Advanced Statistical Learning 2025
Quiz Solution

May 27, 2025

Last Name: _____

First Name: _____

Matriculation Number (Student ID): _____

For every statement, mark if the statement is true (Yes) or false (No) with a cross like in the example below.

There may be several true statements for each topic. The topics are just for orientation.

Every statement marked correctly gives +1 point.

Every statement marked incorrectly gives −1 point.

Every unmarked or ambiguously marked statement gives 0 points.

Example

Yes	No	
☒	☐	True statement 1
☒	☐	True statement 2
☐	☒	False statement

General statistical learning

Yes No

☐ ☐ In supervised learning, the response y of the training data is not known. NO, the response is known in supervised learning.

☐ ☐ In statistical learning, the probability distribution on $\mathcal{X} \times \mathcal{Y}$ that we assume generates the data is typically not known. YES

☐ ☐ A loss function quantifies the difference between true and predicted response values. YES

☐ ☐ To find a good model, we maximize the empirical risk induced by a loss function. NO, we minimize it.

☐ ☐ The training error is a good estimate for the generalization error. NO, it is usually too optimistic because a model is fitted to minimize the training error.

☐ ☐ To assess the generalization performance of a model, we always need to use cross validation. NO, we can split the data into training and test data if we have enough observations.

☐ ☐ $\text{VC}_p(\mathcal{H}) = 3$ means that the hypothesis space $\mathcal{H}$ contains only three hypotheses h. NO, 3 is the largest number of points that can be shattered by members of $\mathcal{H}$.

☐ ☐ The No Free Lunch theorem tells us that no machine learning algorithm is universally better than any other over all possible problems. YES

☐ ☐ Gradient descent can be used for model estimation. Meant to be YES, but one could argue NO if regarding estimation and optimization as separate things.

☐ ☐ Maximum likelihood estimation and empirical risk minimization are equivalent in certain cases. YES

Regression

Yes No

☐ ☐ We always need numerical optimization methods to estimate regression models. NO, there is a closed form solution for e.g. the classical linear model with the quadratic loss function.

☐ ☐ For a regression task, the optimal constant that minimizes the empirical risk induced by the absolute loss function is the median of the response values. YES

☐ ☐ The quantile/pinball loss can weigh positive and negative residuals differently. YES

☐ ☐ The Huber loss combines advantages of the quadratic loss and the absolute loss. YES, differentiability and robustness

☐ ☐ The function $L(y, f(x)) = (y - f(x))^3$ is a valid loss function for regression. NO, a loss function has to be non-negative.

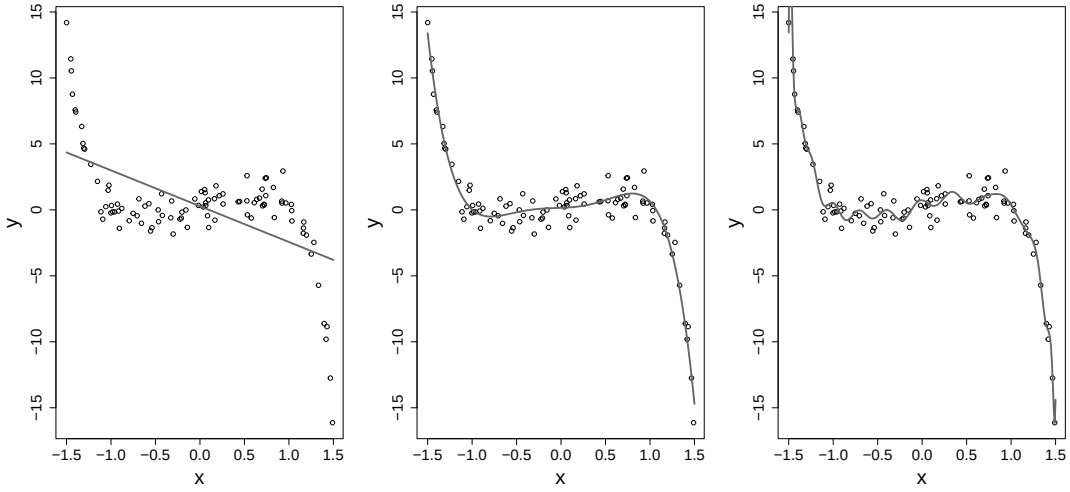
Classification

Yes No

- ☐ ☐ The Bernoulli loss is a suitable loss function for a classification task with a response $y \in \{0, 1\}$. YES
- ☐ ☐ When f is a scoring classifier, $|f(x)|$ is called confidence. YES
- ☐ ☐ In binary classification, probabilities are transformed into class labels via thresholding. YES
- ☐ ☐ In multiclass classification, the softmax function is used to transform class labels into scores. NO, it is used to transform scores into probabilities.
- ☐ ☐ A decision boundary in the multiclass case is a set of points from $\mathcal{X}$ where two scoring functions are equal and larger than or equal to all other scoring functions. YES, this is just the mathematical definition from the lecture put into words.

Regression example

The graphs show a regression data set and three different polynomial models fitted to it.

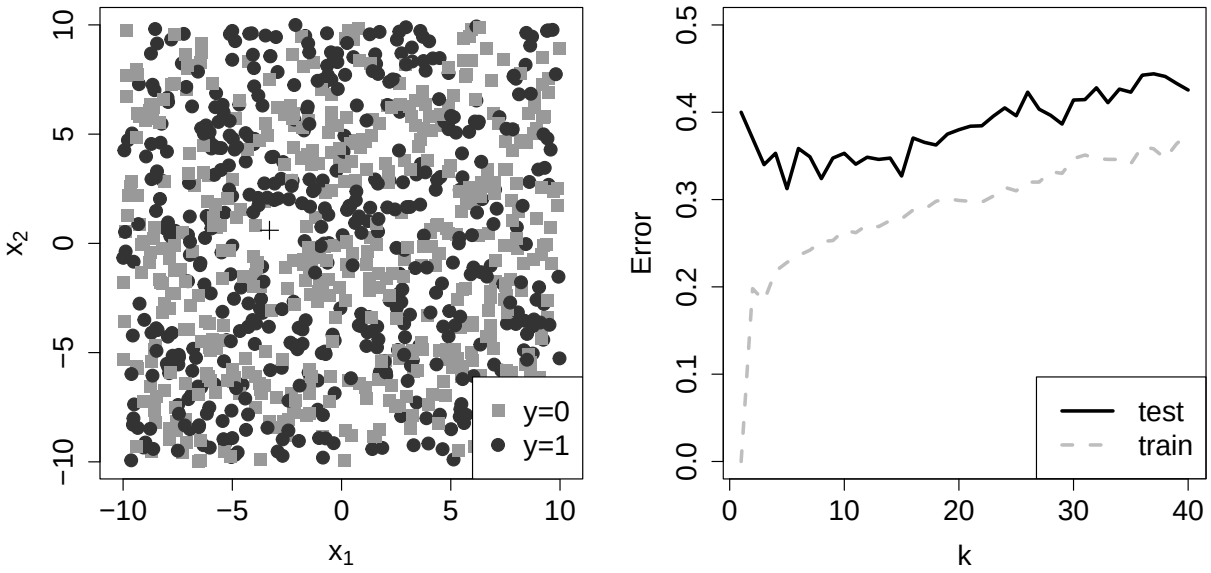


Yes No

- ☐ ☐ The models are from the hypothesis spaces $\mathcal{H}_i = \{f(x) = \theta x^i \,|\, \theta \in \mathbb{R}\}$. NO, these spaces only include single terms of degree i, not full polynomials.
- ☐ ☐ If the order of the fitted polynomial is chosen too large, this leads to underfitting. NO, this might lead to overfitting.
- ☐ ☐ The model on the left has high bias and low variance. YES
- ☐ ☐ The model on the right has a higher capacity than the other two. YES, it is the wiggliest.
- ☐ ☐ If we want to fit a piecewise linear regression model to the data, increasing the number of knots increases the flexibility of the model. YES

Classification example

The left graph shows a binary classification data set of size n with two covariates. The right graph shows estimates for the training error and the test error when applying the k-nearest neighbors (KNN) method with different values of k.



Yes No

☐ ☐ For $k = 1$, KNN predicts class 1 for the new observation marked by +. NO, its nearest neighbor is from class 0.

☐ ☐ If k is chosen too large, this leads to underfitting. YES

☐ ☐ The training error of the k-nearest neighbors method for $k = 1$ is always 0 by construction. YES

☐ ☐ The test error graph suggests that $k = 37$ might be a good choice. NO, we want a small test error.

☐ ☐ For $k = n$, the KNN model is very flexible. NO, that's a constant model.