TMA4268 Statistical Learning V2026

Mock Exam 3 (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 below and pick the best word or phrase for each blank from the choices in parentheses. Each correct fill is worth 1 %.

In supervised statistical learning we typically distinguish between methods that produce an explicit functional form for the relationship between predictors and response, called _____ (1) (*generative* / *parametric* / *nonparametric* / *regularized*) methods, and methods that estimate the relationship more flexibly without assuming such a form, called _____ (2) (*parametric* / *nonparametric* / *linear* / *Bayesian*) methods.

A central concept of the course was that the expected squared test error of a fitted model can be written as the sum of three pieces: a squared bias, a variance, and an _____ (3) (*reducible* / *regularized* / *irreducible* / *validated*) component that does not vanish as we collect more training data. Methods such as ridge regression and lasso are referred to as _____ (4) (*boosting* / *generative* / *shrinkage* / *kernel*) methods because they shrink coefficients toward zero. Of these two, _____ (5) (*ridge* / *lasso* / *both* / *neither*) can produce a model in which some coefficients are exactly zero, and is therefore the natural choice when one also wants automatic variable selection.

When the goal is to estimate the prediction error of a fitted model honestly, we use a resampling technique. The simplest is the validation-set approach; a more reliable variant that splits the training data into k roughly equal pieces and rotates which piece serves as held-out data is called _____(6) (*the bootstrap / leave-one-out cross-validation / nested CV / k-fold cross-validation*). The lecturer in this course generally prefers this technique to penalty-based criteria such as AIC because the latter rely on _____(7) (*shrinkage / distributional assumptions / regularization / cross-validation*) that often fail in practice.

For non-linear regression we discussed methods that allow each predictor to enter the model through its own non-linear shape and are then summed; this family is called _____(8) (*principal components / convolutional networks / boosted trees / generalized additive models*).

Finally, in unsupervised learning we discussed two main families. One reduces the dimensionality of the data by finding orthogonal directions of maximal variance and is called _____(9) (*lasso / principal component analysis / partial least squares / neural networks*); another partitions the observations into groups, of which the agglomerative variant produces a tree-like structure called a _____(10) (*biplot / scree plot / dendrogram / decision tree*).

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 and double descent *(3 %)*

Mark each statement as true or false.

(i) As the flexibility of a fitted model is increased, its squared bias tends to decrease while its variance tends to increase.

(ii) The bias–variance decomposition $\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \text{Bias}^2 + \text{Var} + \sigma^2$ is an algebraic identity that holds for any fitted model $\hat{f}$.

(iii) In the over-parameterized regime where the number of model parameters p is much larger than n, a model that interpolates the training data has test error that grows without bound as p/n increases.

(iv) In a regression setting in which σ^2 is large, a more flexible model is generally preferable because it can chase down more of the noise.

b) Cross-validation variants *(3 %)*

Mark each statement as true or false.

(i) Leave-one-out cross-validation (LOOCV) generally has lower bias but higher variance than 5-fold cross-validation as an estimator of test error.

(ii) The validation-set approach is the same as 2-fold cross-validation.

(iii) For ordinary least squares, LOOCV can be computed without refitting the model n times by exploiting the diagonal of the hat matrix.

(iv) In a feature-selection-then-CV pipeline, it is permissible to use the response y to filter the predictors *once*, before any cross-validation, as long as the CV is then run on the resulting reduced predictor set.

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

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

- 6 input variables,
- one hidden layer of 8 neurons (with biases, ReLU activations),
- one hidden layer of 4 neurons (with biases, ReLU activations),
- one output layer of 2 neurons (with biases, softmax outputs).

(i) *(2 %)* How many parameters does this network have in total, *including all bias terms*? Show the layer-by-layer breakdown.

(ii) *(2 %)* A neuron in the *first* hidden layer has weights $w = (0.5, -1.0, 0.2, 1.0, -0.5, 0.3)$ and bias $b = -0.1$. Its inputs on a given observation are $x = (1, 0, 2, -1, 1, 3)$. Using a *sigmoid* activation $\sigma(z) = 1/(1 + e^{-z})$, compute the output of this neuron, rounded to three decimals.

d) Odds, log-odds, interaction *(3 %)*

(i) *(1 %)* A patient has probability $p = 0.05$ of an adverse event. What are the odds of that event?

(ii) *(1 %)* A model reports odds = 4 for a particular outcome. What is the corresponding probability?

(iii) *(1 %)* In a logistic regression model with the interaction term `bmi:smokerYes`, the fitted coefficients are $\hat{\beta}_{\texttt{bmi}} = 0.02$ and $\hat{\beta}_{\texttt{bmi:smokerYes}} = 0.10$. By what factor do the odds of the response event change for a *smoker* when `bmi` increases by one unit, holding all other predictors fixed? (One numeric value, two decimals.)

e) Splines and degrees of freedom *(3 %)*

Mark each statement as true or false.

(i) A cubic regression spline with K interior knots consumes $K + 4$ parameters (including the intercept).

(ii) A natural cubic spline with K interior knots consumes *fewer* parameters than a plain cubic spline with the same number of knots, because the natural spline is constrained to be linear outside the boundary knots.

(iii) For a cubic regression spline, the third derivative is continuous at every knot.

(iv) In a cubic regression spline the knot locations enter the fit non-linearly and are estimated jointly with the spline coefficients by least squares.

f) Principal component analysis *(5 %)*

You perform PCA on a dataset with five **standardized** variables $X_1, \ldots, X_5$ and obtain the following decomposition:

PC	Eigenvalue	PVE
PC1	2.1	0.42
PC2	1.4	0.28
PC3	0.8	0.16
PC4	0.5	0.10
PC5	0.2	0.04

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

$$\phi_1 = (\,0.55,\ 0.45,\ -0.40,\ 0.30,\ -0.50\,)^\top.$$

A new observation has standardized values $x^* = (\,1.5,\ -1,\ 0.5,\ 2,\ -0.5\,)^\top$.

(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 retained to capture at least 95% of the total variance?

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

(v) *(1 %)* Which one variable contributes *most strongly* to PC1, in absolute terms?

g) Bootstrap *(3 %)*

(i) *(1 %)* For a sample of size $n = 8$, what is the probability that a particular original observation is included in a given bootstrap sample? (Numeric, three decimals.)

(ii) *(1 %)* What does this probability tend to as $n \to \infty$? (Both an exact symbolic expression and a numeric approximation.)

(iii) *(1 %)* Mark each of the following as true or false. (A) Bootstrap samples are drawn from the original data *with* replacement; (B) The bootstrap can be used to correct for bias in the original estimator; (C) A single bootstrap sample is informative about the sampling distribution of an estimator.

h) Random forests and out-of-bag error *(2 %)*

Mark each statement as true or false.

(i) In a random forest, the standard default for the number of predictors sampled at each split is $m = \sqrt{p}$ for classification and $m = p/3$ for regression.

(ii) The number of trees B in a random forest is a tuning parameter that should be chosen by cross-validation.

(iii) The out-of-bag (OOB) error of a random forest provides an estimate of test error *without* needing a separate held-out set, because each observation is left out of approximately 1/3 of the bootstrap samples.

(iv) In a random forest, the individual trees are typically grown deep and *unpruned*; variance reduction is provided by averaging across trees rather than by per-tree pruning.

i) K-means clustering *(2 %)*

Mark each statement as true or false.

(i) The K-means objective is to minimize the sum of within-cluster squared Euclidean distances of every point to its cluster mean.

(ii) The number of clusters K is determined automatically by the K-means algorithm.

(iii) Because the K-means algorithm depends on a random initialization of the cluster centroids, it is recommended to run it several times from different initializations and pick the solution with the lowest within-cluster sum of squares.

(iv) The K-means algorithm produces a hierarchy of clusterings that can be cut at any level to obtain different numbers of clusters.

Problem 3 (16 %) — Theory and hand calculations

a) The mathy one — LDA decision boundary *(8 %)*

Consider a binary classification problem with two classes 0 and 1. Assume that the class-conditional densities are bivariate normal with class-specific means but a *shared* covariance matrix:

$$\mathbf{X} \mid Y = k \;\sim\; \mathcal{N}_2(\boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}), \qquad k = 0, 1,$$

with class priors π_0 and $\pi_1 = 1 - \pi_0$.

For the numerical part below, take

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}, \quad \pi_0 = \pi_1 = 0.5.$$

(i) *(2 %)* State, in one or two sentences each, the two key modelling assumptions of LDA. State also one practical reason why one might prefer LDA to QDA when the number of training observations per class is small.

(ii) *(3 %)* Starting from $\log(\pi_k f_k(\mathbf{x}))$, derive the LDA discriminant function

$$\delta_k(\mathbf{x}) \;=\; \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k \;-\; \tfrac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k \;+\; \log \pi_k.$$

Be *explicit* about which terms in the log-likelihood do and do not depend on k, and explain why $\delta_k(\mathbf{x})$ is *linear* in $\mathbf{x}$.

(iii) *(2 %)* Using the numerical values above, derive the equation of the LDA decision boundary in the form $ax_1 + bx_2 = c$. (You may use $\boldsymbol{\Sigma}^{-1} = \frac{1}{15} \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$.)

(iv) *(1 %)* If you instead allowed the two classes to have *different* covariance matrices $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$, the resulting decision boundary would no longer be linear. What *shape* does it take, and what is the name of this method?

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

Four observations have the following Euclidean dissimilarity matrix:

$$D = \begin{pmatrix} 0 & 3 & 7 & 8 \\ 3 & 0 & 5 & 6 \\ 7 & 5 & 0 & 4 \\ 8 & 6 & 4 & 0 \end{pmatrix}.$$

(i) *(3 %)* Sketch the dendrogram produced by hierarchical clustering with *single* 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 you instead used *complete* linkage on the same matrix, the order of the merges would be unchanged, but the height of the *final* fusion (when all four observations are joined) would differ. Compute the final fusion height under complete linkage and explain in one sentence why it differs from the single-linkage answer.

(iii) *(1 %)* Suppose you cut the single-linkage dendrogram at height 4.5. How many clusters do you obtain, and which observations are in each cluster?

c) The bias–variance decomposition and double descent *(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. 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 randomness the expectations are taken over.*

(ii) *(1 %)* Below are two stylized profiles of test MSE as you increase the flexibility of the fitted model (e.g. the degree of a polynomial regression).

- **Profile A:** test MSE drops, hits a single minimum, and then rises monotonically (the classical U-shape).
- **Profile B:** test MSE drops, then climbs to a peak around the point where the number of fitted parameters equals the training set size, and then *descends a second time* to a value below the original minimum.

Identify Profile B by its standard name, and explain in one or two sentences why both profiles are consistent with the same exact bias–variance decomposition. (Hint: think about what the optimization is implicitly doing once there are infinitely many zero-training-error solutions.)

Problem 4 (22 %) — Data analysis: insurance premiums

A health insurer records $n = 900$ customers' annual insurance premium `charges` (in 1000 EUR). The available predictors are:

- `age` — age in years (continuous, 18–65);
- `bmi` — body-mass index, kg/m^2 (continuous, 16–50);
- `children` — number of dependent children (count, 0–5);
- `smoker` — categorical: *no* (reference), *yes*;
- `sex` — categorical: *female* (reference), *male*;
- `region` — categorical, 4 levels: *northeast* (reference), *northwest*, *southeast*, *southwest*.

The data are split 600/300 into a training set and a test set.

a) Linear regression with interaction *(7 %)*

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

$$\texttt{charges} \sim \texttt{age} + I(\texttt{age}^2) + \texttt{bmi} + \texttt{smoker} + \texttt{bmi:smoker} + \texttt{sex} + \texttt{region} + \texttt{children},$$

where `bmi:smoker` denotes the (continuous × binary) interaction between `bmi` and `smoker`. The fitted output is:

	Estimate	**Std. Error**	**t-value**	**Pr(> |t|)**
(Intercept)	−2.50	1.80	−1.39	0.165
`age`	−0.05	0.10	−0.50	0.617
$I(\texttt{age}^2)$	0.0040	0.0010	4.00	< 0.001
`bmi`	0.10	0.05	2.00	0.046
`smoker_yes`	−22.50	5.00	−4.50	< 0.001
`bmi:smoker_yes`	1.45	0.18	8.06	< 0.001
`sex_male`	0.10	0.21	0.48	0.633
`region_NW`	−0.45	0.30	−1.50	0.134
`region_SE`	0.18	0.30	0.60	0.549
`region_SW`	−0.62	0.30	−2.07	0.039
`children`	0.55	0.10	5.50	< 0.001

Multiple $R^2 = 0.78$, Adjusted $R^2 = 0.776$. Residual standard error: 4.10 on 589 d.f.

(i) *(1 %)* How many parameters does this model estimate, including the intercept?

(ii) *(2 %)* For a *non-smoker*, by how much (in EUR) does an increase of one BMI unit change the predicted annual charges, holding all other predictors fixed? Repeat the calculation for a *smoker*. (Two numeric answers; remember that `charges` is in 1000 EUR.)

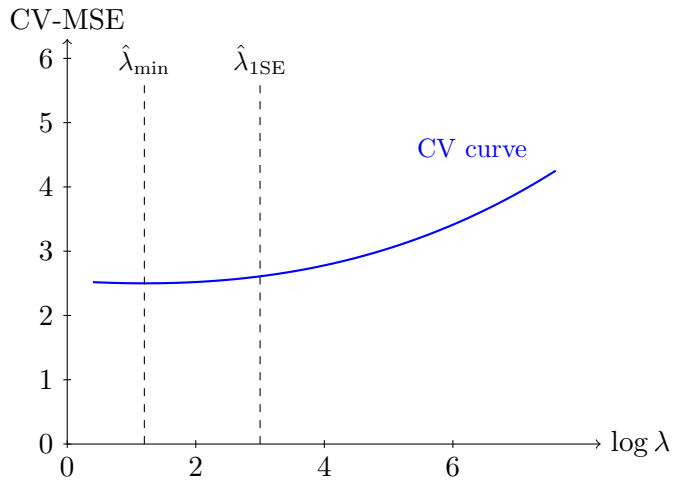
(iii) *(2 %)* Briefly explain why the main-effect coefficient on `smoker_yes` alone (here −22.50) is *not* a useful summary of the average effect of being a smoker on premiums. (Mention the role of the interaction.)

(iv) *(1 %)* An analyst wants to test whether `region` is jointly relevant after controlling for the other predictors. State **which test** you would use and write the null hypothesis. (You do *not* need to compute the test statistic by hand.)

(v) *(1 %)* The model reports $R^2 = 0.78$ and adjusted $R^2 = 0.776$. Briefly comment on what the closeness of these two numbers suggests.

b) Ridge regression with 10-fold cross-validation *(5 %)*

The same predictors are now fed (after standardization) into a ridge regression, with the penalty parameter λ chosen by 10-fold cross-validation. The CV-MSE profile is:



(You may take, on the same test set as in part (a): test MSE for OLS = 38.0; test MSE for ridge at $\hat{\lambda}_{\min} = 38.4$; test MSE for ridge at $\hat{\lambda}_{1SE} = 39.1$.)

(i) *(1 %)* Why is it important to standardize the predictors *before* fitting ridge regression? Give a one-sentence explanation.

(ii) *(1 %)* Write down the ridge regression objective function, and explain (in words) what happens to $\hat{\beta}^R_\lambda$ as $\lambda \to 0$ and as $\lambda \to \infty$.

(iii) *(2 %)* State the *one-standard-error rule* precisely, and give one reason a practitioner might prefer $\hat{\lambda}_{1SE}$ over $\hat{\lambda}_{\min}$.

(iv) *(1 %)* On this dataset, ridge at $\hat{\lambda}_{\min}$ gives test MSE = 38.4 while OLS gives 38.0. Briefly interpret this comparison in bias–variance terms: what does it suggest about the predictors in this dataset?

c) GAM with splines *(5 %)*

A generalized additive model is fit on the same training set:

$$\texttt{charges} = \beta_0 + s(\texttt{age}, \text{df} = 4) + \text{bs}(\texttt{bmi}, \text{knots} = \{25, 30, 35\}) + \beta_{\texttt{smoker}} \cdot \texttt{smoker} + g(\texttt{region}) + \beta_{\texttt{kids}} \cdot \texttt{children}$$

where $s(\cdot, \text{df} = 4)$ denotes a smoothing spline with 4 effective degrees of freedom (intercept removed), $\text{bs}(\cdot)$ denotes a cubic regression spline (intercept removed), and $g(\texttt{region})$ uses dummy coding as in part (a).

(i) *(2 %)* How many degrees of freedom does this GAM consume in total (including the global intercept)? Give a brief term-by-term breakdown.

(ii) *(1 %)* What is the structural assumption that the GAM makes about the relationship between, say, `age` and `bmi`, and in what sense is this a *limitation* compared to (for example) a regression tree?

(iii) *(2 %)* Suppose the GAM achieves test MSE = 32.5 on the same test set as part (a)/(b). Compare this to the OLS result of 38.0. What does the gap suggest about the underlying relationship between predictors and `charges`, and which predictor would you suspect contributes most to the improvement (briefly justify by reference to the model output in part (a))?

d) Tree boosting *(5 %)*

Finally a gradient-boosted regression tree is fit on the same training data; it achieves test MSE = 26.8.

(i) *(2 %)* Gradient boosting has three main hyperparameters: the number of trees B, the interaction depth d (per-tree size), and the shrinkage / learning rate ν. For each, briefly say how you would choose its value and how it qualitatively affects the bias–variance behaviour of the resulting ensemble.

(ii) *(1 %)* For a *random forest*, the number of trees B is *not* a tuning parameter and one simply uses "enough" of them. For *boosting*, B *is* a real tuning parameter. Briefly explain why this difference exists.

(iii) *(2 %)* Comparing the test MSEs — boosting 26.8, GAM 32.5, OLS 38.0 — what does the ranking suggest about the structure of the relationship between predictors and `charges`? If you also wanted *interpretation* (which predictors matter, and roughly in which direction), what additional diagnostic would you compute from the boosted model?

Problem 5 (24 %) — Data analysis: stroke risk classification

A teaching hospital records $n = 4000$ patients followed for five years; the binary response `stroke` is 1 if the patient suffered a stroke during follow-up and 0 otherwise. The available predictors are:

- `age` — age in years (continuous);
- `avg_glucose` — average blood glucose (continuous, mg/dL);
- `bmi` — body-mass index (continuous);
- `hypertension` — binary 0/1 indicator;
- `heart_disease` — binary 0/1 indicator;
- `smoking_status` — categorical, 4 levels: *never* (reference), *former*, *current*, *unknown*.

The data are split 70/30 into training (2800) and test (1200) sets. Among the test set there are 120 true strokes and 1080 true non-strokes.

a) Logistic regression *(8 %)*

A logistic regression model is fit on the training set with the predictors above (no interactions). The fitted output is:

	Estimate	**Std. Error**	**z-value**	**Pr($>\lvert z\rvert$)**
(Intercept)	-8.50	0.50	-17.00	< 0.001
`age`	0.073	0.005	14.60	< 0.001
`avg_glucose`	0.005	0.0008	6.25	< 0.001
`bmi`	-0.005	0.012	-0.42	0.674
`hypertension`	0.45	0.12	3.75	< 0.001
`heart_disease`	0.50	0.13	3.85	< 0.001
`smoking_former`	0.18	0.15	1.20	0.230
`smoking_current`	0.32	0.16	2.00	0.046
`smoking_unknown`	0.05	0.18	0.28	0.781

(i) *(2 %)* By what factor do the odds of stroke multiply for each one-year increase in `age`, holding all other predictors fixed? (Numeric answer, to three decimals.) State briefly in words how the magnitude should be interpreted.

(ii) *(3 %)* Consider a patient with the following profile: `age` = 68, `avg_glucose` = 180, `bmi` = 28, `hypertension` = 1, `heart_disease` = 0, `smoking_status` = *current*. Compute the predicted probability $\hat{p}$ of stroke for this patient. Show your work, including the linear predictor $\hat{\eta}$ and the sigmoid step.

(iii) *(1 %)* At a default classification threshold of $\hat{p} = 0.5$, would this patient be predicted to have a stroke? Justify in one sentence.

(iv) *(2 %)* Define, *in words appropriate to this specific problem*, what *sensitivity* and *specificity* mean for the stroke classifier. Briefly comment on what it would mean clinically if a screening tool had high sensitivity but low specificity.

b) LDA vs. QDA *(6 %)*

LDA and QDA are also fit on the same training set, using the same predictors as in part (a). On the test set they produce the following confusion matrices.

LDA

	Predicted: stroke	**Predicted: no stroke**
Actual: stroke	50	70
Actual: no stroke	100	980

QDA

	Predicted: stroke	**Predicted: no stroke**
Actual: stroke	75	45
Actual: no stroke	250	830

(i) *(2 %)* Compute the sensitivity, specificity, and overall test error rate of the LDA classifier.

(ii) *(2 %)* Compute the same three quantities for the QDA classifier.

(iii) *(2 %)* Briefly compare the two classifiers. State which one you would prefer and *on which criterion*, justifying your answer with reference to (a) the modelling difference between LDA and QDA, and (b) the medical context (cost of a false negative vs. a false positive).

c) A tree-based competitor: random forest *(5 %)*

You also fit a *random forest* with $B = 500$ trees and `mtry` $= 3$ on the training data. On the test set it produces:

	Predicted: stroke	**Predicted: no stroke**
Actual: stroke	65	55
Actual: no stroke	130	950

(i) *(1 %)* Justify briefly the choice `mtry` $= 3$. (Recall: there are 6 predictors at the tree-fitting stage.)

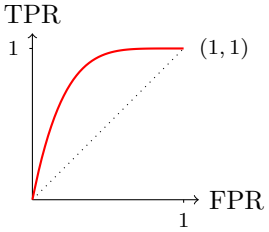
(ii) *(1 %)* Compute the test sensitivity, specificity, and error rate.

(iii) *(1 %)* Briefly explain what the *out-of-bag (OOB) error* is and why it is a useful diagnostic for a random forest.

(iv) *(2 %)* The variable-importance plot from this forest shows `age` and `avg_glucose` as having far higher importance than the other predictors. State (a) what a variable-importance plot from a random forest *shows*, and (b) what kind of question it *cannot* answer.

d) ROC and AUC *(3 %)*

A schematic ROC curve for the logistic regression of part (a), together with the diagonal of chance, is sketched below.



The model achieves AUC $= 0.84$.

(i) *(1 %)* What does AUC = 0.84 *mean* in plain English (i.e. beyond "the area under the ROC curve")? Frame your answer in terms of randomly chosen positive and negative cases.

(ii) *(1 %)* A clinician argues that the default threshold of $\hat{p} = 0.5$ misses too many true stroke cases. If the threshold is lowered from 0.5 to 0.3, in which direction will sensitivity and specificity move?

(iii) *(1 %)* On the ROC curve sketched above, in which direction does the operating point move when the threshold is lowered?

e) Class imbalance *(2 %)*

In the full dataset, only about 5% of patients suffer a stroke during the five-year follow-up.

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

(ii) *(1 %)* In light of (i), comment on whether the 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.