# 5705.25 - Hagfrøðilig læring H25

October 27, 2025

## Contents

# 1 Introduction

# 2 Statistical Learning

Statistical learning is concerned with understanding and estimating the relationship between a target variable $Y$ and one or more predictors $X = (X_1, X_2, \ldots, X_p)$. We typically assume the model:

$$Y = f(X) + \epsilon$$

where $f(X)$ is the systematic part of the relationship, and $\epsilon$ is random noise with mean zero.

## 2.1 Supervised Learning

In supervised learning, both the predictors $X$ and the response $Y$ are observed, and the goal is to estimate the mapping $f : X \mapsto Y$.

### 2.1.1 Types of Supervised Learning

Depending on the type of $Y$, we distinguish between:

| Type of $Y$ | Task | Example |
|---|---|---|
| Continuous (quantitative) | Regression | Predicting sales, temperature, price |
| Discrete (qualitative) | Classification | Predicting spam vs. non-spam, disease vs. healthy |

### 2.1.2 Example: Predicting Sales

Suppose we want to predict sales $Y$ using advertising budgets in three media channels:

$$X_1 = \text{TV}, \quad X_2 = \text{Radio}, \quad X_3 = \text{Newspaper}.$$

We assume

$$Y = f(X_1, X_2, X_3) + \epsilon$$

and start with a linear model:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

### 2.1.3 Estimating $f$ Using Squared Loss

To estimate $f$, we minimize the squared loss:

$$L(y_i, \hat{f}(x_i)) = (y_i - \hat{f}(x_i))^2,$$

and the total (mean) squared error:

$$\text{MSE}(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2.$$

The optimal parameters $\beta_j$ are found by minimizing the MSE, leading to the ordinary least squares (OLS) solution:

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

## 2.2 Unsupervised Learning

In unsupervised learning, only the predictors $X$ are observed; there is no corresponding response variable $Y$. The goal is to discover structure or patterns within the data.

### 2.2.1 Clustering

A common example is **clustering**, where observations are grouped based on similarity among predictors $X_1, X_2, \ldots$. For instance, clustering points $(x_1, x_2)$ may reveal natural groupings or clusters within a dataset.

### 2.2.2 Outcomes

In contrast to supervised learning:

- Continuous outcomes correspond to **regression**.
- Discrete outcomes correspond to **classification**.

## 2.3 Least Squares Regression

In least squares regression, we assume a model such as:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2,$$

which can capture some curvature or nonlinearity.

While more flexible models can better fit training data, they also introduce new issues such as **overfitting**. We can measure performance using the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2.$$

The error can be decomposed into:

- **Training error**: computed on the same data used to fit the model.
- **Test (validation) error**: computed on unseen data.

A model that fits the training data too closely may not generalize well to new data this is overfitting. The aim is to find a balance between underfitting and overfitting.

As a general rule of thumb, more flexible methods tend to perform better than inflexible ones, but they must be validated using separate **validation** and **test** sets.

## 2.4 Bias–Variance Tradeoff

The expected prediction error for a given $x_0$ can be decomposed as:

$$\mathbb{E}\big[(Y - \hat{f}(x_0))^2\big] = \big[\text{Bias}(\hat{f}(x_0))\big]^2 + \text{Var}(\hat{f}(x_0)) + \sigma_\epsilon^2,$$

where:

- $\sigma_\epsilon^2 = \text{Var}(\epsilon)$ is the irreducible error,
- $\text{Bias}(\hat{f}(x_0)) = \mathbb{E}[\hat{f}(x_0)] - f(x_0)$,
- $\text{Var}(\hat{f}(x_0))$ is the variability of the estimate.

We cannot reduce the irreducible error, but we can trade off bias and variance. Accuracy depends on finding the right balance:

- High bias $\Rightarrow$ underfitting.
- High variance $\Rightarrow$ overfitting.

Interpretability often decreases as model flexibility increases.

## 2.5 Classification

In classification, the goal is to predict a qualitative response $Y$ taking values in a finite set of classes $\{1, 2, \ldots, K\}$.

### 2.5.1 K-Nearest Neighbours (KNN)

Given a new observation $x_0$, KNN identifies the $K$ training points closest to $x_0$ (denoted $\mathcal{N}_0$) and estimates:

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(Y_i = j),$$

where $I(\cdot)$ is the indicator function.

The predicted class is:

$$\hat{y}_0 = \arg\max_j \Pr(Y = j \mid X = x_0).$$

### 2.5.2 Classification Error

The classification error rate is:

$$\text{Error} = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(Y_i \neq \hat{Y}_i).$$

### 2.5.3 Bayes Classifier

The Bayes classifier assigns each observation to the most probable class:

$$\hat{y}_{\text{Bayes}} = \arg\max_j \Pr(Y = j \mid X = x),$$

which theoretically minimizes the classification error. In practice, methods like KNN approximate this boundary, sometimes sacrificing perfect classification to reduce noise.

### 2.5.4 Key Idea

A fundamental rule to remember is:

$$Y = f(X) + \epsilon.$$

This holds true for both regression and classification settings the goal of statistical learning is to estimate $f$ as accurately as possible.

# 3 Linear Regression