TMA4268 V2026 Mock Exam 2

Solution Proposal

Compiled for Anders Bekkevard

Companion to `mock-exam-2.tex` (same directory).

Mock for: May 18, 2026

This document is a worked-solution proposal in the style of the official Stefanie/Sara solutions to the 2024 and 2025 finals. Point values are quoted in the heading of each solved sub-part. Partial-credit hints are inline. Refer back to `mock-exam-2.tex` for the problem statements.

Problem 1 (10 %) — Fill-in-the-blank concepts

Solution *(1 P per blank.)*

(1) **validation** (2nd option)

(2) **cross-validation** (3rd option)

(3) **bootstrap** (1st option)

(4) **forward stepwise** (4th option)

(5) **lasso** (2nd option)

(6) **sparse** (1st option)

(7) **principal component analysis** (4th option)

(8) **local** (3rd option)

(9) **dendrogram** (2nd option)

(10) **boosting** (1st option)

Grading: 1 P per correct blank. "Holdout" is acceptable for (1) only if accompanied by "validation"; the passage contrasts validation against the test set. "Bagging" for (3) is wrong: bagging is an ensemble *technique built on top of the bootstrap, but the resampling tool itself is the bootstrap.*

Problem 2 (28 %) — Multiple choice, T/F, short numeric

a) The bootstrap *(3 %)*

Solution *(0.75 P per statement.)*

(i) **True**. The defining feature of the bootstrap is sampling *with* replacement, same size n.

(ii) **False**. A single resample gives one realisation of $\hat{\theta}^*$; the SE estimate is the standard deviation of $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$, which requires B resamples (typically $B \geq 200$ for SEs, ≥ 1000 for percentile CIs).

(iii) **False**. The probability $1/e \approx 0.368$ is the probability that observation i is *absent* from the bootstrap sample: $P(\text{obs } i \notin \text{sample}) = (1 - 1/n)^n \to 1/e$. So the probability that observation i *does* appear approaches $1 - 1/e \approx 0.632$, not 0.368.

(iv) **True**. The *percentile interval* is one of the two standard bootstrap-CI constructions; the other is the normal-approximation interval $\hat{\theta} \pm z_{\alpha/2} \cdot \widehat{\text{SE}}_{\text{boot}}$. The percentile method makes no Gaussianity assumption and is directly read off the bootstrap distribution's quantiles, hence its appeal for statistics with skewed or otherwise non-normal sampling distributions.

b) Cross-validation and its pitfalls *(3 %)*

Solution *(0.75 P per statement.)*

(i) **True**. LOOCV averages n highly-correlated training fits (they share $n-1$ observations), so its variance as a test-error estimator is high. k-fold with $k = 5$ or 10 has more independent folds and lower variance, paid for by a small upward bias (each fit uses only $(k-1)n/k$ observations).

(ii) **False**. Pre-selecting predictors on the *full* training set lets information from each fold's held-out observations leak into the predictor screen. The CV error is then *optimistically biased*. The fix is to put the predictor screening *inside* each CV fold.

(iii) **True**. The validity of k-fold CV rests on the held-out fold being exchangeable with the training fold. Time-series data violate this (autocorrelation, drift), and naive k-fold mixes future and past observations. Time-series CV (rolling/expanding window) is the remedy.

(iv) **False**. The roles are swapped: the *inner* folds are used for hyperparameter tuning, and the *outer* folds are used for unbiased performance assessment of the whole tuning procedure.

c) Subset selection *(3 %)*

Solution *(1 %)* **(i)** Best-subset on $p = 8$ predictors fits every subset, including the null:

$$2^p = 2^8 = \boxed{256} \text{ candidate models.}$$

Solution *(1 %)* **(ii)** Forward stepwise fits the null model plus, at each step $k = 0, 1, \ldots, p-1$, the $p - k$ candidates that add one of the remaining predictors. Total:

$$1 + \sum_{k=0}^{p-1} (p-k) = 1 + p + (p-1) + \cdots + 1 = 1 + \frac{p(p+1)}{2} = 1 + \frac{8 \cdot 9}{2} = \boxed{37} \text{ candidate models.}$$

Solution *(1 %)* **(iii) False.** Backward stepwise starts from the *full* model with all p predictors and removes one at a time. When $p > n$ the full OLS fit is not even defined (the design matrix has more columns than rows), so backward stepwise cannot be applied in the high-dimensional regime. Forward stepwise, by contrast, can run until the model size hits $\min(n-1, p)$.

d) Cubic and natural splines: counting degrees of freedom *(3 %)*

Solution *(1 %)* **(i)** A cubic regression spline with K interior knots, fit on top of a global intercept, uses $K+3$ basis functions for the spline term (the truncated-power basis: x, x^2, x^3 plus one truncated cube per knot). With $K=5$:

$$K+3 = 5+3 = \boxed{8} \text{ coefficients (excluding the intercept).}$$

Solution *(1 %)* **(ii)** A *natural* cubic spline imposes linearity beyond the boundary knots, which removes two effective parameters from each end (four constraints, of which two are absorbed by the already-required smoothness). The standard count is $K+1$ basis functions (excluding the intercept):

$$K+1 = 5+1 = \boxed{6} \text{ coefficients (excluding the intercept).}$$

Solution *(1 %)* **(iii) True.** A cubic spline is, by construction, C^2 at every knot: continuous in value, first derivative, and second derivative. Only the third derivative is permitted to jump, which is what allows the piecewise cubic to bend differently in adjacent intervals.

e) Logistic regression: odds, log-odds, probability *(3 %)*

Solution *(1 %)* **(i)** A change of $\Delta x = 10$ years in `age` multiplies the odds by

$$e^{\hat{\beta}_{\text{age}} \cdot 10} = e^{0.04 \cdot 10} = e^{0.4} \approx \boxed{1.49}.$$

So the older individual's odds are about $1.49\times$ those of the younger.

Solution *(1 %)* **(ii)** Using the logistic transform,

$$\hat{p} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} = \frac{e^{-0.5}}{1+e^{-0.5}} = \frac{0.6065}{1.6065} \approx \boxed{0.38}.$$

Solution *(1 %)* **(iii) False.** β_j is a change in *log-odds*, not probability. The change in probability for a unit increase in x_j depends nonlinearly on the current value of $\hat{p}$ (largest near $\hat{p}=0.5$, smallest near 0 or 1). The clean linear interpretation is on the log-odds / odds-multiplication scale, not on the probability scale.

f) Confusion matrix interpretation *(4 %)*

We have $n = 1500$ patients, with 300 true positives in total (TP = 240, FN = 60) and 1200 true negatives in total (TN = 1080, FP = 120).

Solution *(1 %)* **(i)** Accuracy $= (\text{TP}+\text{TN})/n = (240+1080)/1500 = 1320/1500 = \boxed{0.88}$.

Solution *(1 %)* **(ii)** Sensitivity $= \text{TP}/(\text{TP}+\text{FN}) = 240/300 = \boxed{0.80}$.

Solution *(1 %)* **(iii)** Specificity $= \text{TN}/(\text{TN}+\text{FP}) = 1080/1200 = \boxed{0.90}$.

Solution *(1 %)* **(iv)** Precision $= \text{TP}/(\text{TP}+\text{FP}) = 240/(240+120) = 240/360 = \boxed{0.67}$.

g) K-means clustering *(3 %)*

Solution *(0.75 P per statement.)*

(i) **False**. K-means descends a non-convex objective; for a given initialization it converges to a *local* optimum. Different random starts can land in different basins; the standard remedy is to restart K-means many times and keep the best.

(ii) **True**. Each step (reassign-to-nearest-centroid, recompute-centroid-as-mean) is guaranteed to weakly decrease the within-cluster sum of squares, which is what gives the algorithm its convergence guarantee (to a local minimum).

(iii) **True**. Without standardization, large-scale variables dominate the squared Euclidean distance and the clusters are essentially partitioned along those directions. Standardizing (z-scores) is the usual fix.

(iv) **False**. K-means is not nested across K. Increasing K from 4 to 5 can reorganise the partition globally; the only structure that is monotone in K is hierarchical clustering (with single/complete linkage).

h) Tree ensembles: bagging, random forests, boosting *(4 %)*

Solution *(1 P per statement.)*

(i) **True**. Smaller m forces each split to consider a more restricted predictor set, decorrelating the trees. Per the bagging variance formula $\mathrm{Var}(\bar{f}) = \rho\sigma^2 + (1-\rho)\sigma^2/B$, lowering ρ lowers the floor. (See P3a.)

(ii) **False**. Bagging and random forests do *not* overfit in B: as $B \to \infty$ the variance of the averaged predictor monotonically decreases (toward $\rho\sigma^2$). B is chosen large enough that the OOB or test error stabilizes, not by CV. (Contrast with boosting, where B *is* a real tuning parameter.)

(iii) **True**. For squared loss $L(y, F) = \frac{1}{2}(y - F)^2$, the negative functional gradient with respect to F is $-\partial L/\partial F = y - F$, i.e. the current residual. So gradient boosting with squared loss is residual-fitting.

(iv) **True**. The textbook defaults are $m = \lfloor\sqrt{p}\rfloor$ for classification and $m = \lfloor p/3\rfloor$ for regression.

i) Direction of effect *(2 %)*

Solution *(1 P per statement.)*

(i) **True**. The smoothing-spline effective degrees of freedom $\mathrm{tr}(S_\lambda)$ shrink from n (interpolation, $\lambda = 0$) toward 2 (OLS line, $\lambda \to \infty$). Increasing λ gives a smoother, less-wiggly fit.

(ii) **False**. The direction is the opposite. *Smaller* ν means each tree contributes *less* to the running ensemble, so *more* trees B are needed to fit the data well. The textbook rule of thumb is that halving ν roughly doubles the required B.

Problem 3 (16 %) — Theory and hand calculations

a) The mathy one: variance of an average of correlated predictors *(8 %)*

Solution *(4 %)* **(i)** Let $X_b = \hat{f}^{*b}(x_0)$ for $b = 1, \ldots, B$ to shorten notation. By hypothesis, $\mathrm{Var}(X_b) = \sigma^2$ and $\mathrm{Cov}(X_b, X_{b'}) = \rho\sigma^2$ for $b \neq b'$. Using linearity of variance and the bilinear

expansion of $\mathrm{Var}(\sum_b X_b)$:

$$\begin{aligned}
\mathrm{Var}\big(\bar{f}_{\mathrm{bag}}(x_0)\big) &= \mathrm{Var}\left(\frac{1}{B}\sum_{b=1}^{B} X_b\right) = \frac{1}{B^2}\,\mathrm{Var}\left(\sum_{b=1}^{B} X_b\right) \\
&= \frac{1}{B^2}\left[\sum_{b=1}^{B}\mathrm{Var}(X_b) + \sum_{b\neq b'}\mathrm{Cov}(X_b, X_{b'})\right] \\
&= \frac{1}{B^2}\left[B\,\sigma^2 + B(B-1)\,\rho\,\sigma^2\right] \qquad \text{(there are } B(B-1) \text{ ordered off-diagonal pairs)} \\
&= \frac{\sigma^2}{B} + \frac{B-1}{B}\,\rho\,\sigma^2 \\
&= \frac{1}{B}\sigma^2 + \left(1 - \frac{1}{B}\right)\rho\sigma^2 \\
&= \rho\sigma^2 + \frac{1}{B}\sigma^2 - \frac{1}{B}\rho\sigma^2 \\
&= \boxed{\rho\,\sigma^2 \;+\; \frac{1-\rho}{B}\,\sigma^2.}
\end{aligned}$$

Grading: 1 P for splitting $\mathrm{Var}(\sum X_b)$ *into diagonal + off-diagonal terms; 1 P for the correct count* B *diagonal and* $B(B-1)$ *off-diagonal; 1 P for simplifying to* $\sigma^2/B + (B-1)\rho\sigma^2/B$*; 1 P for the final form* $\rho\sigma^2 + (1-\rho)\sigma^2/B$*. Deduct 1 per algebra slip.*

Solution *(1 %)* **(ii)** Letting $B \to \infty$ with ρ, σ^2 fixed:

$$\lim_{B\to\infty} \mathrm{Var}\big(\bar{f}_{\mathrm{bag}}(x_0)\big) = \boxed{\rho\,\sigma^2}.$$

The $(1-\rho)\sigma^2/B$ term vanishes, leaving the floor $\rho\sigma^2$ that cannot be removed by adding more bootstrap-trained predictors.

Solution *(2 %)* **(iii)** For bagged trees built on B bootstrap samples of the *same* training data, the trees tend to be highly correlated: a strong predictor dominates the top splits in essentially every tree, so the B trees look alike. Then ρ is close to 1 and the limit $\rho\sigma^2$ in (ii) is close to σ^2 — averaging more trees barely helps.

The *random forest* algorithm attacks this floor by, at each split, restricting the candidate predictors to a random subset of size $m < p$. This deliberately stops the dominant predictor from being used at every top split and *decorrelates* the trees: ρ falls, and so does the asymptotic floor $\rho\sigma^2$. The price is a small per-tree bias increase, paid for by the variance reduction.

Solution *(1 %)* **(iv)** The extreme case is $\rho = 0$ (i.i.d. predictions). The formula then collapses to

$$\mathrm{Var}\big(\bar{f}_{\mathrm{bag}}\big) = 0 \cdot \sigma^2 + \frac{1-0}{B}\sigma^2 = \boxed{\sigma^2/B},$$

which is the familiar variance of the mean of B i.i.d. random variables.

b) Hierarchical clustering by hand: single linkage *(5 %)*

Solution *(3 %)* **(i)** Reading off the upper triangle of D:

$$d_{12} = 2,\ d_{13} = 6,\ d_{14} = 10,\ d_{15} = 9,\ d_{23} = 5,\ d_{24} = 9,\ d_{25} = 8,\ d_{34} = 4,\ d_{35} = 5,\ d_{45} = 3.$$

Merge 1 (height 2). Smallest dissimilarity is $d_{12} = 2$. Merge $\{1\} \cup \{2\} \to \{1, 2\}$ at height $\mathbf{h_1 = 2}$. Under *single* linkage the new row/column is the *minimum* of the merged rows:

$$\begin{aligned} d(\{1,2\}, 3) &= \min(d_{13}, d_{23}) = \min(6, 5) = 5, \\ d(\{1,2\}, 4) &= \min(d_{14}, d_{24}) = \min(10, 9) = 9, \\ d(\{1,2\}, 5) &= \min(d_{15}, d_{25}) = \min(9, 8) = 8. \end{aligned}$$

The recomputed 4×4 dissimilarity matrix is

$$D^{(1)} = \begin{array}{c} \\ \{1,2\} \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} \{1,2\} & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 5 & 9 & 8 \\ 5 & 0 & 4 & 5 \\ 9 & 4 & 0 & 3 \\ 8 & 5 & 3 & 0 \end{pmatrix} \end{array}.$$

Merge 2 (height 3). Smallest off-diagonal of $D^{(1)}$ is $d_{45} = 3$. Merge $\{4\} \cup \{5\} \to \{4, 5\}$ at height $\mathbf{h_2 = 3}$. With single linkage:

$$\begin{aligned} d(\{4,5\}, \{1,2\}) &= \min(d_{14}, d_{24}, d_{15}, d_{25}) = \min(10, 9, 9, 8) = 8, \\ d(\{4,5\}, 3) &= \min(d_{34}, d_{35}) = \min(4, 5) = 4. \end{aligned}$$

The recomputed 3×3 dissimilarity matrix is

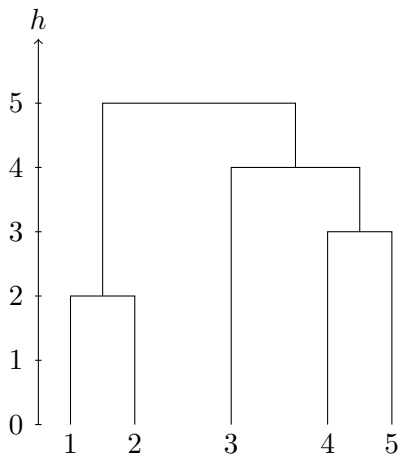
$$D^{(2)} = \begin{array}{c} \\ \{1,2\} \\ 3 \\ \{4,5\} \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2\} & 3 & \{4,5\} \end{array} \\ \begin{pmatrix} 0 & 5 & 8 \\ 5 & 0 & 4 \\ 8 & 4 & 0 \end{pmatrix} \end{array}.$$

Merge 3 (height 4). Smallest off-diagonal of $D^{(2)}$ is $d(\{4,5\}, 3) = 4$. Merge $\{3\} \cup \{4, 5\} \to \{3, 4, 5\}$ at height $\mathbf{h_3 = 4}$. Single-linkage update for the only remaining pair:

$$d(\{1,2\}, \{3,4,5\}) = \min(d_{13}, d_{14}, d_{15}, d_{23}, d_{24}, d_{25}) = \min(6, 10, 9, 5, 9, 8) = 5.$$

Merge 4 (height 5). The final merge of $\{1, 2\}$ with $\{3, 4, 5\}$ occurs at height $\mathbf{h_4 = 5}$.
Grading: 1 P for merge 1 with the correctly updated $D^{(1)}$; 1 P for merge 2 with correctly updated $D^{(2)}$; 1 P for both remaining heights $h_3 = 4$ and $h_4 = 5$. Deduct 0.5 per arithmetic slip on the linkage minimum.

Solution *(1 %)* **(ii)** Dendrogram with the four fusion heights $2, 3, 4, 5$ on the y-axis:



Solution *(1 %)* **(iii) No, the first merge would not change.** The first merge only compares pairs of singletons, and the smallest entry of D ($d_{12} = 2$) is the smallest single-pair dissimilarity regardless of linkage. Linkage rules only start to matter from the second merge onwards, when at least one cluster has size ≥ 2 and the rule decides how to summarise it.

c) The bootstrap, by hand *(3 %)*

Solution *(1 %)* **(i)** Each draw in the bootstrap sample is uniform on $\{1, \ldots, n\}$ with replacement, so the probability that a single draw is *not* observation i is $1-1/n$. The n draws are independent, so the probability that observation i is missed by *all* n draws is

$$P(i \notin \text{boot sample}) = \left(1 - \frac{1}{n}\right)^n.$$

Solution *(1 %)* **(ii)** Standard limit:

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx \boxed{0.368},$$

so $P(i \in \text{boot sample}) \to 1 - e^{-1} \approx \boxed{0.632}$ for large n.

Solution *(1 %)* **(iii)** A bootstrap estimate of $\mathrm{SE}(\hat{\theta})$ for the sample median:

1. Set the number of bootstrap resamples B (e.g. $B = 1000$).
2. For $b = 1, \ldots, B$: draw $(X_1^{*b}, \ldots, X_n^{*b})$ with replacement from the original sample, and compute $\hat{\theta}^{*b} = \text{median}(X_1^{*b}, \ldots, X_n^{*b})$.
3. Form the bootstrap mean $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*b}$.
4. Return

$$\widehat{\mathrm{SE}}_{\text{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \bar{\theta}^*\right)^2}.$$

Grading: full credit for any algorithm that draws with replacement, recomputes $\hat{\theta}$ on each resample, and returns the sample SD of the bootstrap replicates. Deduct 0.5 if sampling without replacement is used.

Problem 4 (22 %) — Data analysis: regression on house prices

a) Ordinary least squares with an interaction *(5 %)*

Solution *(1 %)* **(i)** Count the rows of the coefficient table: intercept, `crim`, `rm`, `age`, `dis`, `tax`, `ptratio`, `chas`, `rm:chas` = $\boxed{9}$ parameters. Consistency check: residual d.f. = $n_{\text{train}} - p = 400 - 9 = 391$, matching the printed residual d.f. exactly. ✓

Solution *(2 %)* **(ii)** The classmate misreads the table. In a model with an interaction term `rm:chas`, the coefficient on `chas` alone is the river effect *at* `rm` $= 0$ (standardized, i.e. at the mean number of rooms); the full effect of being next to the river is $\hat{\beta}_{\text{chas}} + \hat{\beta}_{\text{rm:chas}} \cdot \texttt{rm}$, which depends on `rm`. A non-significant main effect on `chas` does not imply that `chas` is irrelevant globally; the interaction is significant ($p = 0.027$) and pulls the river effect upward for high-`rm` tracts.

Implied river effects (in 1000 USD):

- (a) at standardized `rm` $= 0$ (the mean): $\hat{\beta}_{\text{chas}} + \hat{\beta}_{\text{rm:chas}} \cdot 0 = 1.20 + 0 = \boxed{1.20}$.
- (b) at standardized `rm` $= +1$ (one SD above mean): $\hat{\beta}_{\text{chas}} + \hat{\beta}_{\text{rm:chas}} \cdot 1 = 1.20 + 2.00 = \boxed{3.20}$.

Solution *(1 %)* **(iii)** Because the predictors are standardized, the coefficient $\hat{\beta}_{\text{ptratio}} = -1.80$ is the model's predicted change in `price` (in 1000 USD) for a one-standard-deviation increase in the pupil–teacher ratio, holding the other predictors fixed. A 95% Wald confidence interval is

$$\hat{\beta}_{\text{ptratio}} \pm 1.96 \cdot \widehat{\text{SE}}\big(\hat{\beta}_{\text{ptratio}}\big) \;=\; -1.80 \pm 1.96 \cdot 0.30 \;=\; -1.80 \pm 0.588 \;=\; \boxed{[\,-2.39,\ -1.21\,]} \text{ (in 1000 USD).}$$

So a one-SD rise in the pupil–teacher ratio is associated with a price drop of about 1,800 USD on average; the interval excludes zero, consistent with the highly significant t-value of -6.00 in the table.
Grading: 1 P for the correct CI (accept any minor rounding, e.g. $[-2.39, -1.21]$ *or* $[-2.388, -1.212]$*). Deduct* 0.5 *if the standardization rationale is omitted.*

Solution *(1 %)* **(iv)** A single marginal p-value reflects `age`'s contribution *conditional on the other predictors.* `age` may be correlated with `dis`, `tax`, or other predictors (e.g. old neighbourhoods are clustered, and proportion of pre-1940 housing is collinear with distance to employment centres), and its signal can be soaked up. Dropping it on the basis of one t-test alone is inappropriate; one should check whether the cross-validated test error rises after removal, or compare predictive performance with and without the term, before dropping it.

b) Forward stepwise vs. best subset *(3 %)*

Solution *(1 %)* **(i)** Forward stepwise fits the null model plus, at each step $k = 0, 1, 2, 3$, the $7 - k$ candidates that add one predictor on top of the current best size-k model. Through step 4 (i.e. choosing the best 4-predictor model) the total is

$$1 + 7 + 6 + 5 + 4 = \boxed{23} \text{ distinct candidate models.}$$

Solution *(1 %)* **(ii)** Best-subset fits every subset up to and including size 4:

$$\binom{7}{0} + \binom{7}{1} + \binom{7}{2} + \binom{7}{3} + \binom{7}{4} = 1 + 7 + 21 + 35 + 35 = \boxed{99} \text{ candidate models.}$$

Solution *(1 %)* **(iii)** Forward stepwise is *greedy*: each step locks in a predictor and never revisits the choice. The best 4-predictor model under best subset is the global RSS-minimiser over all $\binom{7}{4} = 35$ size-4 subsets, while the forward path is constrained to extend the chosen 3-predictor subset. When the optimal 4-subset cannot be reached by sequentially adding to the optimal 3-subset — a classic outcome with correlated predictors — the two methods disagree.

c) Ridge with 10-fold cross-validation *(7 %)*

Solution *(1 %)* **(i)** Ridge minimises (with the intercept *not* penalised, since standardized predictors have mean zero so $\hat{\beta}_0 = \bar{y}$ regardless):

$$\hat{\beta}^R_\lambda \;=\; \arg\min_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} \big(y_i - \beta_0 - x_i^\top \beta\big)^2 \;+\; \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \qquad \lambda \geq 0.$$

The penalty sum runs over $j = 1, \ldots, p$ (here $p = 7$), *excluding* the intercept. Closed-form: $\hat{\beta}^R_\lambda = (X^\top X + \lambda I_p)^{-1} X^\top y$ on centred data.

Solution *(2 %)* **(ii)** The ridge penalty $\lambda \sum_j \beta_j^2$ is smooth and strictly convex in β. Its gradient is zero only at $\beta = 0$ jointly; for finite λ the minimiser shrinks each $\hat{\beta}_j$ *toward* zero but generically does not set any of them *exactly* to zero (it would require the unpenalised gradient to perfectly cancel the linear ridge derivative at exactly $\beta_j = 0$, a measure-zero event). Hence all seven coefficients remain nonzero at both $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{1\text{SE}}$.

The *lasso* penalty $\lambda \sum_j |\beta_j|$ is non-differentiable at zero, and its subdifferential contains an interval $[-\lambda, \lambda]$ at $\beta_j = 0$. Whenever the gradient of the data term $|2x_j^\top (y - X\beta)|$ is smaller than λ, the optimum is at $\beta_j = 0$ exactly. So lasso *does* produce sparse coefficient vectors and would zero out one or more of the seven predictors at the lasso CV-optimal λ.

Solution *(1 %)* **(iii)** The ridge penalty $\lambda \sum_j \beta_j^2$ treats the coefficients as directly comparable. Without standardization, a predictor measured on a large numeric scale has a tiny OLS coefficient and is penalised lightly, while one on a small scale has a large OLS coefficient and is penalised heavily. Standardising every predictor to mean 0, SD 1 makes the penalty scale-invariant in the units of the data and gives every predictor a fair chance of being shrunk.

Solution *(2 %)* **(iv)** Test MSE drops from 19.40 (OLS) to 16.20 (ridge at $\hat{\lambda}_{\min}$), a 16.5% reduction. Since ridge *adds bias* relative to OLS but reduces variance, the only way ridge can beat OLS in test MSE is if OLS is *variance-dominated* on this dataset. With $n_{\text{train}} = 400$ and $p = 7$ standardized predictors, OLS is mostly fine but is sufficiently variance-heavy — likely due to correlation among the continuous predictors (`age`, `dis`, `tax` are correlated, for instance) — that shrinking coefficients toward zero buys more in variance reduction than it pays in bias inflation. Ridge is winning the bias–variance trade-off here.

Solution *(1 %)* **(v) Comparable or slightly lower than OLS.** If most predictors carry real signal (which the OLS t-table suggests — only `age` and `chas` have $p > 0.05$), the lasso will zero out at most one or two coefficients and shrink the rest, with a test MSE roughly between the OLS value of 19.40 and the ridge value of 16.20. A defensible alternative answer: "slightly higher than ridge but lower than OLS, since the data has weak sparsity."

d) PCR and spline degrees of freedom *(4 %)*

Solution *(2 %)* **(i)** Total variance (sum of eigenvalues of the correlation matrix):

$$\sum_{j=1}^{7} \lambda_j = 2.80 + 1.60 + 1.00 + 0.70 + 0.40 + 0.30 + 0.20 = 7.00.$$

(Sanity check: standardized predictors have variance 1, so the trace of the correlation matrix is exactly $p = 7$. ✓) The 80% threshold is $0.80 \cdot 7.00 = 5.60$. Cumulative sums:

$$\begin{aligned}
\text{PC1:} &\quad 2.80/7.00 = 40.00\% \\
\text{PC1+2:} &\quad 4.40/7.00 = 62.86\% \\
\text{PC1+3:} &\quad 5.40/7.00 = 77.14\% \quad (< 80\%) \\
\text{PC1+4:} &\quad 6.10/7.00 = 87.14\% \quad (\geq 80\%) \ \checkmark
\end{aligned}$$

Need $\boxed{4}$ principal components to explain at least 80% of the total variance.

Solution *(1 %)* **(ii)** Both PCR and ridge place *less weight on low-variance directions* of the predictor space: write the standardized design matrix as $X = UDV^\top$ (SVD). PCR keeps the top M principal components — it gives full weight to the first M directions and *zero* weight to the rest. That is the "discrete" aspect: directions are either fully in or fully out. Ridge instead applies the weight $d_j^2/(d_j^2 + \lambda)$ to the jth SVD direction, which is a smooth, monotone shrinkage from 1 (large d_j, large variance) down toward 0 (small d_j, small variance). Same shape of pressure, continuous (ridge) versus discrete (PCR).

Solution *(1 %)* **(iii)** Counting d.f. including the intercept:

- Intercept: 1 d.f.
- Cubic spline on `rm` with $K = 3$ interior knots: $K + 3 = 6$ d.f. (excluding intercept).
- Five remaining continuous predictors `crim`, `age`, `dis`, `tax`, `ptratio` entered linearly: $5 \cdot 1 = 5$ d.f.
- Binary `chas` (one dummy): 1 d.f.

Total:

$$1 + 6 + 5 + 1 \;=\; \boxed{13} \text{ d.f.}$$

e) Boosting interpretation *(3 %)*

Solution *(1 %)* **(i)** Boosting at test MSE 11.40 versus OLS at 19.40 and ridge at 16.20 suggests the relationship between predictors and `price` contains substantial *nonlinearities and/or interactions* that the linear models cannot capture. Depth-4 trees can model up to 4-way interactions and the smooth (often nonlinear) effect of `rm` on price.

Solution *(1 %)* **(ii)** Not a good idea. Unlike random forests, gradient boosting *can* overfit for very large B: each new tree fits residuals of the running ensemble and, with enough trees, those residuals are just noise. B is a genuine tuning parameter; "$B = 10{,}000$ just to be safe" will hurt test MSE. The right procedure is to monitor CV / OOB / validation error as B grows and stop when it stops decreasing.

Solution *(1 %)* **(iii)** Two diagnostics for a boosted ensemble:

- **Variable importance plot** (mean decrease in node impurity or permutation importance). *Answers*: which predictors the ensemble uses most heavily. *Does not answer*: the *direction* or shape of each predictor's effect, or which interactions are present.
- **Partial dependence plot** on a single predictor. *Answers*: the marginal shape of how the predicted `price` depends on that predictor, averaged over the others. *Does not answer*: how the effect changes within subgroups (i.e. interactions, unless you compute 2-D PDPs) or whether the marginal effect is causal.

Problem 5 (24 %) — Data analysis: classification of heart disease

a) Discriminant analysis: LDA vs. QDA *(7 %)*

Solution *(2 %)* **(i)** With shared $\hat{\sigma}^2 = 2$ and equal priors $\hat{\pi}_0 = \hat{\pi}_1 = 0.5$ (so $\log \hat{\pi}_0 = \log \hat{\pi}_1 = \log 0.5 = -\log 2$):

$$\begin{aligned}
\delta_0(x) &= x \cdot \frac{\hat{\mu}_0}{\hat{\sigma}^2} - \frac{\hat{\mu}_0^2}{2\hat{\sigma}^2} + \log \hat{\pi}_0 = x \cdot \frac{4}{2} - \frac{16}{4} + \log(0.5) \\
&= 2x - 4 + \log(0.5), \\
\delta_1(x) &= x \cdot \frac{\hat{\mu}_1}{\hat{\sigma}^2} - \frac{\hat{\mu}_1^2}{2\hat{\sigma}^2} + \log \hat{\pi}_1 = x \cdot \frac{6}{2} - \frac{36}{4} + \log(0.5) \\
&= 3x - 9 + \log(0.5).
\end{aligned}$$

Solution *(1 %)* **(ii)** Set $\delta_0(x) = \delta_1(x)$. The shared $\log(0.5)$ cancels:

$$\begin{aligned}
2x - 4 &= 3x - 9 \\
9 - 4 &= 3x - 2x \\
5 &= x.
\end{aligned}$$

Decision boundary at $\boxed{x = 5}$. Interpretation: with equal priors and equal variances, the LDA boundary sits at the *midpoint* $(\hat{\mu}_0 + \hat{\mu}_1)/2 = (4 + 6)/2 = 5$ of the two class means; below 5 we predict class 0 (no CHD), above 5 we predict class 1 (CHD).

Solution *(1 %)* **(iii)** With $\hat{\pi}_0 = 0.8$, $\hat{\pi}_1 = 0.2$, the priors no longer cancel. Solving $\delta_0(x) = \delta_1(x)$:

$$\begin{aligned} 2x - 4 + \log(0.8) &= 3x - 9 + \log(0.2) \\ 2x - 3x &= -9 + \log(0.2) - (-4 + \log(0.8)) \\ -x &= -5 + \log(0.2) - \log(0.8) \\ -x &= -5 + \log\left(\tfrac{0.2}{0.8}\right) = -5 + \log(0.25) = -5 - \log 4 \\ x &= 5 + \log 4 \approx 5 + 1.386 = \boxed{6.386}. \end{aligned}$$

The boundary moves *toward* the class-1 mean ($\hat{\mu}_1 = 6$). This makes sense: with a strong prior $\hat{\pi}_0 = 0.8$ on "no CHD," we need more evidence (a higher x) before classifying as CHD, so the threshold moves rightward.

Solution *(3 %)* **(iv)** With $\hat{\sigma}_0^2 = 1$, $\hat{\sigma}_1^2 = 4$, $\hat{\mu}_0 = 4$, $\hat{\mu}_1 = 6$, $\hat{\pi}_0 = \hat{\pi}_1 = 0.5$:

$$\begin{aligned} \delta_0(x) &= -\frac{x^2}{2\hat{\sigma}_0^2} + \frac{x\hat{\mu}_0}{\hat{\sigma}_0^2} - \frac{\hat{\mu}_0^2}{2\hat{\sigma}_0^2} - \tfrac{1}{2}\log\hat{\sigma}_0^2 + \log\hat{\pi}_0 \\ &= -\frac{x^2}{2 \cdot 1} + \frac{4x}{1} - \frac{16}{2 \cdot 1} - \tfrac{1}{2}\log 1 + \log(0.5) \\ &= -\tfrac{1}{2}x^2 + 4x - 8 + 0 + \log(0.5), \\ \delta_1(x) &= -\frac{x^2}{2\hat{\sigma}_1^2} + \frac{x\hat{\mu}_1}{\hat{\sigma}_1^2} - \frac{\hat{\mu}_1^2}{2\hat{\sigma}_1^2} - \tfrac{1}{2}\log\hat{\sigma}_1^2 + \log\hat{\pi}_1 \\ &= -\frac{x^2}{2 \cdot 4} + \frac{6x}{4} - \frac{36}{2 \cdot 4} - \tfrac{1}{2}\log 4 + \log(0.5) \\ &= -\tfrac{1}{8}x^2 + \tfrac{3}{2}x - \tfrac{9}{2} - \tfrac{1}{2}\log 4 + \log(0.5). \end{aligned}$$

The QDA boundary is quadratic because, when class variances differ, the $-x^2/(2\sigma_k^2)$ term has a different coefficient in δ_0 and δ_1. Subtracting them leaves an x^2 term that does *not* cancel:

$$\delta_0(x) - \delta_1(x) = \left(-\tfrac{1}{2} + \tfrac{1}{8}\right)x^2 + \cdots = -\tfrac{3}{8}x^2 + \cdots,$$

so the boundary $\delta_0 = \delta_1$ is a quadratic equation in x (generically two roots, hence a quadratic decision region). Under the LDA assumption $\sigma_0^2 = \sigma_1^2$, the x^2 coefficients are equal and cancel, leaving a linear boundary.
Grading: 1 P each correct discriminant; 1 P for the structural explanation (the x^2 coefficients differ when variances differ). Deduct 1 P per algebra mistake.

b) Logistic regression with an interaction *(8 %)*

Solution *(2 %)* **(i)** Encoding: `sex` $= 0$ for male (reference), `sex` $= 1$ for female.

For one extra year of `age`, holding everything else fixed, the log-odds change is $\hat{\beta}_{\text{age}} + \hat{\beta}_{\text{age:sex}} \cdot$ `sex`:

- (a) Male (`sex` $= 0$): change in log-odds $= \hat{\beta}_{\text{age}} = 0.06$, so odds factor $= e^{0.06} \approx \boxed{1.06}$.
- (b) Female (`sex` $= 1$): change in log-odds $= \hat{\beta}_{\text{age}} + \hat{\beta}_{\text{age:sex}} = 0.06 + (-0.04) = 0.02$, so odds factor $= e^{0.02} \approx \boxed{1.02}$.

Grading: 1 P male, 1 P female. Deduct 1 if both factors are $e^{0.06}$, which ignores the interaction.

Solution *(1 %)* **(ii)** For one extra mmol/L of `ldl`, the odds change by

$$e^{\hat{\beta}_{\text{ldl}}} = e^{0.40} \approx \boxed{1.49}.$$

This factor is the same for males and females because the model has *no* `ldl:sex` interaction: the slope on `ldl` on the log-odds scale is a single number $\hat{\beta}_1 = 0.40$ shared across both sex strata. (Only `age` interacts with `sex` in this model.)

Solution *(2 %)* **(iii)** The patient: male (`sex` $= 0$), `age` $= 55$, `ldl` $= 5$, `famhist` $= 1$. The linear predictor is

$$\begin{aligned}
\hat{\eta} &= \hat{\beta}_0 + \hat{\beta}_1\,\texttt{ldl} + \hat{\beta}_2\,\texttt{age} + \hat{\beta}_3\,\texttt{sex} + \hat{\beta}_4\,(\texttt{age:sex}) + \hat{\beta}_5\,\texttt{famhist} \\
&= -5.00 + 0.40 \cdot 5 + 0.06 \cdot 55 + 1.20 \cdot 0 + (-0.04) \cdot (55 \cdot 0) + 0.80 \cdot 1 \\
&= -5.00 + 2.00 + 3.30 + 0 + 0 + 0.80 \\
&= \boxed{1.10}.
\end{aligned}$$

Fitted probability:

$$\hat{p} = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = \frac{e^{1.10}}{1 + e^{1.10}} = \frac{3.0042}{4.0042} \approx \boxed{0.7503}.$$

At threshold 0.5, $\hat{p} > 0.5$, so the predicted class is $\boxed{\hat{y} = 1}$ (predicted CHD).
Grading: 1 P linear predictor; 1 P probability + class label. Accept $\hat{p} \in [0.74, 0.76]$ for rounding.

Solution *(1 %)* **(iv)** When a model contains an interaction `age:sex`, the main-effect coefficient $\hat{\beta}_{\text{sex}} = 1.20$ is the sex effect *at* `age` $= 0$, which is far outside the observed data range. Its p-value tests whether the sex effect is zero *at age zero*, which is essentially meaningless here. Dropping the main-effect `sex` term while keeping the interaction would force the sex effect to be zero at `age` $= 0$, distorting the fitted effects at realistic ages. The standard rule ("hierarchy / marginality principle") is: do not drop a main effect whose interaction is in the model.

Solution *(2 %)* **(v)** In the CHD setting:

- **Sensitivity** $= P(\hat{Y} = 1 \mid Y = 1)$: the proportion of patients who *actually develop CHD* that the classifier correctly flags as CHD. ("Of the patients who really get sick, how many do we catch?")
- **Specificity** $= P(\hat{Y} = 0 \mid Y = 0)$: the proportion of patients who *never develop CHD* that the classifier correctly clears. ("Of the healthy patients, how many do we leave alone?")

c) ROC curve and threshold tuning *(5 %)*

The test set has $n = 200$ patients: 80 actual CHD (TP $= 30$, FN $= 50$) and 120 actual non-CHD (TN $= 104$, FP $= 16$).

Solution *(2 %)* **(i)**

$$\begin{aligned}
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{30}{30 + 50} = \frac{30}{80} = \boxed{0.375}, \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{104}{104 + 16} = \frac{104}{120} \approx \boxed{0.867}, \\
\text{Test error rate} &= \frac{\text{FP} + \text{FN}}{n} = \frac{16 + 50}{200} = \frac{66}{200} = \boxed{0.33}.
\end{aligned}$$

Grading: 1 P for sensitivity, 1 P for specificity + error.

Solution *(1 %)* **(ii)** The **AUC** is the area under the ROC curve, swept over all classification thresholds in $[0, 1]$. Equivalently, AUC equals the probability that the classifier ranks a randomly chosen positive observation higher (by predicted probability) than a randomly chosen negative observation. $\boxed{\text{AUC} = 0.5}$ corresponds to a classifier that ranks no better than random guessing (the diagonal line in ROC space).

Solution *(2 %)* **(iii)** Lowering the threshold from 0.5 to 0.3 flags *more* patients as CHD-positive:

(a) Sensitivity *increases* (more true positives are caught — jumping from 0.375 to roughly 0.66 per the marked point on the ROC curve), and specificity *decreases* (more false alarms, jumping from ≈ 0.867 to ≈ 0.64, i.e. FPR rises from ~ 0.13 to ~ 0.36).

(b) On the ROC curve the operating point moves *up and to the right*, from the marked $\hat{p} = 0.5$ point to the marked $\hat{p} = 0.3$ point.

A clinical context where this is the right move: *screening* for CHD in a setting where missing a true case is much more costly (delayed treatment, future events) than a false alarm, which only triggers a follow-up test.

d) A tree-based competitor *(4 %)*

Solution *(2 %)* **(i)** I would fit a **random forest** for binary classification. Tuning parameters and their justifications:

- B (number of trees) = 500. Random forests do not overfit in B; we keep growing until the OOB / test error stabilises, and 500 is the conventional plateau point.
- m (predictors sampled per split) = $\lfloor \sqrt{p} \rfloor$. With $p = 6$ candidate predictors that gives $m = 2$ (or 3). The role of m is to *decorrelate* the trees (lower ρ in the bagging variance formula from P3a) and reduce variance.
- Minimum terminal node size = 1 (i.e. trees grown deep). Variance is controlled by averaging across the B trees rather than by per-tree pruning; deep trees give low-bias base learners.

Grading: 2 P for naming the method and giving a concrete value for each of three parameters with one-sentence justification. Deduct 0.5 per missing parameter or missing justification.

Solution *(1 %)* **(ii)** Forest: sensitivity 0.50, specificity 0.85, error rate 0.21. Logistic at threshold 0.5: sensitivity 0.375, specificity 0.867, error 0.33. The random forest *wins* on test error rate ($0.21 < 0.33$) and *wins* on sensitivity ($0.50 > 0.375$), at a small cost in specificity ($0.85 < 0.867$). In a clinical CHD-screening setting where catching true cases matters more than avoiding the occasional false alarm, the higher sensitivity and lower overall error rate make **random forest** the recommended classifier. The decisive criterion is the combination of test error and sensitivity in a context where false negatives are clinically costly.

Solution *(1 %)* **(iii)** With $\Pr(\texttt{chd} = 1) \approx 0.27$, the trivial "always predict no CHD" classifier already attains test error ≈ 0.27 — which actually *beats* the logistic regression's error of 0.33. Test error alone is therefore a misleading metric: a useless classifier can "win" it on imbalanced data. The right metrics here are the *(sensitivity, specificity)* pair (which separately measure performance on each class) and the threshold-independent **AUC**; in a cost-sensitive screening context, one should privilege sensitivity (or the F_1 score, which weighs precision and recall) over raw error rate.

End of solution proposal. Total awarded: $10 + 28 + 16 + 22 + 24 = 100$ points.