TMA4268 Statistical Learning V2026

Mock Exam 2 (estimated final exam)

Compiled by Claude for Anders Bekkevard

Based on the Apr 28 exam-review lecture, the 2023–2025 finals, and the prof's stated scope rule

Mock for: May 18, 2026 (real exam date)

Instructions.

- **Duration:** 4 hours. **Open book.** Permitted aids: ISLP (2nd ed.), one handwritten A5 sheet of notes, calculator.
- **No code required.** Write answers as math, plain English, or pseudocode. Do *not* memorize R/Python package names.
- **Show your work** — partial credit is generously available, including when calculator slips spoil a numeric answer but the setup is correct.
- **No negative scoring.** Always answer, even when unsure.
- **If a question seems broken or ambiguous**, state the assumption you are making in one short sentence and proceed.
- Total: **100 points = 100 %**. Per-problem weights given in parentheses.

Grade boundaries (NTNU *prosentvurderingsmetoden*, advisory): A: 89–100 % B: 77–88 % C: 65–76 % D: 53–64 % E: 41–52 % F: 0–40 %.

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

Read the passage and pick the best word or short phrase for each blank from the choices in parentheses. Each correct fill is worth 1 %.

In statistical learning we often partition the data into three disjoint pieces: a training set used to fit a candidate model, a _____ (1) (*bootstrap / validation / holdout-of-the-residuals / oob*) set used to choose between candidate models or tune a hyperparameter, and a test set used to estimate the chosen model's final generalization error. When data are too scarce for a clean three-way split we replace the validation set by _____ (2) (*regularization / permutation testing / cross-validation / label smoothing*); the most common choice in this course is k-fold with $k = 5$ or 10.

A complementary resampling tool, the _____ (3) (*bootstrap / bagging / Bayes rule / Markov chain*), draws repeated samples *with replacement* of the same size as the original training set, and uses the empirical sampling distribution of a statistic to estimate its standard error and confidence intervals.

For model selection within the regression module we distinguished discrete and continuous approaches. Best-subset selection considers all 2^p subsets of predictors, while _____ (4)

(*lasso shrinkage / principal component / boosted / forward stepwise*) selection performs a greedy search that adds one predictor at a time and is computationally tractable even when $p > n$. Among continuous approaches, ridge regression shrinks coefficients toward zero but never sets them exactly to zero, whereas _____(5) (*ridge / lasso / PCR / GAM*) can yield _____(6) (*sparse / dense / orthogonal / biased*) coefficient vectors because the geometry of its constraint set has corners on the coordinate axes.

For unsupervised analyses on the same data, dimensionality reduction can be performed by _____(7) (*ridge regression / the bootstrap / the F-test / principal component analysis*), which finds orthogonal directions of maximal variance. Clustering can be done by K-means (which always returns a flat partition into K clusters and can converge to a _____(8) (*global / Bayes-optimal / local / deterministic*) minimum depending on initialization) or by hierarchical clustering, which produces a tree-structured nested family of clusterings called a _____(9) (*biplot / dendrogram / scree plot / ROC curve*).

In ensemble methods for trees, _____(10) (*boosting / bagging / random forests / stacking*) builds many trees *sequentially*, each one trained on residuals (or, more generally, negative gradients) of the running ensemble, and combines them with a small shrinkage factor.

Problem 2 (28 %) — Multiple choice, true/false, and short numeric

For each subproblem, write *True* or *False* for each statement (or the requested numeric answer). You may add a one-sentence justification but only if you think it helps; do not write essays.

a) The bootstrap *(3 %)*

Mark each statement true or false.

(i) Bootstrap samples are drawn from the training data *with* replacement, and each bootstrap sample has the same size n as the original.

(ii) A single bootstrap resample is enough to obtain a reasonable estimate of the standard error of a statistic; the role of the number of resamples B is purely computational.

(iii) For large n, the probability that a given training observation *appears* in a particular bootstrap sample approaches $1/e \approx 0.368$.

(iv) A common bootstrap-based 95% confidence interval for $\hat{\theta}$ is the *percentile interval*: the 2.5% and 97.5% quantiles of the bootstrap distribution $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.

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

Mark each statement true or false.

(i) Compared with leave-one-out CV (LOOCV), 5- or 10-fold CV typically has lower *variance* as an estimator of the test error, at the cost of a small additional bias.

(ii) Suppose you first rank all p predictors by their marginal correlation with y on the *full* training data, keep the top 20, and then run 10-fold CV on a logistic regression fit using just those 20. The resulting CV error is an unbiased estimate of the test error of the procedure.

(iii) For ordinary k-fold CV to give a sensible estimate of the test error, the observations must be (at least approximately) independent and identically distributed. With strong time-series autocorrelation, naive k-fold CV is misleading.

(iv) In nested cross-validation, the *outer* folds are used for model selection (hyperparameter tuning) and the *inner* folds for performance assessment.

c) Subset selection *(3 %)*

(i) *(1 %)* How many distinct candidate models does *best-subset selection* fit on $p = 8$ predictors (counting the null model)? *One numeric answer.*

(ii) *(1 %)* How many distinct candidate models does *forward stepwise selection* fit on $p = 8$ predictors (counting the null model)? *One numeric answer.*

(iii) *(1 %)* True or false, with one short justification: "Backward stepwise selection can be applied in the high-dimensional regime $p > n$."

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

A predictor x is modeled by a piecewise polynomial fit on top of an intercept.

(i) *(1 %)* A *cubic regression spline* with $K = 5$ interior knots is fit using the truncated-power basis. How many coefficients does the spline term consume (*not* including the global intercept of the model)?

(ii) *(1 %)* A *natural* cubic spline with $K = 5$ interior knots is fit on the same predictor. How many coefficients does the spline term consume (not including the global intercept)?

(iii) *(1 %)* True or false: "A cubic spline at every knot is continuous in value, in first derivative, and in second derivative; only the third derivative is allowed to jump."

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

(i) *(1 %)* A logistic regression on the binary outcome "heart attack within ten years" includes a coefficient $\hat{\beta}_{\text{age}} = 0.04$ for `age` measured in *years*. By what factor do the odds of a heart attack change when comparing two otherwise identical individuals whose ages differ by 10 years? (One numeric value, rounded to two decimals.)

(ii) *(1 %)* For a particular individual the fitted linear predictor equals $\hat{\eta} = -0.5$. What is the predicted probability $\hat{p}$ of a heart attack within ten years? (One numeric value, rounded to two decimals.)

(iii) *(1 %)* True or false: "A logistic-regression coefficient β_j has the interpretation that a one-unit increase in x_j changes the *probability* of $Y = 1$ by approximately β_j, holding the other predictors fixed."

f) Confusion matrix interpretation *(4 %)*

A binary classifier is applied to a test set of 1500 patients. Of these, 300 truly have the disease. At a chosen threshold the confusion matrix is:

	Predicted: disease	**Predicted: no disease**
Actual: disease	240	60
Actual: no disease	120	1080

Compute, with two decimals:

(i) *(1 %)* Overall test *accuracy* of the classifier.

(ii) *(1 %) Sensitivity* (true positive rate).

(iii) *(1 %) Specificity* (true negative rate).

(iv) *(1 %) Precision* (positive predictive value) $= \text{TP}/(\text{TP} + \text{FP})$.

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

Mark each statement true or false.

(i) Given a fixed initialization, the K-means algorithm converges to the *global* minimum of the within-cluster sum of squares.

(ii) The within-cluster sum of squares (the K-means objective) is *non-increasing* from one iteration to the next.

(iii) For predictors measured in very different units (e.g. kilograms and meters), running K-means on the raw data without standardizing can lead to clusters dominated by the largest-scale variable.

(iv) Going from $K = 4$ to $K = 5$ in K-means produces a 5-cluster solution in which one of the original 4 clusters is split, with the other three unchanged.

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

Mark each statement true or false.

(i) In a random forest, the parameter m (number of predictors sampled per split) controls the correlation between trees: smaller m tends to give *less* correlated trees and therefore more variance reduction.

(ii) In bagging and random forests, increasing the number of trees B can eventually cause the ensemble to *overfit* the training data.

(iii) In gradient boosting with squared-error loss, fitting the next tree to the current residuals is equivalent to fitting a tree to the *negative gradient* of the loss with respect to the current ensemble's predictions.

(iv) Random forests for classification are typically fit with $m \approx \sqrt{p}$, whereas for regression the standard default is $m \approx p/3$.

i) Direction of effect *(2 %)*

Mark each statement true or false.

(i) In a smoothing-spline fit, increasing the smoothing parameter λ decreases the effective degrees of freedom $\mathrm{tr}(S_\lambda)$ and produces a smoother (less wiggly) fit.

(ii) In gradient boosting, decreasing the shrinkage / learning rate ν *decreases* the number of trees B required to fit the data well; smaller ν means each tree contributes more, so fewer trees are needed.

Problem 3 (16 %) — Theory and hand calculations

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

Let $\hat{f}^{*1}(x_0), \ldots, \hat{f}^{*B}(x_0)$ be the predictions of B models trained on B different bootstrap samples drawn from the same training set, evaluated at a fixed query point x_0. Assume that the $\hat{f}^{*b}(x_0)$ are identically distributed with common variance

$$\operatorname{Var}\Big(\hat{f}^{*b}(x_0)\Big) = \sigma^2 \quad \text{for every } b,$$

and pairwise positive correlation

$$\operatorname{Cov}\Big(\hat{f}^{*b}(x_0),\, \hat{f}^{*b'}(x_0)\Big) = \rho\,\sigma^2 \quad \text{for every } b \neq b'.$$

Let $\bar{f}_{\text{bag}}(x_0) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x_0)$ denote the bagged prediction.

(i) *(4 %)* Show, starting from the definition $\operatorname{Var}(Z) = \mathbb{E}\big[(Z - \mathbb{E}Z)^2\big]$ *or* the identity $\operatorname{Var}(\sum_b X_b) = \sum_b \operatorname{Var}(X_b) + \sum_{b \neq b'} \operatorname{Cov}(X_b, X_{b'})$, that

$$\boxed{\operatorname{Var}\big(\bar{f}_{\text{bag}}(x_0)\big) \;=\; \rho\,\sigma^2 \;+\; \frac{1-\rho}{B}\,\sigma^2.}$$

(ii) *(1 %)* What is the limit of $\operatorname{Var}\big(\bar{f}_{\text{bag}}(x_0)\big)$ as $B \to \infty$, holding ρ and σ^2 fixed?

(iii) *(2 %)* Briefly explain why, in light of (ii), *bagging alone* cannot reduce variance below a positive floor in the typical case where bootstrap-trained trees are highly correlated (for instance because every tree splits on the same dominant predictor at the root). What does the *random forest* algorithm do differently to attack this floor?

(iv) *(1 %)* In what (extreme) special case does the formula above reduce to the familiar $\operatorname{Var}(\bar{X}) = \sigma^2/B$?

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

Five observations have the following Euclidean dissimilarity matrix:

$$D = \begin{pmatrix} 0 & 2 & 6 & 10 & 9 \\ 2 & 0 & 5 & 9 & 8 \\ 6 & 5 & 0 & 4 & 5 \\ 10 & 9 & 4 & 0 & 3 \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix}.$$

(i) *(3 %)* Run agglomerative hierarchical clustering with *single* linkage. List *each fusion* and its height, and show the recomputed dissimilarity matrix after *both* of the first two merges.

(ii) *(1 %)* Sketch the resulting dendrogram, labelling the four fusion heights on a properly scaled y-axis. The horizontal ordering of the leaves $\{1, 2, 3, 4, 5\}$ is not unique — pick any valid ordering.

(iii) *(1 %)* If you instead used *complete* linkage on the same matrix D, would the *first* merge change? Justify in one sentence.

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

(i) *(1 %)* Let n training observations be given. Show that the probability a specific observation i is *not* in a given bootstrap sample equals $(1 - 1/n)^n$. (One line of reasoning is enough.)

(ii) *(1 %)* Evaluate the limit of $(1 - 1/n)^n$ as $n \to \infty$, and state the resulting (approximate) probability that observation i *is* in a given bootstrap sample. (Two numeric values.)

(iii) *(1 %)* You want to estimate the standard error of the sample median $\hat{\theta} = \text{median}(X_1, \ldots, X_n)$, for which no clean closed-form sampling distribution is available. Write down (math or pseudocode is fine) a bootstrap algorithm that returns an estimate $\widehat{\text{SE}}_{\text{boot}}(\hat{\theta})$.

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

A statistical-learning class collects a dataset of $n = 506$ neighbourhoods in a metropolitan area. The response variable is the median `price` of an owner-occupied home (in 1000 USD). The continuous predictors are:

- `crim` — per-capita crime rate;
- `rm` — average number of rooms per dwelling;
- `age` — proportion of homes built before 1940 (in %);
- `dis` — weighted distance to five employment centres;
- `tax` — property-tax rate per 10,000 USD;
- `ptratio` — pupil-teacher ratio in local schools;

plus one binary predictor `chas` (1 if the tract borders a river, 0 otherwise).

The data is split into a training set ($n_{\text{train}} = 400$) and a test set ($n_{\text{test}} = 106$). All continuous predictors are standardized to mean 0, standard deviation 1 *before* fitting any of the models below.

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

The course staff first fits an OLS model with an interaction between `rm` (rooms) and `chas` (river dummy):

$$\texttt{price} \sim \texttt{crim} + \texttt{rm} + \texttt{age} + \texttt{dis} + \texttt{tax} + \texttt{ptratio} + \texttt{chas} + \texttt{rm:chas}.$$

The output is:

	Estimate	**Std. Error**	**t-value**	**Pr(>|t|)**
(Intercept)	22.50	0.40	56.25	< 0.001
`crim`	-1.20	0.30	-4.00	< 0.001
`rm`	3.80	0.45	8.44	< 0.001
`age`	-0.20	0.35	-0.57	0.569
`dis`	-1.60	0.45	-3.56	< 0.001
`tax`	-1.10	0.35	-3.14	0.002
`ptratio`	-1.80	0.30	-6.00	< 0.001
`chas`	1.20	0.85	1.41	0.159
`rm:chas`	2.00	0.90	2.22	0.027

Residual standard error: 4.05 on 391 degrees of freedom. Multiple $R^2 = 0.735$, Adjusted $R^2 = 0.730$.

(i) *(1 %)* How many parameters (counting the intercept) does this model estimate? Verify the count against the printed residual degrees of freedom.

(ii) *(2 %)* A classmate says: "Being next to the river raises the price by $\hat{\beta}_{\texttt{chas}} = 1.20$ thousand USD, but the p-value is 0.159, so location on the river is irrelevant." Explain in one or two sentences why this reading is wrong *given the interaction in the model*, and write down the implied effect on `price` of being next to the river for a neighbourhood whose standardized `rm` equals (a) 0 and (b) +1.

(iii) *(1 %)* Holding the other predictors fixed, by approximately how much does the model predict that `price` (in thousands of USD) changes for a 1-standard-deviation increase in `ptratio`? Compute an approximate 95% confidence interval for this effect, using 1.96 as the critical value.

(iv) *(1 %)* Briefly say why `age` ($p = 0.569$) is not by itself sufficient evidence to drop `age` from the model in this dataset. (One sentence.)

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

Suppose we instead apply forward stepwise selection on the $p = 7$ *main-effect* predictors (no interactions). At step 4, the procedure has selected the four predictors {`rm`, `ptratio`, `crim`, `dis`}.

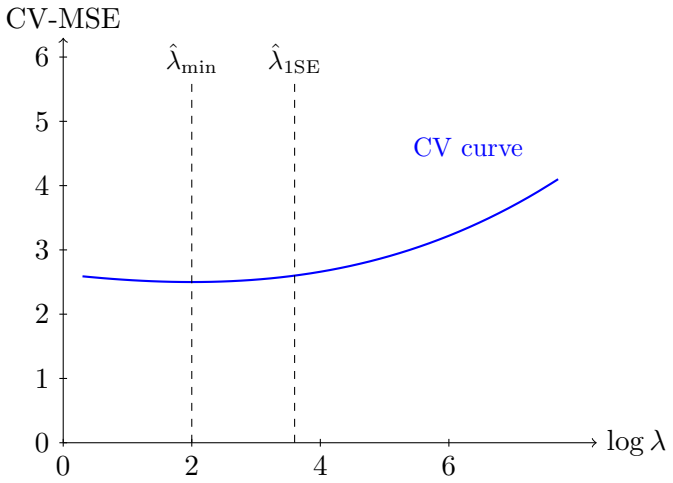
(i) *(1 %)* How many *distinct* candidate models has forward stepwise fit by the time it has chosen its best 4-predictor model? (Recall: include the null model.)

(ii) *(1 %)* How many distinct candidate models would best-subset selection fit by the time it has chosen its best 4-predictor model?

(iii) *(1 %)* The course staff observes that, on this dataset, forward stepwise and best subset *disagree* on which 4-predictor model is best. Give one reason this can happen.

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

The same 7 standardized main-effect predictors are now fed into ridge regression. A 10-fold cross-validation gives the following CV-MSE profile against $\log \lambda$:



The corresponding test MSEs at the two marked λ values, evaluated on the held-out 106 observations, are:

	Test MSE	**No. of nonzero coefficients**
OLS (full main-effect model, no penalty)	19.40	7
Ridge at $\hat{\lambda}_{\min}$	16.20	7
Ridge at $\hat{\lambda}_{1\mathrm{SE}}$	16.95	7

(i) *(1 %)* Write down (math or pseudocode) the objective function that the ridge estimator $\hat{\beta}^R_\lambda$ minimises for a fixed λ, being careful about whether the intercept is or is not penalized.

(ii) *(2 %)* Briefly explain why all seven coefficients remain nonzero in the ridge fits at both $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{1\mathrm{SE}}$, and contrast with what you would expect from a lasso fit on the same data.

(iii) *(1 %)* Why was it important to *standardize* the predictors before running ridge regression?

(iv) *(2 %)* Compare the test MSE of ridge-at-$\hat{\lambda}_{\min}$ (16.20) to OLS (19.40). Interpret in bias–variance terms: what does this tell you about the variance of the OLS estimator on this dataset?

(v) *(1 %)* Without doing any new computation, would you expect the test MSE of a 7-predictor *lasso* fit at the lasso CV-optimal λ to be *lower*, *higher*, or *comparable* to that of the OLS fit, and why? (Answer in one short sentence; a defensible answer earns full credit.)

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

(i) *(2 %)* The same 7 standardized predictors are now fed into principal components regression (PCR). The first few principal-component eigenvalues are

$$\lambda_1 = 2.80, \quad \lambda_2 = 1.60, \quad \lambda_3 = 1.00, \quad \lambda_4 = 0.70, \quad \lambda_5 = 0.40, \quad \lambda_6 = 0.30, \quad \lambda_7 = 0.20.$$

How many principal components must be retained to explain *at least* 80% of the total variance of the standardized X-matrix? Show your calculation.

(ii) *(1 %)* The prof has called principal components regression "*a discretized version of ridge regression.*" Briefly explain in what sense both methods place "pressure" on the same directions of the predictor space, and what makes PCR "discrete" while ridge is "continuous."

(iii) *(1 %)* Suppose, instead, you fit a generalized additive model with a *cubic regression spline* on `rm` with 3 interior knots, and treat all six remaining continuous predictors linearly. How many degrees of freedom does this GAM consume in total, including the intercept? (Treat `chas` as a single dummy.)

e) Boosting interpretation *(3 %)*

A gradient-boosted tree model with $B = 900$ trees, interaction depth $d = 4$, and shrinkage $\nu = 0.05$ is fit on the same training data. The test MSE comes out to 11.40, lower than every other method tried above.

(i) *(1 %)* Explain in one or two sentences what the lower test MSE of boosting (relative to OLS and ridge) suggests about the structure of the relationship between the predictors and `price`.

(ii) *(1 %)* A junior colleague proposes setting $B = 10{,}000$ "just to be safe." Is this a good idea? Why or why not?

(iii) *(1 %)* You want to recover some interpretability from the boosted ensemble. Name one diagnostic plot and one summary statistic the model can produce that help with this. State, for each, what question it does *and does not* answer.

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

A clinic records $n = 600$ male patients followed over ten years. The binary response `chd` (coronary heart disease) equals 1 if the patient developed CHD during follow-up. Predictors of interest:

- `ldl` — LDL cholesterol (mmol/L);
- `age` — in years;
- `sbp` — systolic blood pressure (mmHg);
- `tobacco` — cumulative tobacco use (kg);
- `famhist` — family history (1 = positive, 0 = none);
- `sex` — in this question we additionally include a small companion subgroup with `sex` = 1 (female) recorded; `sex` = 0 (male) is the reference level.

The data is split 400/200 into a training and a test set.

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

You first consider a generative model on a *single* predictor $x = \texttt{ldl}$, with two classes $Y = 0$ (no CHD) and $Y = 1$ (CHD), and you estimate from the training data:

$$\hat{\mu}_0 = 4, \qquad \hat{\mu}_1 = 6, \qquad \hat{\sigma}^2 = 2, \qquad \hat{\pi}_0 = 0.5, \qquad \hat{\pi}_1 = 0.5,$$

under the LDA assumption that the within-class densities are Gaussian with the *shared* variance $\hat{\sigma}^2$.

(i) *(2 %)* Write down the two LDA discriminant scores $\delta_0(x)$ and $\delta_1(x)$ as functions of x. (Use the formula $\delta_k(x) = x\mu_k/\sigma^2 - \mu_k^2/(2\sigma^2) + \log \pi_k$.)

(ii) *(1 %)* Find the LDA decision boundary by solving $\delta_0(x) = \delta_1(x)$ for x. Interpret the answer in one sentence.

(iii) *(1 %)* Suppose now that the priors change to $\hat{\pi}_0 = 0.8$, $\hat{\pi}_1 = 0.2$ but all other parameters stay the same. Where does the decision boundary move to? (You may leave $\log 4$ unevaluated, or use $\log 4 \approx 1.386$.) Briefly say (one sentence) why the direction of the shift makes sense.

(iv) *(3 %)* Now relax the shared-variance assumption: assume $\hat{\sigma}_0^2 = 1$, $\hat{\sigma}_1^2 = 4$ (so the class-1 density is wider), with $\hat{\mu}_0, \hat{\mu}_1, \hat{\pi}_0, \hat{\pi}_1$ as in the original LDA setting. Write down the corresponding *QDA* discriminant scores

$$\delta_k(x) = -\frac{x^2}{2\sigma_k^2} + \frac{x\,\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} - \tfrac{1}{2}\log \sigma_k^2 + \log \pi_k,$$

and explain in one short sentence why the QDA boundary is *quadratic* in x even though the LDA boundary was linear. (You are not required to solve the QDA boundary equation explicitly.)

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

A logistic regression model is fit on the training set using

$$\operatorname{logit}\big(\Pr(\texttt{chd} = 1 \mid X)\big) \;=\; \beta_0 + \beta_1 \texttt{ldl} + \beta_2 \texttt{age} + \beta_3 \texttt{sex} + \beta_4 (\texttt{age:sex}) + \beta_5 \texttt{famhist},$$

with `sex` $= 0$ (male) as the reference. The coefficient table is:

	Estimate	**Std. Error**	**z-value**	**Pr($>\lvert z\rvert$)**
(Intercept)	-5.00	0.80	-6.25	< 0.001
`ldl`	0.40	0.10	4.00	< 0.001
`age`	0.06	0.012	5.00	< 0.001
`sex`	1.20	0.90	1.33	0.183
`age:sex`	-0.04	0.014	-2.86	0.004
`famhist`	0.80	0.22	3.64	< 0.001

(`ldl` is in mmol/L; `age` is in years.)

(i) *(2 %)* State your encoding assumption for `sex` explicitly. For each additional year of `age`, by what factor do the odds of CHD change for (a) a *male* patient and (b) a *female* patient, holding the other predictors fixed? Give both odds-multiplication factors to two decimals.

(ii) *(1 %)* For each additional mmol/L of `ldl`, by what factor do the odds of CHD change, holding the other predictors fixed? Briefly explain why this factor does *not* depend on `sex` in the fitted model.

(iii) *(2 %)* Consider a 55-year-old *male* patient with `ldl` $= 5$ mmol/L and a positive family history. Compute the linear predictor $\hat{\eta}$, the fitted probability $\hat{p}$ of CHD, and the predicted class label at threshold 0.5.

(iv) *(1 %)* A classmate writes: "Because $p_{\text{sex}} = 0.183 > 0.05$, the main effect of sex is statistically insignificant, so we can drop the `sex` term from the model." Briefly say why this reasoning is wrong when the model contains an `age`$\times$`sex` interaction.

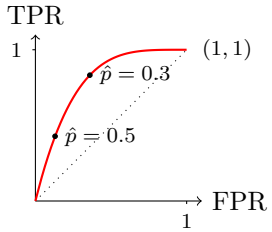
(v) *(2 %)* Define, in plain words *appropriate to this CHD setting*, what *sensitivity* and *specificity* mean for this classifier.

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

At the default threshold $\hat{p} = 0.5$ the logistic model's test-set confusion matrix is:

	Predicted: chd ($\hat{y} = 1$)	**Predicted: no chd** ($\hat{y} = 0$)
Actual: chd	30	50
Actual: no chd	16	104

A clinician complains: "We are missing too many true cases of CHD." To explore alternatives, you produce the ROC curve below:



(i) *(2 %)* From the confusion matrix above, compute the *sensitivity*, *specificity*, and *test error rate* at the threshold $\hat{p} = 0.5$.

(ii) *(1 %)* Briefly define the *AUC* of the ROC curve and state what AUC = 0.5 corresponds to.

(iii) *(2 %)* The clinician proposes lowering the threshold from 0.5 to 0.3. Explain (a) how you would expect *sensitivity* and *specificity* to change, and (b) where the operating point moves on the ROC curve. In one short sentence, name a clinical context in which lowering the threshold like this is the right thing to do.

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

You want to fit a tree-based ensemble to compete with the logistic regression. Choose one method and justify your hyperparameter choices.

(i) *(2 %)* State which tree-based method you would use (e.g. bagging, random forest, or gradient boosting). Give a concrete numeric or symbolic value for *each* of its main tuning parameters and justify each choice with one short sentence. "Sufficiently many trees" alone is not a sufficient justification; tie the choice to the role the parameter plays.

(ii) *(1 %)* Suppose your chosen method yields, on the same test set, sensitivity 0.50, specificity 0.85, and test error rate 0.21. Compare these to the logistic regression result at threshold 0.5 from part (c)(i). Which model would you recommend to the clinic, and based on what criterion?

(iii) *(1 %)* The marginal frequency of CHD in this population is approximately $\Pr(\texttt{chd} = 1) \approx 0.27$. A trivial classifier that always predicts "no CHD" would already achieve a test error rate of about 0.27. Comment in one or two sentences on whether *test error rate* is the most appropriate metric for choosing between classifiers on this dataset, and what alternative metric or pair of metrics you would privilege.

End of exam. Total: $10 + 28 + 16 + 22 + 24 = 100$ points.