathcal{X} = \{(-4,2),\ (-1,3),\ (0,4),\ (2,6),\ (3,5)\}, \qquad \Longrightarrow \qquad \bar{x} = (0,4).$$

$$\text{Var}_{\mathbf{e}_1} = \frac{1}{5}\sum (x_i - \bar{x})^2 = 6, \qquad \text{Var}_{\mathbf{e}_2} = \frac{1}{5}\sum (y_i - \bar{y})^2 = 2.$$

The larger variance lies along $\mathbf{e}_1$; hence the first PC direction $\mathbf{u}_1$ is closer to $\mathbf{e}_1$ (the horizontal axis).

(c) (7 points). Performing PCA on a dataset yields 7 principal components with variances:

$$28,\ 16,\ 9,\ 3,\ 2,\ 1,\ 1.$$

Sketch a scree plot and compute the cumulative proportion of variance explained (PVE) by the top 3 components. Then briefly discuss the trade-offs of keeping too few or too many components.

$$\text{Variances: } 28, 16, 9, 3, 2, 1, 1, \qquad \Longrightarrow \qquad \text{Total} = 60.$$

- *Scree plot:* Plot component index (1–7) on the x-axis versus variance on the y-axis; expect a sharp elbow after the 3rd component.
- *Cumulative PVE (top 3):* $\dfrac{28 + 16 + 9}{60} = 0.883$ (88.3%).
- *Trade-off:* Too few PCs may omit meaningful signal (high bias); too many complicate models and possibly re-introduce noise (higher variance, interpretation difficulty).

Problem 8 (20 points total). Clustering

(a) (14 points total). You have a two-dimensional dataset:

$$z_1 = (0,\ 2), \quad z_2 = (2,\ 1), \quad z_3 = (3,\ -2), \quad z_4 = (6,\ 0).$$

(i) **(7 points)** Briefly explain the k-means algorithm and illustrate *two iterations* of k-means with $k = 2$ on $\{z_1, z_2, z_3, z_4\}$, assuming the initial cluster assignment:

$$C_1 = \{z_1,\ z_3\}, \quad C_2 = \{z_2,\ z_4\}.$$

Initial assignment $C_1 = \{z_1, z_3\},\ C_2 = \{z_2, z_4\}$.

Iteration 1:

Compute centroids $\mu_1^{(0)} = (1.5, 0),\ \mu_2^{(0)} = (4, 0.5)$.

Re-label points z_1, z_2, z_3 are closer to cluster 1, and z_4 is closer to cluster 2.

Iteration 2:

Compute centroids $\mu_1^{(1)} = (1.667, 0.333),\ \mu_2^{(1)} = (6, 0)$.

Re-label points Assignments do not change, so the algorithm has converged:
$\boxed{C_1 = \{z_1, z_2, z_3\}}$, $\boxed{C_2 = \{z_4\}}$.

(ii) **(7 points)** Using *complete linkage*, construct a dendrogram for $\{z_1, z_2, z_3, z_4\}$. Then explain how you would form two clusters from this dendrogram and identify those clusters.

- Pairwise distances (Euclidean):

Pair	Distance	Pair	Distance
(1,2)	$\sqrt{5}$	(2,3)	$\sqrt{10}$
(1,3)	5	(2,4)	$\sqrt{17}$
(1,4)	$\sqrt{40}$	(3,4)	$\sqrt{13}$

- Merge $\{z_1, z_2\}$ first (smallest max-distance $\sqrt{5}$) at height $\sqrt{5} \approx 2.236$. Update pairwise distance using complete linkage:

Pair	Distance
({1,2},3)	$\max\{5, \sqrt{10}\} = 5$
({1,2},4)	$\max\{\sqrt{40}, \sqrt{17}\} = \sqrt{40}$
(3,4)	$\sqrt{13}$

- Next merge $\{z_3, z_4\}$ at height $\sqrt{13} \approx 3.606$.
- Final merge of the two clusters $\{z_1, x_2\}$ and $\{z_3, z_4\}$ at height $\max\{5, \sqrt{40}\} = \sqrt{40} \approx 6.325$.

Cutting the dendrogram below 6.325 but above 3.606 yields the two clusters: $\boxed{\{z_1, z_2\}}$ and $\boxed{\{z_3, z_4\}}$.

(b) (6 points). State one advantage and one drawback of k-means compared to hierarchical clustering.

- *Advantage of k-means:* It is computationally fast and handles large datasets efficiently.
- *Drawback of k-means:* One must choose k in advance, and the final result can depend on the initial assignment of clusters.