TMA4268 Statistical Learning V2026

Mock Exam 5 (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 multiple linear regression, when two or more predictors are highly correlated with each other, the matrix $\mathbf{X}^\top\mathbf{X}$ becomes near-singular and the variances of the estimated coefficients $\hat{\boldsymbol{\beta}}$ inflate. This phenomenon is known as _____ (1) (*collinearity* / *heteroscedasticity* / *leverage* / *influence*). Two classical remedies discussed in this course are ridge regression, which adds a penalty proportional to the squared L2 norm of the coefficient vector and is therefore also called a _____ (2) (*forward-selection* / *shrinkage* / *decorrelation* / *bagging*) method, and principal components regression, which first replaces the correlated predictors by their orthogonal _____ (3) (*principal components* / *residuals* / *Tukey directions* / *leverages*) and then regresses on a subset of those.

In neural networks, fitting a feedforward model with many more parameters than training observations is only possible because we always combine training with one or more forms of regularization. Common explicit choices include L1 / L2 weight decay; randomly zeroing a fraction of node outputs at each training iteration, called _____ (4) (*pruning* / *pooling* / *dropout*

/ *thinning*); halting training when validation error stops improving, called _____ (5) (*warm starting / early stopping / tolerance gating / patience tuning*); and softening one-hot targets so the network does not become over-confident, called _____ (6) (*label smoothing / soft margins / class reweighting / batch normalization*). In addition, training the network with stochastic gradient descent on _____ (7) (*full passes / leave-one-out folds / out-of-bag samples / mini-batches*) provides an *implicit* regularization effect for free, which is one of the reasons over-parameterized networks generalize.

For estimating the test error of a chosen procedure honestly while also using cross-validation to *select* a hyperparameter, the recommended approach is to run two layers of cross-validation, with outer folds for assessment and inner folds for selection. This procedure is known as _____ (8) (*nested cross-validation / stratified CV / leave-one-out CV / bootstrap aggregation*).

In ensemble methods for classification, _____ (9) (*bagging / AdaBoost / random forests / LDA*) builds a sequence of weak classifiers, each fit on training data that have been *re-weighted* so that observations the previous classifiers have misclassified receive higher weight. A modern unifying view interprets this and related methods as fitting a tree at each step to the *negative gradient* of a chosen differentiable loss, giving the general framework called _____ (10) (*gradient boosting / coordinate descent / Newton stepping / posterior averaging*).

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) Bias–variance decomposition and double descent *(3 %)*

Mark each statement true or false.

(i) In the decomposition $\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \text{Bias}^2 + \text{Var} + \sigma^2$, the irreducible term σ^2 can be made smaller by collecting more training data.

(ii) In the over-parameterized regime ($p \gg n$), the bias–variance decomposition no longer holds as an algebraic identity, which is why test MSE can exhibit a second descent past the interpolation peak.

(iii) For a fixed problem and a fixed model class, switching from no regularization to a well-tuned regularizer typically lowers variance *more* than it raises bias, and therefore lowers expected test error.

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

Mark each statement true or false.

(i) Leave-one-out CV (LOOCV) has *lower* bias than 10-fold CV as an estimator of test error, because each LOOCV fit uses $n - 1$ observations.

(ii) In nested cross-validation, the *inner* folds are used to select a hyperparameter (or model), and the *outer* folds are used to give an honest estimate of the chosen procedure's test error.

(iii) For temporally autocorrelated data, randomly assigning observations to k folds gives an *honest* estimate of test error provided k is large enough (say $k \geq 10$).

c) Bootstrap, standard errors and CIs *(3 %)*

(i) *(1 %)* For a training sample of size $n = 10$, what is the probability (rounded to three decimals) that a given original observation is included in a particular bootstrap sample?

(ii) *(1 %)* You wish to obtain a 95% confidence interval for a scalar statistic $\hat{\theta}$ using $B = 1000$ bootstrap resamples. Specify the *percentile* confidence interval as two specific quantiles of the bootstrap distribution $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.

(iii) *(1 %)* Mark each of the following true or false. (A) The bootstrap standard error of $\hat{\theta}$ is the sample standard deviation across the bootstrap replicates $\hat{\theta}_b^*$. (B) The bootstrap can correct for the bias of $\hat{\theta}$ as an estimator of θ. (C) Bootstrap resamples should be drawn *without* replacement to avoid duplicating observations.

d) Neural network regularization *(3 %)*

Mark each statement true or false.

(i) In a feedforward network trained with dropout, the dropout layer is active during both training and test-time prediction.

(ii) Label smoothing replaces hard one-hot targets like $(0, 0, 1, 0)$ with softened targets like $(\varepsilon/3, \varepsilon/3, 1-\varepsilon, \varepsilon/3)$ for a small $\varepsilon > 0$; this tends to make the model less over-confident on the training set and improves generalization in classification tasks.

(iii) Adding L2 weight decay decreases the network's training error.

e) Mini-batch SGD and backpropagation *(3 %)*

Mark each statement true or false.

(i) Backpropagation is the algorithm that *performs* the parameter update at each iteration of mini-batch SGD, by stepping each weight in the direction that most decreases the per-batch loss.

(ii) Mini-batch stochastic gradient descent yields, on each update, an *unbiased* estimator of the full-data gradient.

(iii) In typical practice, batch sizes for mini-batch SGD are chosen as powers of two (e.g. $32, 128, 256$) primarily for hardware-efficiency reasons rather than statistical reasons.

f) Boosting: AdaBoost, gradient boosting, XGBoost *(4 %)*

(i) *(1 %)* In the AdaBoost algorithm, at iteration m the weak classifier $G_m(x)$ has weighted misclassification rate $\text{err}_m = 0.20$ on the current sample weights w_i. Using $\alpha_m = \log \frac{1-\text{err}_m}{\text{err}_m}$, compute the resulting classifier weight α_m (two decimals).

(ii) *(1 %)* For a gradient boosting model with squared-error loss $L(y, f) = \frac{1}{2}(y - f)^2$, give the negative gradient $-\partial L/\partial f$ evaluated at the current ensemble's prediction $\hat{f}^{(m-1)}(x_i)$, and identify what familiar quantity it equals.

(iii) *(2 %)* Mark each statement true or false. (A) In gradient boosting, the number of trees M is a real tuning parameter; using M much larger than necessary will typically over-fit. (B) For a random forest, in contrast, the number of trees B is *not* a serious tuning parameter and one can simply "use enough." (C) Decreasing the learning rate ν in gradient boosting usually *decreases* the number of trees M needed to reach a good fit. (D) XGBoost augments vanilla gradient boosting with second-order (Newton-like) gradient information and with L1 / L2 penalties on the leaf weights.

g) Logistic regression: odds, log-odds, interaction *(3 %)*

(i) *(1 %)* A logistic regression of `chd` (1 = developed coronary heart disease, 0 = did not) on `ldl` (LDL cholesterol, mmol/L) reports $\hat{\beta}_{\text{ldl}} = 0.30$. By what factor (two decimals) do the odds of CHD change for a 2 mmol/L increase in LDL, holding other predictors fixed?

(ii) *(1 %)* A model includes the predictors `age` and `smoker` and an interaction `age:smoker`. The fitted coefficients are $\hat{\beta}_{\text{age}} = 0.04$ and $\hat{\beta}_{\text{age:smoker}} = 0.03$. By what factor (two decimals) do the odds of the event change for a one-year increase in `age` for a *smoker*?

(iii) *(1 %)* True or false: "In a binary logistic regression with a one-unit increase in x_j, the *probability* of $Y = 1$ changes by approximately $\hat{\beta}_j$."

h) PCA: explained variance and loadings *(3 %)*

You perform PCA on $p = 6$ *standardized* variables and obtain eigenvalues

$$\lambda_1 = 2.7,\ \lambda_2 = 1.5,\ \lambda_3 = 0.9,\ \lambda_4 = 0.5,\ \lambda_5 = 0.3,\ \lambda_6 = 0.1.$$

(i) *(1 %)* What is the total variance in the standardized data, and what proportion of this total variance is explained by the first two principal components combined (two decimals)?

(ii) *(1 %)* How many principal components must be retained to capture at least 90% of the total variance? Show your cumulative sum.

(iii) *(1 %)* True or false: "Because the variables were standardized before PCA, the loadings ϕ_{jk} for the k-th principal component depend only on $\mathbf{X}$ and not on the response y."

i) Discriminant analysis and splines *(3 %)*

(i) *(1 %)* True or false: "Under the LDA assumption that the class-conditional densities share a common covariance matrix $\mathbf{\Sigma}$, the resulting Bayes-optimal decision boundary between two classes is linear in $\mathbf{x}$."

(ii) *(1 %)* True or false: "The QDA decision boundary is *quadratic* in $\mathbf{x}$ because the class-specific covariances $\mathbf{\Sigma}_k$ make the term $\mathbf{x}^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}$ in the discriminant fail to cancel between classes."

(iii) *(1 %)* A cubic regression spline (using the truncated-power basis, with a global intercept) on a single predictor x has K interior knots. How many parameters does this spline term consume, *including* the global intercept of the model?

Problem 3 (16 %) — Theory and pseudocode

a) The mathy one — bias–variance decomposition *(7 %)*

Let

$$y_0 = f(x_0) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2, \quad \varepsilon \perp x_0,$$

and let $\hat{f}$ be a model fit on a random training set $\mathcal{D}$, independent of ε. All expectations and variances below are jointly over the random training set $\mathcal{D}$ *and* the noise ε at the test point.

(i) *(2 %)* Show that

$$\mathbb{E}\Big[(y_0 - \hat{f}(x_0))^2\Big] \;=\; \mathbb{E}\Big[(f(x_0) - \hat{f}(x_0))^2\Big] \;+\; \sigma^2.$$

Be explicit about why the cross term vanishes.

(ii) *(3 %)* Starting from the right-hand side of (i) and the trick of adding and subtracting $\mathbb{E}[\hat{f}(x_0)]$, derive

$$\boxed{\mathbb{E}\Big[(y_0 - \hat{f}(x_0))^2\Big] \;=\; \big(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\big)^2 \;+\; \mathrm{Var}\Big(\hat{f}(x_0)\Big) \;+\; \sigma^2.}$$

Identify each term with its standard name (*Bias*^{2}, *Variance*, *Irreducible error*) and explain in one sentence which randomness each term captures.

(iii) *(1 %)* A lasso regression is fit with λ chosen by CV. Compared with ordinary least squares on the same data, the lasso typically has *higher* bias and *lower* variance. In one short sentence each, explain (a) why the bias is generally higher under lasso, and (b) why the variance is generally lower.

(iv) *(1 %)* A student writes: "Bias–variance is just a *trade-off*; you cannot improve both at the same time." Briefly explain why this assertion is too strong in light of regularization and the over-parameterized / double-descent regime. (One short sentence.)

b) Cross-validation: pseudocode and nested CV *(4 %)*

(i) *(2 %)* Write *pseudocode* (math, plain English, or programming-language-agnostic notation) for k-fold cross-validation of a regression model $\mathcal{M}$ at a fixed hyperparameter setting θ, returning the cross-validation MSE estimate $\widehat{\mathrm{CV}}_k(\theta)$. Make explicit (a) how the data is partitioned, (b) what is held out and what is fit at each iteration, and (c) how the per-fold errors are aggregated.

(ii) *(1 %)* Using the pseudocode in (i) as a primitive, write a short description of how to use it to *choose* a hyperparameter θ from a finite grid $\theta \in \{\theta_1, \ldots, \theta_G\}$, and how to use the chosen θ to produce a final fitted model.

(iii) *(1 %)* A colleague proposes the following pipeline: "Step 1: compute the marginal correlation of each of $p = 5000$ predictors with the response y on the full training set; keep the top 25. Step 2: run 10-fold CV on a logistic regression fit using just those 25. The CV misclassification rate is my honest estimate of test error." Briefly state why this estimate is biased *downward*, and describe in one or two sentences the corrected procedure.

c) Bootstrap by hand *(3 %)*

(i) *(1 %)* Show that, for a sample of size n, the probability that a specific observation i is *not* included in a given bootstrap sample is $(1 - 1/n)^n$, and that this probability tends to $1/e \approx 0.368$ as $n \to \infty$. (One line of reasoning is sufficient.)

(ii) *(1 %)* You wish to estimate the standard error of the sample median $\hat{\theta} = \widehat{\text{med}}(X_1, \ldots, X_n)$, for which no clean closed-form expression is available. Describe in 3–4 short sentences (or short pseudocode lines) the bootstrap procedure that estimates $\widehat{\text{SE}}_{\text{boot}}(\hat{\theta})$ using B resamples. Be explicit about whether resampling is with or without replacement.

(iii) *(1 %)* A logistic regression of `default` on `balance` produces $\hat{\beta}_{\text{balance}} = 0.005$ with closed-form $\widehat{\text{SE}}(\hat{\beta}) = 0.00020$ from the inverse Fisher information. A bootstrap with $B = 5000$ resamples gives $\widehat{\text{SE}}_{\text{boot}}(\hat{\beta}) = 0.00050$. Give *one* substantive reason (one sentence) why these two SEs can disagree.

d) Hierarchical clustering by hand *(2 %)*

Four observations have the following Euclidean dissimilarity matrix:

$$D = \begin{pmatrix} 0 & 2 & 8 & 7 \\ 2 & 0 & 9 & 6 \\ 8 & 9 & 0 & 3 \\ 7 & 6 & 3 & 0 \end{pmatrix}.$$

(i) *(1 %)* Run agglomerative hierarchical clustering with *complete* linkage. List the two earliest fusions (which observations / clusters fuse, and at what height).

(ii) *(1 %)* At what height does the *final* fusion (all four observations joined into one cluster) occur under complete linkage? Show the inter-cluster distance you used.

Problem 4 (22 %) — Data analysis: energy use of homes

An energy utility collects $n = 800$ single-family houses' annual heating cost `cost` (in 1000 EUR). The available predictors are:

- `area` — floor area (m^2, continuous);
- `age` — age of the house (years, continuous);
- `insulation` — categorical, 3 levels: *poor* (reference), *standard*, *premium*;
- `heatpump` — binary 0/1: 1 if the house uses a heat pump, 0 otherwise;
- `rooms` — number of rooms (count).

The data are split 600/200 into a training and a test set. All continuous predictors are standardized to mean 0, standard deviation 1 *before* fitting any of the models below; the standardized variables are denoted $\widetilde{\texttt{area}}, \widetilde{\texttt{age}}, \widetilde{\texttt{rooms}}$ (so that one unit in the model corresponds to one standard deviation in the raw variable).

a) Linear regression with a polynomial and an interaction *(7 %)*

The course staff first fit, on the training set, the model

$$\texttt{cost} \sim \widetilde{\texttt{area}} + \widetilde{\texttt{age}} + \widetilde{\texttt{age}}^2 + \texttt{insulation} + \texttt{heatpump} + \widetilde{\texttt{area}}\texttt{:heatpump} + \widetilde{\texttt{rooms}}.$$

The fitted output is:

	Estimate	**Std. Error**	**t-value**	**Pr(>|t|)**
(Intercept)	2.80	0.10	28.00	< 0.001
$\widetilde{\texttt{area}}$	0.80	0.09	8.89	< 0.001
$\widetilde{\texttt{age}}$	−0.05	0.10	−0.50	0.617
$\widetilde{\texttt{age}}^2$	0.30	0.06	5.00	< 0.001
`insulation_standard`	−0.40	0.10	−4.00	< 0.001
`insulation_premium`	−0.80	0.12	−6.67	< 0.001
`heatpump`	−0.60	0.15	−4.00	< 0.001
$\widetilde{\texttt{area}}$`:heatpump`	−0.50	0.12	−4.17	< 0.001
$\widetilde{\texttt{rooms}}$	0.10	0.09	1.11	0.268

Multiple $R^2 = 0.71$, Adjusted $R^2 = 0.706$. Residual standard error: 0.45 on 591 d.f.

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

(ii) *(2 %)* For a house *without* a heat pump, by how much does an increase of one standard deviation in $\widetilde{\texttt{area}}$ change the predicted `cost` (in 1000 EUR), holding all other predictors fixed? Repeat the calculation for a house *with* a heat pump, and briefly comment on the sign of the difference.

(iii) *(2 %)* Consider two houses, both at $\widetilde{\texttt{area}} = 0$, $\widetilde{\texttt{rooms}} = 0$, `insulation` = *poor*, `heatpump` = 0. House A has $\widetilde{\texttt{age}} = -1$ (one SD younger than average); house B has $\widetilde{\texttt{age}} = +2$ (two SDs older than average). Compute the predicted `cost` of each (two numerics, 1000 EUR). Briefly comment on what the $\widetilde{\texttt{age}}^2$ term implies about the marginal effect of age at older versus younger houses.

(iv) *(1 %)* A classmate writes: "$\widetilde{\texttt{age}}$ has a p-value of 0.617, so we should drop it from the model." In one or two sentences, explain why this is wrong given the rest of the fitted model. (Hint: marginal effect of $\widetilde{\texttt{age}}$ depends on both $\hat{\beta}_{\widetilde{\texttt{age}}}$ and $\hat{\beta}_{\widetilde{\texttt{age}}^2}$, and this course follows the hierarchical-principle convention.)

(v) *(1 %)* The main-effect coefficient on `heatpump` is $\hat{\beta}_{\text{heatpump}} = -0.60$. A classmate concludes: "On average, houses with a heat pump cost 600 per year less to heat than houses without." Briefly explain why this is misleading *given the interaction in the model*, and write down the actual change in predicted `cost` (in 1000 EUR) of installing a heat pump for a house at standardized area $\widetilde{\texttt{area}} = +1$.

b) Collinearity diagnosis *(4 %)*

In a separate, more realistic version of the same dataset, the predictors $\widetilde{\texttt{area}}$ and $\widetilde{\texttt{rooms}}$ are found to have empirical correlation $r = 0.92$. The course staff fit two models, both *without* the interaction and quadratic terms, on the same training set:

Model	$\hat{\beta}_{\widetilde{\texttt{area}}}$	**SE**	$\hat{\beta}_{\widetilde{\texttt{rooms}}}$	**SE**	**Adjusted** R^2
$\widetilde{\texttt{area}}$ only	0.90	0.06	—	—	0.43
$\widetilde{\texttt{area}} + \widetilde{\texttt{rooms}}$	0.55	0.15	0.42	0.16	0.44

(i) *(1 %)* Briefly identify the *two* numerical symptoms in the second row of the table above that are characteristic of collinearity.

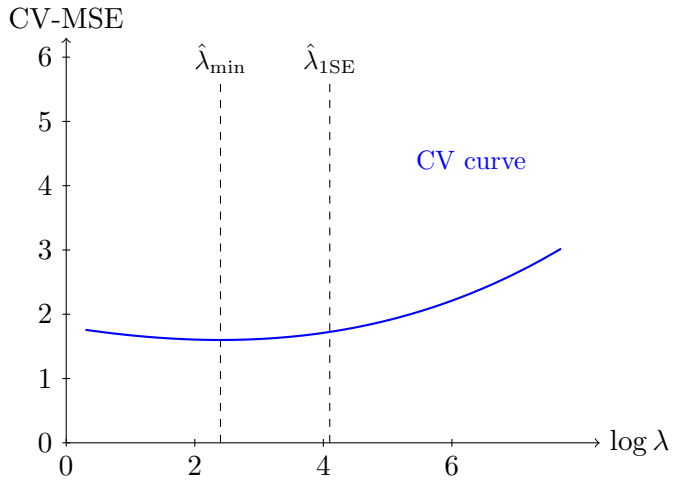
(ii) *(1 %)* Explain in one short sentence the source of this behavior in terms of $(\mathbf{X}^\top\mathbf{X})^{-1}$ and the fact that $\widetilde{\texttt{area}}$ and $\widetilde{\texttt{rooms}}$ are nearly the same column of the design matrix.

(iii) *(1 %)* Briefly comment on whether *predictions* of `cost` from the second model are also expected to be unstable, given that the joint contribution $\hat{\beta}_{\widetilde{\texttt{area}}} \cdot \widetilde{\texttt{area}} + \hat{\beta}_{\widetilde{\texttt{rooms}}} \cdot \widetilde{\texttt{rooms}}$ is well-determined.

(iv) *(1 %)* Name two methods discussed in this course that one could apply to address the collinearity *without* simply dropping one of the two predictors, and for each one state in one short sentence the trade-off it makes (in terms of bias, variance, or interpretability).

c) Lasso with 10-fold CV *(5 %)*

The same $p = 8$ *standardized* predictors from part (a) (including the $\widetilde{\texttt{age}}^2$ and the $\widetilde{\texttt{area}}$: `heatpump` interaction) are now fed into a lasso regression. A 10-fold cross-validation gives the CV-MSE profile



The corresponding test MSEs on the held-out 200 observations:

	Test MSE	**No. of nonzero coefs (excl. intercept)**
OLS (full model from (a))	0.205	8
Lasso at $\hat{\lambda}_{\min}$	0.196	6
Lasso at $\hat{\lambda}_{1\mathrm{SE}}$	0.203	4

(i) *(1 %)* The `insulation` factor has been entered into the design matrix as two dummies (`insulation_standard` and `insulation_premium`). Briefly explain why the resulting lasso path is *not* invariant to the choice of reference level (i.e. which level is dropped). What practical consequence does this have for interpretation when one of the dummies is shrunk to exactly zero?

(ii) *(1 %)* State the *one-standard-error rule* precisely, and give one reason a practitioner might choose $\hat{\lambda}_{1\mathrm{SE}}$ over $\hat{\lambda}_{\min}$.

(iii) *(1 %)* On this dataset, lasso at $\hat{\lambda}_{\min}$ improves test MSE from 0.205 to 0.196, a small improvement, and drops 2 of 8 coefficients. Interpret in bias–variance terms: what does the small improvement plus a modest reduction in active predictors suggest about the importance of the dropped predictors?

(iv) *(2 %)* Suppose the practitioner cares primarily about *interpretability* and would prefer a simpler model. Which of $\hat{\lambda}_{\min}$ and $\hat{\lambda}_{1\mathrm{SE}}$ would you recommend, and why? Give a concrete claim about the trade-off the recommendation makes in test MSE terms (refer to the numbers in the table).

d) Gradient boosting interpretation and pseudocode *(6 %)*

A gradient-boosted regression tree (squared-error loss) is fit on the same training data, achieving test MSE = 0.151, lower than the lasso.

(i) *(2 %)* Write *pseudocode* for the squared-error gradient-boosting algorithm, with B trees, learning rate ν and tree depth d as hyperparameters. Make explicit (a) the initialization, (b) the per-iteration step (a fit-residuals view is fine for squared-error loss), and (c) the final returned function. Two short sentences of accompanying explanation are fine if helpful.

(ii) *(1 %)* For *squared-error* loss, derive in one or two lines that fitting a tree to the current residuals $r_i^{(m)} = y_i - \hat{f}^{(m-1)}(x_i)$ is equivalent to fitting a tree to the negative gradient of the loss with respect to the function values $\hat{f}^{(m-1)}(x_i)$.

(iii) *(1 %)* Briefly describe how you would tune the three boosting hyperparameters (B, d, ν) in practice, and tie each one's role to a single sentence about bias–variance.

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

(v) *(1 %)* The gradient boosting test MSE (0.151) is substantially below the OLS / lasso test MSE (≈ 0.20). Briefly interpret what this gap suggests about the structure of the underlying relationship between predictors and `cost`, and name one diagnostic plot you would compute from the boosted ensemble to recover some *interpretability*.

Problem 5 (24 %) — Data analysis: hospital-readmission classification

A regional hospital records $n = 3000$ patients discharged after a cardiac event and follows each for 90 days. The binary response `readmit` is 1 if the patient is readmitted within 90 days and 0 otherwise. The available predictors are:

- `age` — age in years (continuous);
- `los` — length of initial hospital stay (days, continuous);
- `prev_adm` — number of prior admissions in the past year (count, 0–10);
- `diabetes` — binary 0/1 indicator;
- `insulin` — binary 0/1 indicator (insulin therapy during stay);
- `sex` — binary 0/1 (female = 0 reference, male = 1).

The data are split 70/30 into training (2100) and test (900) sets. Among the test set, 180 patients are truly readmitted and 720 are not.

a) Logistic regression with an interaction *(7 %)*

A logistic regression model is fit on the training set:

$$\operatorname{logit}\big(\Pr(\texttt{readmit} = 1 \mid X)\big) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{los} + \beta_3 \texttt{prev\_adm} + \beta_4 \texttt{diabetes} + \beta_5 \texttt{insulin} + \beta_6 (\texttt{diabetes:insul}$$

	Estimate	**Std. Error**	**z-value**	**Pr(> |z|)**
(Intercept)	−4.20	0.45	−9.33	< 0.001
`age`	0.030	0.005	6.00	< 0.001
`los`	0.080	0.020	4.00	< 0.001
`prev_adm`	0.40	0.06	6.67	< 0.001
`diabetes`	0.10	0.18	0.56	0.578
`insulin`	0.05	0.20	0.25	0.804
`diabetes:insulin`	0.90	0.30	3.00	0.003
`sex`	0.15	0.12	1.25	0.211

(i) *(1 %)* State your encoding assumption for the binary predictors `diabetes` and `insulin` explicitly. By what factor (two decimals) do the odds of readmission multiply for each additional prior admission `prev_adm`, holding the other predictors fixed?

(ii) *(2 %)* For each of the four combinations of (`diabetes`, `insulin`) $\in \{(0,0),(1,0),(0,1),(1,1)\}$, write down the joint contribution of the three diabetes / insulin terms to the linear predictor $\hat{\eta}$. By what factor do the odds of readmission change for a diabetic-and-on-insulin patient compared with a reference patient (`diabetes` = 0, `insulin` = 0), holding all other predictors fixed? (Two decimals.)

(iii) *(2 %)* Consider a 70-year-old male patient with `los` = 6 days, `prev_adm` = 2, `diabetes` = 1, `insulin` = 1. Compute the linear predictor $\hat{\eta}$ and the predicted probability $\hat{p}$ of readmission. Show your work.

(iv) *(1 %)* A classmate writes: "$\hat{\beta}_{\text{diabetes}} = 0.10$ with p-value 0.578 and $\hat{\beta}_{\text{insulin}} = 0.05$ with p-value 0.804. Both are insignificant, so we should drop both `diabetes` and `insulin` from the model." In one or two sentences, explain why this is wrong given that the model also contains an `diabetes:insulin` interaction.

(v) *(1 %)* Define, *in plain words appropriate to this readmission setting*, what *sensitivity* means for this classifier. (One sentence.)

b) Confusion matrix and threshold tuning *(3 %)*

On the held-out test set, the logistic model with default threshold $\hat{p} = 0.5$ produces the confusion matrix

	Predicted: readmit	**Predicted: no readmit**
Actual: readmit	54	126
Actual: no readmit	72	648

(i) *(1 %)* Compute, to two decimals, the test *sensitivity*, *specificity*, and overall *error rate* of the classifier.

(ii) *(1 %)* The care-coordination team argues that "we are missing too many true readmissions." Indicate the direction (up / down) in which *sensitivity* and *specificity* are expected to move if the classification threshold is lowered from 0.5 to 0.3.

(iii) *(1 %)* In the readmission context (where a false negative means a high-risk patient is sent home without targeted follow-up), explain in one sentence which of *sensitivity* and *specificity* you would prioritize, and why.

c) AdaBoost: pseudocode and a small numerical iteration *(6 %)*

To compare with logistic regression, an AdaBoost classifier with binary weak learners (stumps) is also fit. The classifier returns a label $G(x) \in \{-1, +1\}$ (with $+1$ encoding readmission).

(i) *(2 %)* Write *pseudocode* for the AdaBoost algorithm with M rounds. Make explicit (a) the initialization of the sample weights w_i, (b) the weighted misclassification rate err_m, (c) the classifier weight α_m, (d) the per-observation weight update, and (e) the final ensemble classifier.

(ii) *(2 %)* A tiny illustrative example: at iteration $m = 1$ on a training sample of $N = 10$ observations, the initial sample weights are $w_i = 1/10$. The stump $G_1(x)$ misclassifies exactly two of the ten observations. Compute the weighted misclassification rate err_1, the classifier weight α_1, and the *updated, un-normalized* weights w_i for (a) a correctly-classified observation and (b) a misclassified one. Round numerics to two decimals.

(iii) *(1 %)* On the same test set, the AdaBoost ensemble (with $M = 200$ stumps) achieves sensitivity 0.55, specificity 0.86, and error rate 0.16. Compare these to the logistic regression numbers from (b)(i). Which model would you recommend to the hospital, and on which criterion?

(iv) *(1 %)* True or false, with one sentence of justification: "AdaBoost can be derived as a special case of gradient boosting, in which the loss function is the *exponential loss* $L(y, f) = e^{-yf}$."

d) A neural network classifier and a backprop mini-derivation *(8 %)*

You also fit a feed-forward neural network with the following architecture (using the standardized version of the six predictors above, plus a derived $\widetilde{\texttt{prev\_adm}^2}$ to capture non-linearity, for $p = 7$ inputs):

- input layer of 7 standardized predictors;
- hidden layer of $M = 10$ neurons (with biases, ReLU activations);
- output layer of 1 neuron (with bias, sigmoid activation, returning $\hat{p}$).

(i) *(1 %)* How many parameters does this network have in total, *including all bias terms*? Show your layer-by-layer breakdown.

(ii) *(1 %)* A neuron in the hidden layer has weights $w = (0.20, -0.50, 1.00, 0.40, -0.60, 0.30, 0.10)$ and bias $b = -0.20$. Its inputs on a given observation are $x = (0.5, -0.5, 2, 1, 0, 1, -1)$. Compute the output of this neuron under a ReLU activation, rounded to two decimals.

(iii) *(3 %)* A mini backpropagation derivation. Consider, for a single observation (x_i, y_i), the simple network

$$z_k = g\left(\alpha_{k0} + \sum_{j=1}^{p} \alpha_{kj}\, x_{ij}\right), \qquad \hat{f}(x_i) = \beta_0 + \sum_{k=1}^{M} \beta_k z_k,$$

with squared-error loss $L_i = \frac{1}{2}(y_i - \hat{f}(x_i))^2$ and a generic differentiable activation $g(\cdot)$.

(a) *(1 %)* Compute $\partial L_i / \partial \hat{f}(x_i)$.

(b) *(1 %)* Using the chain rule and your answer to (a), compute $\partial L_i / \partial \beta_k$ for $k \geq 1$ in terms of z_k, y_i, and $\hat{f}(x_i)$.

(c) *(1 %)* Using the chain rule once more, compute $\partial L_i / \partial \alpha_{kj}$ in terms of y_i, $\hat{f}(x_i)$, β_k, x_{ij}, and $g'(v_{ik})$, where $v_{ik} = \alpha_{k0} + \sum_j \alpha_{kj} x_{ij}$ is the pre-activation of hidden unit k.

(iv) *(2 %)* Briefly state *two* regularization techniques from this course that you would apply to this network. For each, give a concrete numeric value of any associated hyperparameter where applicable, and one short sentence on what role it plays. *(One of the two must be a technique other than L2 weight decay.)*

(v) *(1 %)* A friend trains the same network with mini-batch SGD (batch size 128) and no explicit weight-decay penalty, and observes that it generalizes *better* than the much smaller logistic regression in (a). Briefly state one feature of mini-batch SGD that, according to the lectures, can rationalize this surprising outcome.

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