TMA4268 Statistical Learning V2026

Mock Exam 1 (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 for each blank from the choices in parentheses. Each correct fill is worth 1 %.

In this course we studied methods that learn from data. Our methods were broadly divided into _____ (1) (*supervised / unsupervised / parametric / deterministic*) and unsupervised learning. In the supervised setting we distinguished problems with a quantitative response (regression) from problems with a qualitative response (_____ (2) (*clustering / classification / dimensionality reduction*)).

We also separated two motivations: when we care mostly about the actual coefficients of the fitted model and what they mean, we are doing _____ (3) (*prediction / inference / interpolation / regularization*). When we instead only care about $\hat{y}$ being close to the true y on unseen data, we are doing _____ (4) (*prediction / inference / cross-validation / shrinkage*).

A central theme of the course was the bias–variance decomposition. The expected squared test error at a fixed point x_0 splits into three pieces: a squared-bias term, a variance term, and an _____ (5) (*irreducible / reducible / biased / averaged*) error. Methods that explicitly

trade a little extra bias for a large reduction in variance are called _____ (6) (*regularization methods / boosted methods / generative methods / bootstrap methods*).

For classification we discussed both *generative* approaches (which model $P(X \mid Y = k)$ and the class priors) and *discriminative* approaches (which model $P(Y \mid X)$ directly). Among the methods we studied, _____ (7) (*logistic regression / LDA / KNN / ridge regression*) belongs to the discriminative family, while LDA and QDA are _____ (8) (*generative / nonparametric / distance-based*).

Finally, in unsupervised learning we studied _____ (9) (*principal component analysis / ridge regression / the bootstrap*) as a linear dimensionality-reduction tool, and clustering methods such as K-means and _____ (10) (*hierarchical clustering / logistic clustering / kernel PCA*).

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

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

a) Bias–variance reasoning *(3 %)*

Which of the following statements are true?

(i) As the flexibility of a fitted model increases, its squared bias generally decreases.

(ii) As the flexibility of a fitted model increases, its variance generally decreases.

(iii) In a high-noise (large σ^2) setting, a very flexible model is generally preferable to a less flexible one.

(iv) The irreducible error σ^2 can be driven arbitrarily close to zero by collecting more training data.

b) Ridge vs. Lasso *(3 %)*

Which of the following are true?

(i) Increasing the penalty parameter λ in ridge regression shrinks coefficients toward zero but, for any finite λ, never makes any coefficient exactly zero.

(ii) Lasso can produce sparse models because the ℓ_1 penalty ball has corners on the coordinate axes.

(iii) Ridge regression is more flexible than ordinary least squares.

(iv) Standardizing the predictors before fitting ridge or lasso is generally recommended.

c) Neural network parameters and forward pass *(4 %)*

Suppose you build a fully-connected feed-forward neural network with:

- 5 input variables,
- one hidden layer of 6 neurons (with bias),
- one output neuron (with bias).

(i) *(2 %)* How many weights (parameters) does the network have in total, *including bias terms*?

(ii) *(2 %)* A neuron in the hidden layer has weights $w_1 = 0.5$, $w_2 = -1.0$, $w_3 = 0.0$, $w_4 = 2.0$, $w_5 = -0.5$ and bias $b = 0.3$. Its inputs on a given observation are $x_1 = 2$, $x_2 = 1$, $x_3 = -3$, $x_4 = 0.5$, $x_5 = 4$. Using a ReLU activation, what is the output of this neuron?

d) Odds and probability *(3 %)*

(i) *(1 %)* A study reports that the odds of an event are 2.5. What is the probability of the event?

(ii) *(1 %)* An individual has a 30% probability of defaulting on her credit card. What are her odds of defaulting?

(iii) *(1 %)* In a logistic-regression fit, the coefficient on a binary predictor D (smoker = 1, non-smoker = 0) equals $\hat{\beta} = 0.8$. By what factor do the odds of the response event change for a smoker relative to a non-smoker, holding all other predictors fixed? (One numeric value.)

e) Smoothing splines and degrees of freedom *(3 %)*

Consider a smoothing spline fit with smoothing parameter λ.
Which of the following are true?

(i) As $\lambda \to 0$, the fit approaches the rough interpolating spline that passes through every training point.

(ii) As $\lambda \to \infty$, the fit approaches the ordinary-least-squares straight line.

(iii) Increasing λ *increases* the effective degrees of freedom of the fit.

(iv) The effective degrees of freedom $\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda)$ of a smoothing spline is, in general, not an integer.

f) Principal component analysis *(5 %)*

You perform PCA on a dataset with six **standardized** variables $X_1, \ldots, X_6$ and obtain:

PC	Eigenvalue	PVE
PC1	2.4	0.40
PC2	1.5	0.25
PC3	1.2	0.20
PC4	0.5	≈0.083
PC5	0.3	0.05
PC6	0.1	≈0.017

The loading vector of the first principal component (entries rounded to two decimal places) is

$$\phi_1 = (\,0.60,\ 0.50,\ 0.40,\ 0.30,\ 0.30,\ 0.20\,)^\top.$$

A new observation has standardized values

$$x^* = (\,1,\ -1,\ 0.5,\ 2,\ 0,\ -1\,)^\top.$$

(i) *(1 %)* What is the total variance in the standardized data?

(ii) *(1 %)* What proportion of the total variance is explained by PC1 and PC2 together?

(iii) *(1 %)* How many principal components must be kept to retain at least 85% of the total variance?

(iv) *(1 %)* Compute the score z_1^* of the new observation on PC1.

(v) *(1 %)* Which one variable contributes *most* to PC1?

g) Sensitivity, specificity, ROC, and AUC *(3 %)*

A classifier is applied to a test set of 1000 patients, of whom 200 truly have the disease. At a chosen threshold the confusion matrix is:

	Predicted: disease	**Predicted: no disease**
Actual: disease	140	60
Actual: no disease	80	720

(i) *(1 %)* Compute the *sensitivity* of the classifier.

(ii) *(1 %)* Compute the *specificity* of the classifier.

(iii) *(1 %)* Which of the following is true about the AUC (area under the ROC curve)? Mark all that apply.

(A) AUC $= 0.5$ corresponds to a classifier no better than random guessing.

(B) Two classifiers with the same AUC must produce ROC curves of the same shape.

(C) The ROC curve plots true positive rate (TPR) on the y-axis against false positive rate (FPR) on the x-axis, varying the classification threshold.

(D) AUC is unaffected by class imbalance in the way raw accuracy is.

h) Why are nonlinear activations necessary? *(2 %)*

Mark all that are true. In a feed-forward neural network, the hidden-layer activation function is nonlinear because...

(i) it allows the network to approximate complex, non-linear functions of the inputs;

(ii) without it the entire network collapses to an affine map of the input, no matter how many hidden layers are stacked;

(iii) it acts as an implicit regularizer that shrinks the hidden weights toward zero;

(iv) it makes the empirical loss surface convex in the network weights.

i) LDA vs. QDA *(2 %)*

Mark all that are true.

(i) LDA assumes a single shared within-class covariance matrix Σ, whereas QDA allows a separate Σ_k for each class.

(ii) Under LDA's assumption, the resulting Bayes decision boundary is *linear* in the features because the quadratic term $\mathbf{x}^\top \Sigma^{-1} \mathbf{x}$ is identical for every class and therefore cancels when discriminant functions are compared.

(iii) In a regime with very few training observations per class, QDA generally outperforms LDA because QDA is more flexible.

(iv) LDA and QDA are both *generative* classifiers: they model $P(\mathbf{X} \mid Y = k)$ and the class priors π_k and then apply Bayes' rule.

Problem 3 (16 %) — Theory and hand calculations

a) The mathy one — maximum likelihood and least squares *(8 %)*

Consider a regression setting in which the response is modelled as

$$Y_i = f_\theta(X_i) + \varepsilon_i, \qquad i = 1, \dots, n,$$

where f_θ is a regression function indexed by an unknown parameter vector θ, the inputs X_i are treated as fixed, and the errors ε_i are independent and identically distributed (i.i.d.).

(i) *(2 %)* Write down the optimization problem that defines the *ordinary least-squares* estimator $\hat{\theta}_{\mathrm{LS}}$ of θ. State the minimal assumptions on ε_i needed for this estimator to be well-defined.

(ii) *(2 %)* Now additionally assume $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Write down the log-likelihood $\ell(\theta, \sigma^2)$ of the sample $(Y_1, \dots, Y_n)$.

(iii) *(3 %)* Show that the maximum-likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ and the least-squares estimator $\hat{\theta}_{\mathrm{LS}}$ *coincide* under the Gaussian-error assumption. Be explicit about which terms in your log-likelihood do and do not depend on θ.

(iv) *(1 %)* Does this equivalence still hold if you replace the Gaussian assumption with $\varepsilon_i \sim$ Laplace$(0, b)$? Briefly justify your answer (one or two sentences).

b) Hierarchical clustering by hand *(5 %)*

Four observations have the following Euclidean dissimilarity matrix:

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

(i) *(3 %)* Sketch the dendrogram produced by hierarchical clustering with *complete* linkage. Label the leaves $1, 2, 3, 4$, indicate the fusion height of each merge on a properly labelled y-axis, and show the recomputed dissimilarity matrix after the first merge.

(ii) *(1 %)* If instead you use *single* linkage on the same matrix, what would change about the *first* merge? Justify in one sentence.

(iii) *(1 %)* Suppose you cut the complete-linkage dendrogram at height 7. How many clusters do you obtain, and which observations are in each cluster?

c) The bias–variance decomposition *(3 %)*

(i) *(2 %)* Let $y_0 = f(x_0) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$, and let $\hat{f}$ be a model fit on a random training set, with $\hat{f}$ independent of ε. Write down the bias–variance decomposition of the expected squared test error $\mathbb{E}\Big[(y_0 - \hat{f}(x_0))^2\Big]$ as a sum of three terms. *Identify each term by name* and *state what the expectations are taken over.*

(ii) *(1 %)* In one or two sentences, give a reason why one might prefer the label *bias–variance decomposition* over *bias–variance trade-off.* (You may, for example, appeal to the squared form of the bias term, the role of a clever model choice such as regularization, or the over-parameterized regime in which both bias and variance can simultaneously decrease.)

Problem 4 (22 %) — Data analysis: regression on used-car prices

A dealership records $n = 500$ used-car transactions. The response is the sale `price` (in 1000 EUR). The available predictors are:

- `mileage` — total mileage in 1000 km (continuous);
- `age` — age of the car in years (continuous);
- `horsepower` — engine power, hp (continuous);
- `engine` — engine displacement, litres (continuous);
- `mpg` — combined fuel economy (continuous);
- `fuel` — categorical: *petrol* (reference), *diesel*, *electric*.

The course staff fit the linear model

$$\texttt{price} \sim \texttt{mileage} + \texttt{age} + I(\texttt{age}^2) + \texttt{horsepower} + \texttt{engine} + \texttt{mpg} + \texttt{fuel}$$

on a training set ($n_{\text{train}} = 350$, the remaining 150 cars form the test set). The output is:

	Estimate	**Std. Error**	**t-value**	**Pr($>\lvert t\rvert$)**
(Intercept)	24.50	2.30	10.65	< 0.001
`mileage`	-0.085	0.012	-7.08	< 0.001
`age`	-1.20	0.40	-3.00	0.003
$I(\texttt{age}^2)$	0.030	0.020	1.50	0.135
`horsepower`	0.045	0.011	4.09	< 0.001
`engine`	1.80	0.65	2.77	0.006
`mpg`	0.12	0.08	1.50	0.135
`fuel_diesel`	-0.65	0.55	-1.18	0.239
`fuel_electric`	3.20	1.10	2.91	0.004

Residual standard error: 2.10 on 341 degrees of freedom. Multiple $R^2 = 0.621$, Adjusted $R^2 = 0.612$. F-statistic on the categorical `fuel` variable (joint test of both dummies against `fuel = petrol`): $F = 4.85$ on $(2, 341)$ d.f., $p = 0.008$.

a) Reading the table *(6 %)*

(i) *(1 %)* How many parameters does this model estimate? (i.e. how many degrees of freedom does the model consume, including the intercept?) Justify your count briefly.

(ii) *(2 %)* An electric car is, on average, how much more or less expensive than an otherwise identical petrol car? Quantify both the point estimate and an approximate 95% confidence interval (you may use 1.96 as the critical value).

(iii) *(2 %)* Is there evidence that the categorical variable `fuel` is *jointly* associated with `price`, after controlling for the other predictors? **Name the test you would use** and state your conclusion based on the output above. (You do *not* need to compute the test statistic by hand.)

(iv) *(1 %)* A friend reads the table and concludes: "`mpg` is irrelevant to the price." Briefly say in what sense she is right and in what sense she may be wrong (one or two sentences).

b) Training vs. test error: the keyword *(3 %)*

Suppose you fit a second model that uses the same predictors as above but *also* adds the cubic term $I(\texttt{age}^3)$ and the redundant term $I(\texttt{age}/12)$ (i.e. `age` re-expressed in different units).

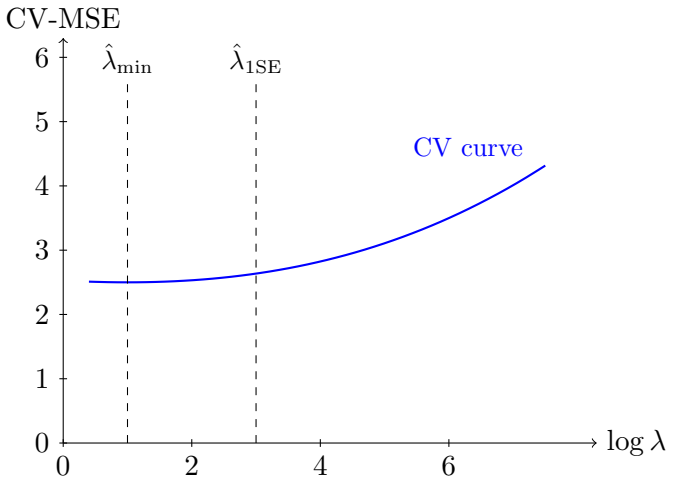
(i) *(1 %)* Will the *training* RSS of the larger model be lower than, equal to, or higher than that of the smaller model? Justify.

(ii) *(1 %)* What about the *test* RSS? Can you be certain about the direction? Justify.

(iii) *(1 %)* What problem does the redundant term $I(\texttt{age}/12)$ create at fitting time, and what would happen if you added an ℓ_2 penalty?

c) Lasso with 10-fold cross-validation *(6 %)*

The same predictors are now fed into a lasso regression. A 10-fold cross-validation gives the following profile of CV-MSE against $\log \lambda$:



(You may assume that the prediction-test MSEs you would obtain if you fitted the lasso at $\hat{\lambda}_{\text{min}}$ and at $\hat{\lambda}_{1SE}$ are 5.20 and 5.55 respectively. The unregularized OLS model from part (a), evaluated on the same test set, gives test MSE 5.15.)

(i) *(1 %)* Write down, in math or pseudocode, how the 10-fold CV-MSE estimate at a fixed λ is computed.

(ii) *(1 %)* Which value of λ would you pick from the plot, and why? (Either $\hat{\lambda}_{\text{min}}$ or $\hat{\lambda}_{1SE}$ is acceptable, but justify.)

(iii) *(2 %)* Compare the test MSE of the lasso (at $\hat{\lambda}_{\text{min}}$) to the test MSE of OLS. Interpret the result in bias–variance terms: what does it tell you about the predictors in this dataset?

(iv) *(2 %)* On the plot above, the vertical dashed lines mark $\hat{\lambda}_{\text{min}}$ (the λ with the smallest CV error) and $\hat{\lambda}_{\text{1SE}}$ (the largest λ whose CV error is within one standard error of the minimum). Briefly state the *one-standard-error rule* and explain when a practitioner might prefer $\hat{\lambda}_{\text{1SE}}$ over $\hat{\lambda}_{\text{min}}$.

d) Tree boosting *(5 %)*

The same training set is now fit with a gradient-boosted tree model, evaluated on the same test set. The resulting test MSE is 3.85, lower than the OLS and lasso test MSEs.

(i) *(2 %)* Gradient boosting has three main hyperparameters: the number of trees B, the interaction depth d, and the shrinkage (learning rate) η. For each, briefly say how you would choose its value and how it affects the bias–variance behaviour of the final model.

(ii) *(1 %)* Why is it incorrect to claim that B "does not matter as long as it is large enough"? (One sentence.)

(iii) *(2 %)* What does the lower test MSE of boosting compared to OLS suggest about the structure of the relationship between the predictors and `price`? What would you check next if you also wanted *interpretation*, not just prediction?

e) GAM degrees of freedom *(2 %)*

Finally, you fit a generalized additive model (GAM) of the form

$$\texttt{price} = \beta_0 + f_1(\texttt{mileage}) + f_2(\texttt{age}) + f_3(\texttt{horsepower}) + \beta_4 \cdot \texttt{engine} + \beta_5 \cdot \texttt{mpg} + g(\texttt{fuel}) + \varepsilon,$$

where f_1, f_2, f_3 are cubic regression splines, each with 4 internal knots (and an intercept absorbed into β_0). The categorical `fuel` term $g(\texttt{fuel})$ uses dummy coding as in part (a).

(i) *(1 %)* How many degrees of freedom does the spline $f_1(\texttt{mileage})$ alone consume? (Count basis functions, excluding the global intercept.)

(ii) *(1 %)* How many total degrees of freedom does the GAM consume (including the intercept)?

Problem 5 (24 %) — Data analysis: classification of loan defaults

A bank has data on $n = 10{,}000$ customers; the response `default` is binary (1 if the customer defaulted within the year, 0 otherwise). Predictors include:

- `balance` — average outstanding monthly balance, in 1000 EUR;
- `income` — annual income, in 1000 EUR;
- `sex` — categorical: *male* (reference), *female*;
- six payment-history variables `pay_0`, ..., `pay_5`.

The data are split 70/30 into a training set (7000) and a test set (3000).

a) Logistic regression with interaction *(10 %)*

A logistic regression model is fit on the training set, using all predictors plus the interaction `balance:sex`. The coefficient table for the most important rows is:

	Estimate	**Std. Error**	**z-value**	**Pr($>\|z\|$)**
(Intercept)	-10.80	0.45	-24.0	< 0.001
`balance`	1.00	0.05	20.0	< 0.001
`income`	-0.010	0.0025	-4.0	< 0.001
`sex_female`	-0.18	0.35	-0.51	0.610
`balance:sex_female`	0.25	0.08	3.13	0.002
`pay_0`	0.55	0.04	13.8	< 0.001
`pay_1`	0.20	0.04	5.0	< 0.001
$\cdots$ (other `pay_k` omitted for brevity)				

(Recall: "balance" is measured in *thousands of EUR*; "income" in thousands of EUR.)

(i) *(3 %)* For a *male* customer, by what factor do the odds of default change for each 1000 EUR increase in `balance`, holding all other predictors fixed? Repeat the calculation for a *female* customer. (Two numeric factors, each rounded to two decimals.)

(ii) *(2 %) State* your encoding assumption explicitly, then explain why the coefficient on `sex_female` alone (-0.18) is *not* a useful summary of the average gender effect when an interaction is present.

(iii) *(2 %)* Define, *in words appropriate to this specific problem*, what *sensitivity* and *specificity* mean for the default classifier.

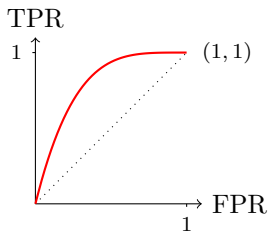
(iv) *(3 %)* At a threshold of $\hat{p} = 0.5$, the logistic-regression model produces the following confusion matrix on the test set:

	Predicted: default ($\hat{y} = 1$)	**Predicted: no default ($\hat{y} = 0$)**
Actual: default	270	330
Actual: no default	80	2320

Compute the *sensitivity*, *specificity*, and *test error rate*. (Three numbers.)

b) ROC and AUC interpretation *(3 %)*

A second researcher complains that the chosen threshold $\hat{p} = 0.5$ "misses too many defaulters". She produces the ROC curve below (schematic):



(i) *(1 %)* What is plotted on the two axes of the ROC curve?

(ii) *(1 %)* Define the AUC and state what AUC $= 0.5$ corresponds to.

(iii) *(1 %)* The researcher proposes lowering the threshold from 0.5 to 0.3. How would you expect this to affect sensitivity and specificity, and where on the ROC curve does the operating point move?

c) K-nearest neighbours *(4 %)*

A classmate fits a K-NN classifier with $K = 35$ on the same training data, using the predictors above (standardized). He produces the following test-set confusion matrix:

	Predicted: default	**Predicted: no default**
Actual: default	200	400
Actual: no default	120	2280

(i) *(1 %)* Briefly describe how a K-NN classifier produces a predicted class label for a new observation. (Two or three sentences.)

(ii) *(2 %)* Compute the sensitivity, specificity, and test error rate of this KNN classifier.

(iii) *(1 %)* The classmate did not standardize `balance` (measured in thousands) together with `pay_0` (measured on a small integer scale). Why is this a problem for KNN, in particular?

d) A tree-based competitor *(5 %)*

You want to fit a tree-based method to compete with the logistic regression. Choose one method we discussed in the course, and justify your choice.

(i) *(2 %)* State which tree-based method you would use (e.g. bagging, random forest, gradient boosting) and give a *concrete numeric value* for each of its main tuning parameters (for a random forest, for instance: `mtry`, `ntree`, and the terminal-node size). **Justify each numeric choice** with one short sentence; a vague phrase like "sufficiently many trees" will not earn full credit.

(ii) *(2 %)* Suppose your chosen method gives test sensitivity 0.42, test specificity 0.96, and test error rate 0.14. Compare to the logistic regression results from part (a). Which model would you prefer, and why? Make explicit which criterion drives your choice.

(iii) *(1 %)* Briefly describe what a *variable-importance plot* from a random forest shows, and what kind of question it can (and cannot) answer.

e) Class imbalance *(2 %)*

The marginal distribution of `default` on the full dataset is approximately

$$P(\texttt{default} = 1) \approx 0.20, \qquad P(\texttt{default} = 0) \approx 0.80.$$

(i) *(1 %)* A naive classifier that always predicts "no default" would already achieve a test error rate of approximately what value? Justify in one sentence.

(ii) *(1 %)* In light of (i), comment on whether *test error rate* is the most appropriate metric for choosing between classifiers on this dataset. What metric or pair of metrics would you privilege instead, and why?

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