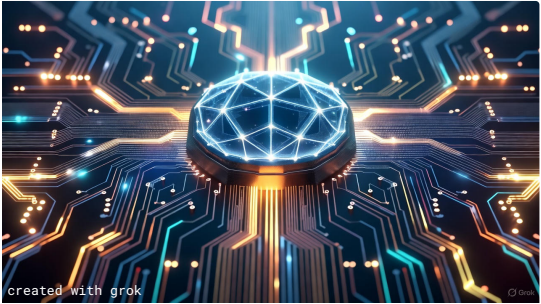


Advanced Machine Learning

Silvan Stadelmann - 28. Oktober 2025 - v0.0.1

github.com/silvasta/summary-aml



Contents

Representation	2
1 Learning objectives	2
2 Expected risk	2
3 Empirical risk	2
4 Empirical test error and expected risk	2
5 Comparing algorithm performance on test data	2
6 Data	2
6.1 Feature space	2
6.2 Example of Data	2
7 Mathematical Spaces	2
Regression	2
8 Linear Regression	2
9 Gauss Markov Theorem	3
10 Bias/Variance Dilemma	3
11 Bayesian Maximum A Posteriori (MAP) estimates	3
11.1 Ridge Regression	3
11.2 LASSO	3
11.3 Ridge vs. LASSO Estimation	3
12 Remarks on Shrinkage Methods	3
Support Vector Machines	3
Neural Networks	3
Transformer	3
Exercises	3
13 Problem 1 - Regression	3

Representation

1 Learning objectives

Estimation of Dependences Based on Empirical Data

What is the learning problem?

$$y = f_{\theta}(x) + \eta \quad \text{with} \quad \nu \sim P(\eta|0, \sigma^2)$$

2 Expected risk

- Conditional expected risk

- Total expected risk

3 Empirical risk

- Test and Train Data

Test data cannot be used before the final estimator has been selected!

Training error $\hat{R}(\hat{f}_n, \mathcal{Z}^{\text{train}})$ for Empirical Risk Minimizer (ERM)

4 Empirical test error and expected risk

Distinguish

5 Comparing algorithm performance on test data

6 Data

6.1 Feature space

- Measurement space \mathcal{X}

+ numerical $\mathcal{X} \subset \mathbb{R}^d$

+ boolean $\mathcal{X} = \mathbb{B}$

+ categorical $\mathcal{X} = \{1, \dots, k\}$

Features are derived quantities or indirect observations which often significantly compress the information content of measurements.

Remark The selection of a specific feature space predetermines the metric to compare data; this choice is the first significant design decision in a machine learning system.

Taxonomy of Data

6.2 Example of Data

- monadic data

- dyadic data

- pairwise data

- polyadic data

7 Mathematical Spaces

- Topological spaces

- Metric space

- Euclidean vector spaces

- Probability Spaces

Regression

8 Linear Regression

- Statistical model

$$Y = X^T \beta. \quad Y \in \mathbb{R}. \quad X, \beta \in \mathbb{R}^{d+1}$$

- Residual Sum of Squares (RSS)

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

9 Gauss Markov Theorem

10 Bias/Variance Dilemma

- Tradeoff, split Error
- Identify error components

11 Bayesian Maximum A Posteriori (MAP) estimates

11.1 Ridge Regression

- Cost function
- Bayesian view
- Solution

Tikhonov regularization

11.2 LASSO

- Cost function
- Bayesian view
- Solution

11.3 Ridge vs. LASSO Estimation

12 Remarks on Shrinkage Methods

- Generalized Ridge Regression

Idea behind shrinkage When white noise is added to the data then all Fourier coefficients are increased by a constant on average. ☒ Shrink all coefficients by the estimated noise amount to derive a robust predictor.

Support Vector Machines

Neural Networks

Transformer

Exercises

13 Problem 1 - Regression

- Linear Regression
- Ridge Regression
- Noisy Regression

E1.2.c - An Engineer's rule of thumb is to choose K as $\min \sqrt{n}, 10$

- Overfitting
- Cross Validation
- Generative vs. Discriminative Modeling