

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK-IN4300/9300 — Statistical Learning
Methods in Data Science

Day of examination: Wednesday, December 10, 2025

Examination hours: 9.00–13.00

This problem set consists of 4 pages.

Appendices: None.

Permitted aids: None.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 Bias-Variance Tradeoff

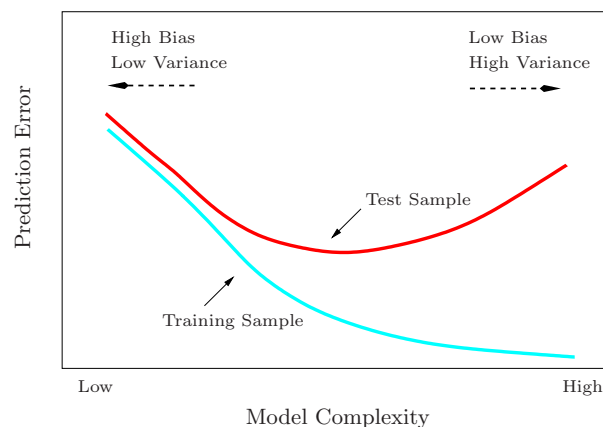


Figure 1: From Hastie et al. (2009). *The Elements of Statistical Learning*.

- (a) Based on Figure 1, explain the bias-variance tradeoff.
- (b) The (cubic) smoothing spline is the minimizer $\hat{f}(x)$ of the criterion

$$\text{RSS}_\lambda(f) = \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$$

over all functions with two continuous derivatives. Now:

- (i) Explain how λ controls the bias-variance tradeoff.
- (ii) What happens when $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$?
- (c) k -nearest neighbors (KNN) is a fundamental non-parametric supervised learning method. Now:

(Continued on page 2.)

- (i) Briefly explain how KNN works, both in the context of regression and classification.
- (ii) Explain how the bias-variance tradeoff can be controlled when using KNN.

Problem 2 Linear Regression

Assume that we have a linear model $Y = \beta^T x + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and x is an input vector of length $p + 1$ (including the constant). Furthermore, assume we have training data in the form of N input vectors, collected into an $N \times (p + 1)$ matrix $\mathbf{X}$, and an associated outcome vector $\mathbf{y}$ of length N . Assuming $\mathbf{X}$ has full rank, the least squares estimate of the parameters is uniquely given by

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- (a) Let $\hat{\theta} = \hat{\beta}^T x_0$ be the least squares estimate of the true conditional mean $\theta = \beta^T x_0$ at target point x_0 . Considering $\mathbf{X}$ to be fixed, show that:
 - (i) $\hat{\theta}$ is a linear function of $\mathbf{y}$ (and hence a linear estimator).
 - (ii) $\hat{\theta}$ is an unbiased estimator for θ , that is, $\mathbb{E}(\hat{\theta}) = \theta$.
- (b) Let $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ be any linear unbiased estimator for θ (including $\hat{\theta}$). The expected prediction error (EPE) at x_0 can under the given assumptions be decomposed as

$$EPE(x_0) = \sigma_\epsilon^2 + (\theta - \mathbb{E}[\tilde{\theta}])^2 + \mathbb{E}[(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])^2]$$

Now, under the given assumptions, the Gauss-Markov theorem states that $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta})$. In terms of the EPE at x_0 :

- (i) Explain why the Gauss-Markov theorem favours the least squares estimator over any other linear unbiased estimator.
- (ii) Explain why one still might be better off using another estimator.
- (c) Consider three methods for restricting the linear regression model: best subset selection, ridge regression and the lasso. In the special case of an orthonormal input matrix $\mathbf{X}$ (for which $\mathbf{X}^T \mathbf{X} = \mathbf{I}$), the three methods have explicit solutions in the form of a simple transformation of the least squares estimate, as shown in Figure 2. Based on Figure 2, explain how each method regularises the full least squares model.

Problem 3 Effective Number of Parameters

Let $\{(x_i, y_i)\}_{i=1}^N$ be a set of N input-output training observations. Furthermore, let $\mathbf{y} = (y_1, \dots, y_N)$ denote the N -vector of output values and $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_N)$ the corresponding predictions. Then, a linear fitting method is one for which we can write

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y},$$

(Continued on page 3.)

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\hat{\beta}_j \geq \hat{\beta}_{(M)})$
Ridge	$\hat{\beta}_j / (1 + \lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\hat{\beta}_j - \lambda)_+$

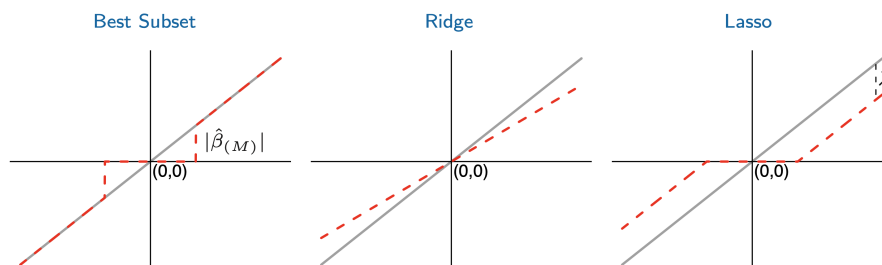


Figure 2: From Hastie et al. (2009). *The Elements of Statistical Learning*. Note: The least squares estimate is on the x-axis and the restricted estimate on the y-axis.

where $\mathbf{S}$ is an $N \times N$ matrix that depends on the inputs x_i but not on the outputs y_i . For a linear fitting method we can define the *effective number of parameters* as

$$\text{df}(\mathbf{S}) = \text{trace}(\mathbf{S}),$$

where the trace operator is the sum of the diagonal elements.

- Show that k -nearest neighbors regression is a linear fitting method.
- Derive an expression for the effective number of parameters for k -nearest neighbors regression and evaluate the expression for $k = 1$ and $k = N$.

Problem 4 Model Selection and Evaluation

- Let $\mathcal{T}$ denote a set of training data containing N joint observations of the input-output pair (X, Y) and let (X^0, Y^0) denote a new test data pair (all generated from the same $p(X, Y)$). Further let $\hat{f}(X)$ be a model fitted on the training data. Given some loss function $L(y, f(x))$, consider the following definitions of errors for the fitted model:

$$(i) \text{ Err} = \mathbb{E}_{\mathcal{T}} [\mathbb{E}_{X^0, Y^0} [L(Y^0, \hat{f}(X^0)) | \mathcal{T}]]$$

$$(ii) \text{ Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{Y^0} [L(Y_i^0, \hat{f}(x_i)) | \mathcal{T}]$$

Explain what errors (i)–(ii) measure and describe some method for estimating each of them.

- Consider a linear regression model $y = \beta^T x$, where β is a $p + 1$ parameter vector that includes the constant. Let $\hat{\beta}$ be the least squares estimate based on training observations $(x_1, y_1), \dots, (x_N, y_N)$, drawn independently from some distribution $p(X, Y)$. Let $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_M, \tilde{y}_M)$

(Continued on page 4.)

be a set of test observations, drawn independently from the same distribution as the training data. Now, prove that

$$\mathbb{E}_{X,Y}[R_{tr}(\hat{\beta})] \leq \mathbb{E}_{X,Y}[R_{te}(\hat{\beta})],$$

where $R_{tr}(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = \frac{1}{M} \sum_{i=1}^M (\tilde{y}_i - \beta^T \tilde{x}_i)^2$.

Problem 5 Gradient Tree Boosting

A gradient tree boosting algorithm for regression given a generic loss function $L(y, f(x))$ is shown below in Algorithm 10.3.

Algorithm 10.3 Gradient Tree Boosting Algorithm.

1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$.

2. For $m = 1$ to M :

(a) For $i = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm} , $j = 1, 2, \dots, J_m$.

(c) For $j = 1, 2, \dots, J_m$ compute

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.

(a) Explain steps 1 – 3 of the gradient tree boosting algorithm as defined in Algorithm 10.3.

(b) Assume that we have a global J , that is, $J_m = J$ for $m = 1, \dots, M$. Now:

(i) Explain how the parameters M and J are used to to regularize the fitted model.

(ii) Explain the implications on the obtained model when setting $J = 1$.

THE END - GOOD LUCK!