Chapter ML:I

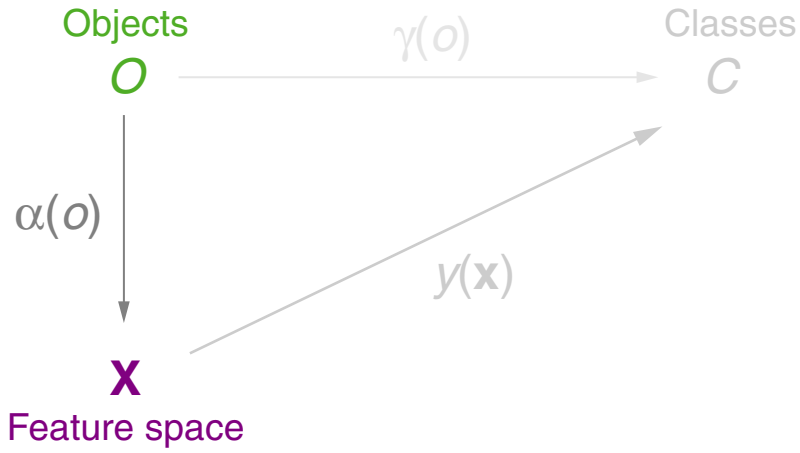
I. Introduction

- Examples of Learning Tasks
- Specification of Learning Tasks
- Elements of Machine Learning
- Notation Overview
- Classification Approaches Overview

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(1) Model Formation: Real World → Model World



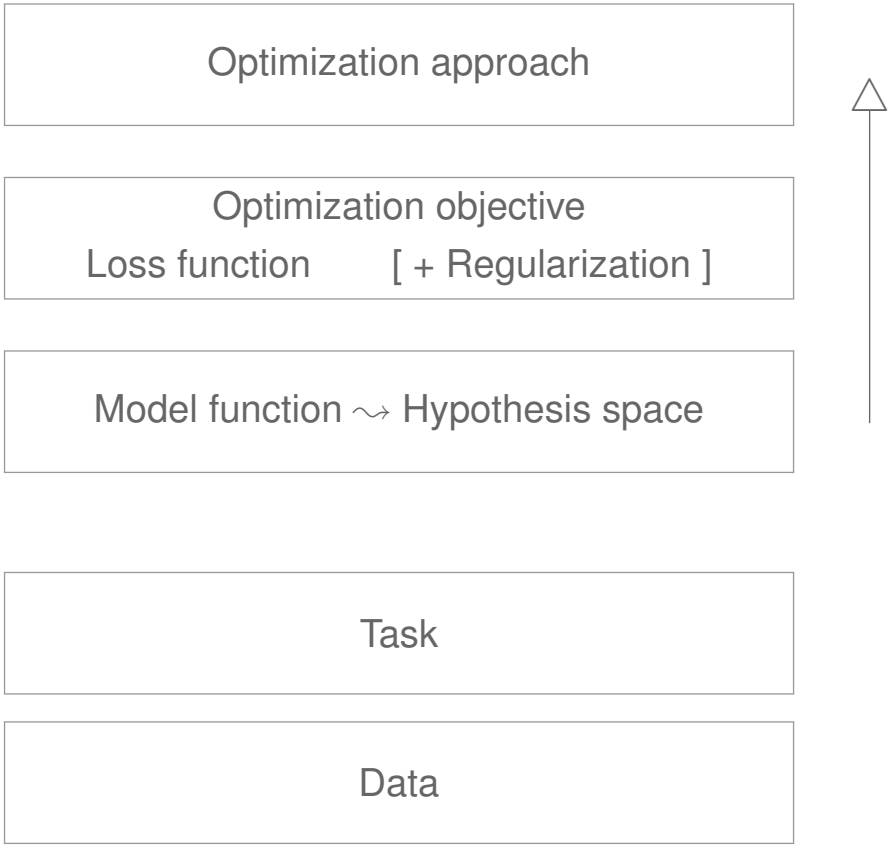
Related questions:

- From what kind of experience should be learned?
- Which level of fidelity is sufficient to solve a certain task?

©STEIN/VÖLSKE 2026

Elements of Machine Learning

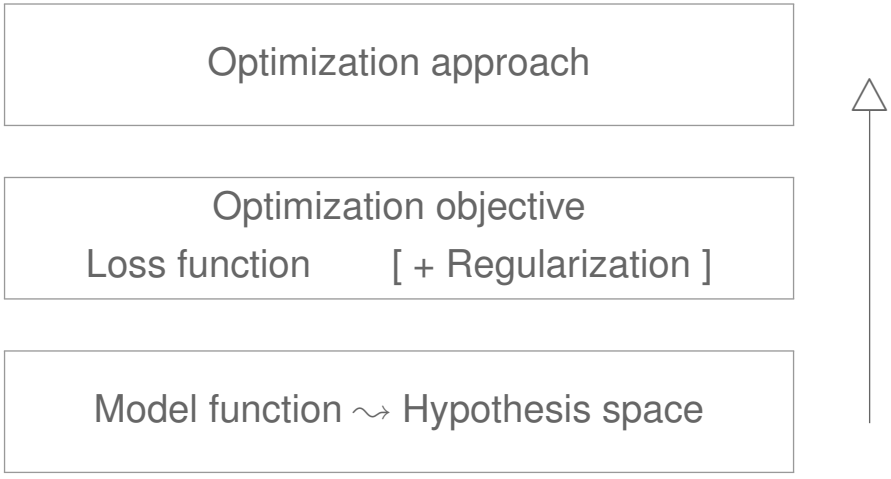
(2) ML Stack: LMS [ML stack: LMS, log. regression, loss comp., regularization, GD]



©STEIN/VÖLSKE 2026

Elements of Machine Learning

(2) ML Stack: LMS [ML stack: LMS, log. regression, loss comp., regularization, GD]



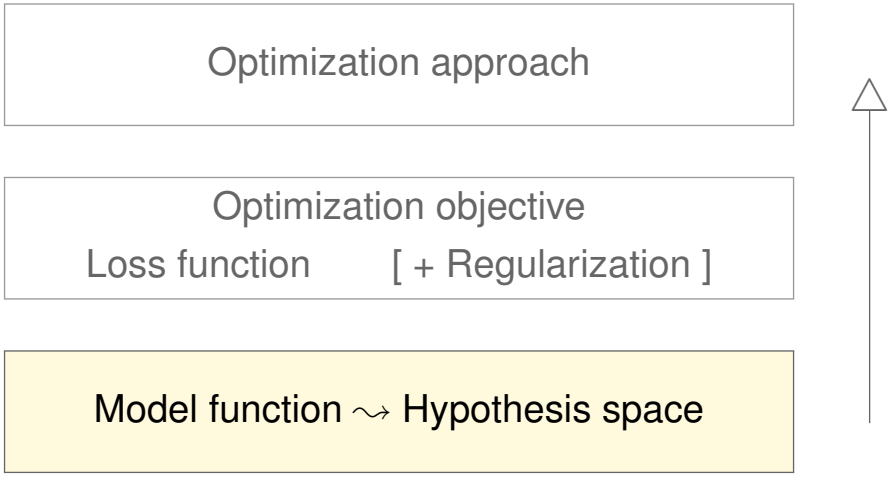
Binary classification

$$D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times \{-1, 1\}$$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(2) ML Stack: LMS [ML stack: LMS, log. regression, loss comp., regularization, GD]



- Hypothesis space: $\mathbf{w} \in \mathbf{R}^{p+1}$
- Linear model: $y(\mathbf{x}) = w_0 + \sum_{i=1}^{p} w_i x_i$

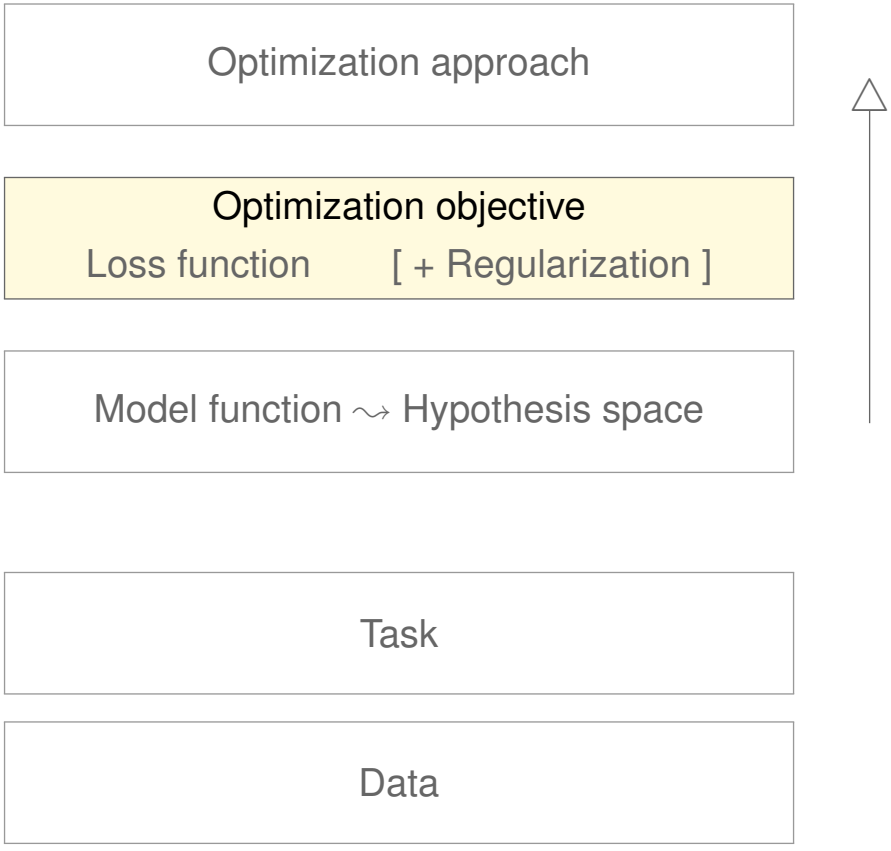
Binary classification

$$D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times \{-1, 1\}$$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(2) ML Stack: LMS [ML stack: LMS, log. regression, loss comp., regularization, GD]



- Objective: minimize squared loss (RSS)
- Regularization: none
- Loss: $l_2(c, y(\mathbf{x})) = (c - y(\mathbf{x}))^2, \ (\mathbf{x}, c) \in D$

- Hypothesis space: $\mathbf{w} \in \mathbf{R}^{p+1}$
- Linear model: $y(\mathbf{x}) = w_0 + \sum_{i=1}^{p} w_i x_i$

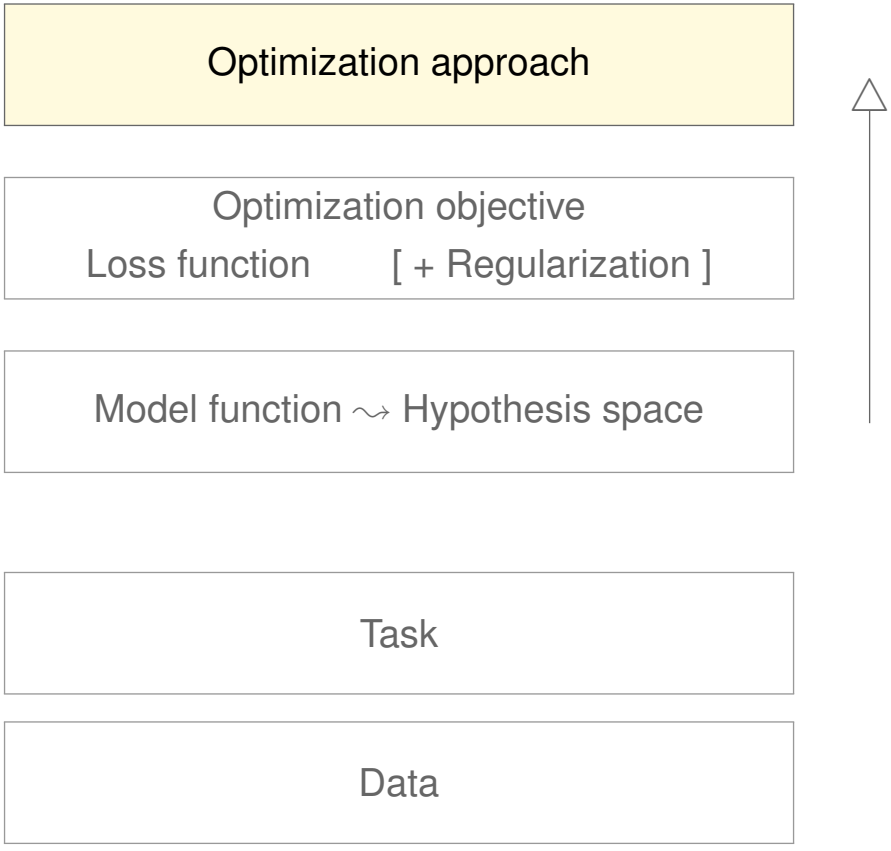
Binary classification

$D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times \{-1, 1\}$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(2) ML Stack: LMS [ML stack: LMS, log. regression, loss comp., regularization, GD]



Stochastic gradient descent (SGD)

- Objective: minimize squared loss (RSS)
- Regularization: none
- Loss: $l_2(c, y(\mathbf{x})) = (c - y(\mathbf{x}))^2, \ (\mathbf{x}, c) \in D$

- Hypothesis space: $\mathbf{w} \in \mathbf{R}^{p+1}$
- Linear model: $y(\mathbf{x}) = w_0 + \sum_{i=1}^{p} w_i x_i$

Binary classification

$D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times \{-1, 1\}$

©STEIN/VÖLSKE 2026

Related questions:

- What are useful classes of model functions?
- What are methods to fit (= learn) model functions?
- What are measures to assess the goodness of fit?

- How does (label) noise affect the learning process?
- How does the example number affect the learning process?
- How to deal with extreme class imbalance?

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(3) Feature Space Structure

Example 1. The feature space is an inner product space.

- An inner product space is a vector space with an operation called “inner product”.
- Example: Euclidean vector space equipped with the dot product.
- Enables algorithms such as gradient descent and support vector machines.

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(3) Feature Space Structure

Example 1. The feature space is an inner product space.

- An inner product space is a vector space with an operation called “inner product”.
- Example: Euclidean vector space equipped with the dot product.
- Enables algorithms such as gradient descent and support vector machines.

Example 2. The feature space is a σ-algebra.

- A σ-algebra on a set Ω is a collection of subsets of Ω that includes Ω itself, is closed under complement, and is closed under countable unions.
- Enables probability spaces and statistical learning, such as naive Bayes.

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(3) Feature Space Structure

Example 1. The feature space is an inner product space.

- An inner product space is a vector space with an operation called “inner product”.
- Example: Euclidean vector space equipped with the dot product.
- Enables algorithms such as gradient descent and support vector machines.

Example 2. The feature space is a σ-algebra.

- A σ-algebra on a set Ω is a collection of subsets of Ω that includes Ω itself, is closed under complement, and is closed under countable unions.
- Enables probability spaces and statistical learning, such as naive Bayes.

Example 3. The feature space is a finite set of vectors with nominal dimensions.

- Requires concept learning via set splitting as done by decision trees.

©STEIN/VÖLSKE 2026

Remarks:

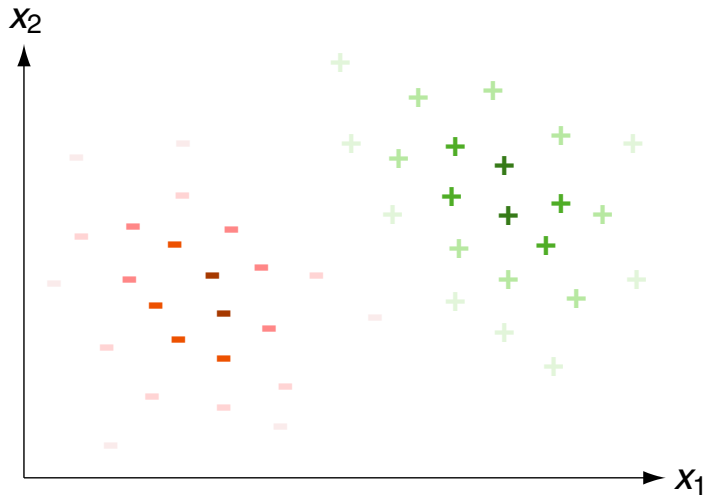
- An inner product space is also called pre-Hilbert space.
- The aforementioned examples of feature spaces are not meant to be complete. But, they illustrate a broad range of structures underlying the example sets we want to learn from.
- The structure of a feature space constrains the applicable learning algorithm. Usually, this structure is inherently determined by the application domain and cannot be chosen.

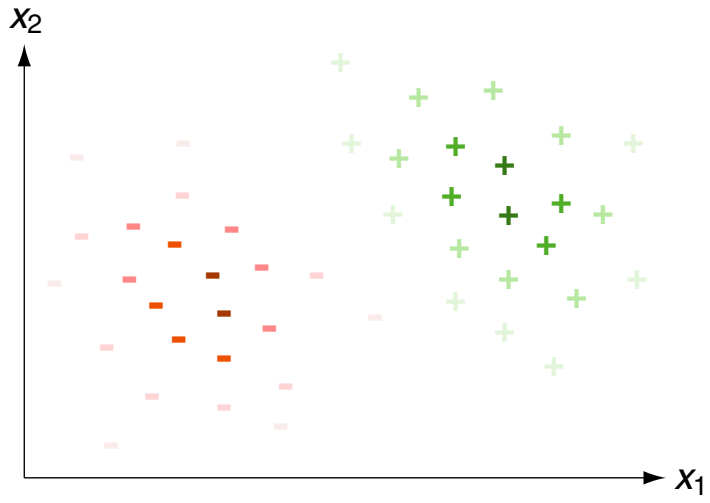
©STEIN/VÖLSKE 2026

Elements of Machine Learning

(4) Discriminative versus Generative Approach to Classification

- Discriminative classifiers (models) learn a boundary between classes.
- Generative classifiers exploit the distributions underlying the classes.



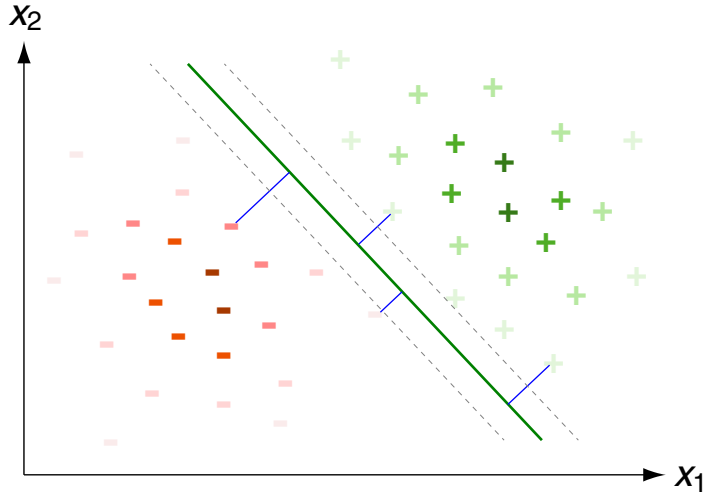


©STEIN/VÖLSKE 2026

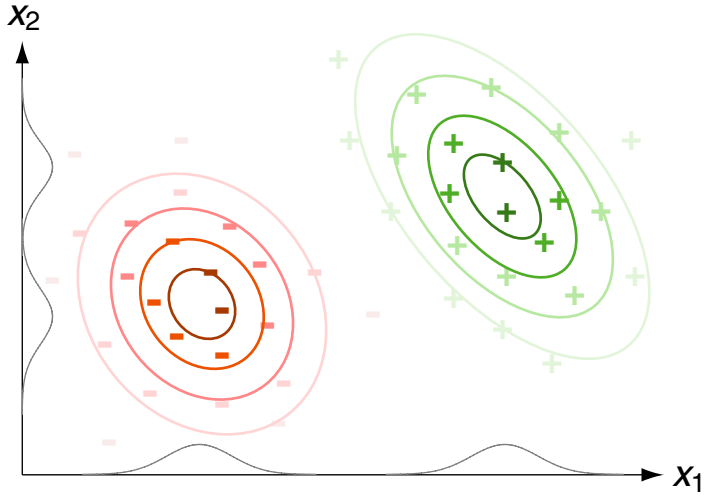
Elements of Machine Learning

(4) Discriminative versus Generative Approach to Classification

- Discriminative classifiers (models) learn a boundary between classes.
- Generative classifiers exploit the distributions underlying the classes.



discriminative
⇝ classification rule



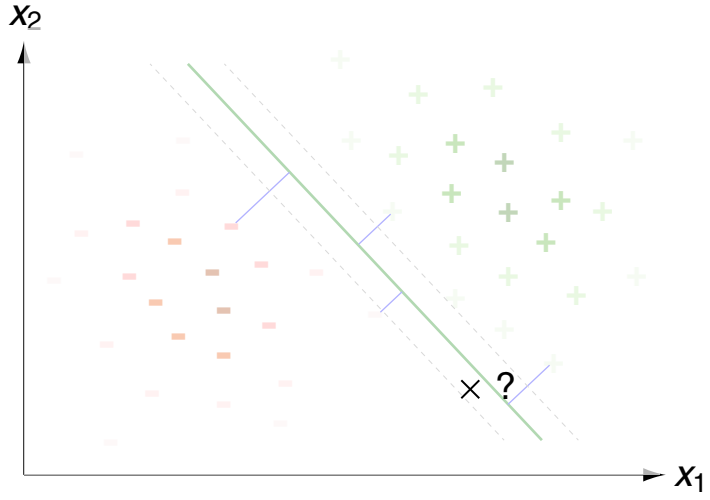
generative
⇝ class membership probability

©STEIN/VÖLSKE 2026

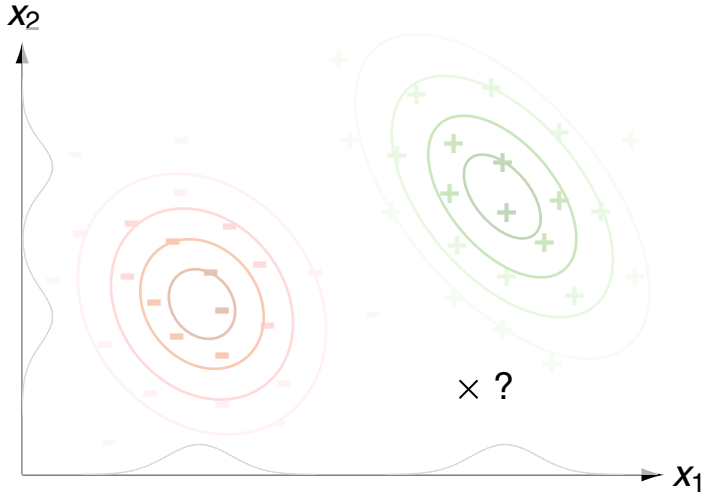
Elements of Machine Learning

(4) Discriminative versus Generative Approach to Classification

- Discriminative classifiers (models) learn a boundary between classes.
- Generative classifiers exploit the distributions underlying the classes.



discriminative
$\rightsquigarrow$ classification rule



generative
$\rightsquigarrow$ class membership probability

©STEIN/VÖLSKE 2026

Remarks:

- When classifying a new example $\mathbf{x}$, then
  (1) discriminative classifiers apply a decision rule that was learned via minimizing the misclassification rate given training examples D, while
  (2) generative classifiers maximize the probability of the combined event $p(\mathbf{x}, y)$, or, similarly, the posterior probability $p(y \mid \mathbf{x})$, $y \in \{\ominus, \oplus\}$.
- The LMS algorithm computes "only" a decision boundary, i.e., it constructs a discriminative classifier. A Bayes classifier is an example for a generative model.
- Yoav Freund provides an excellent video illustrating the pros and cons of discriminative and generative models respectively. [YouTube]
- Discriminative models may be further differentiated in models that also determine the posterior class probabilities $p(y \mid \mathbf{x})$ (without computing the joint probabilities $p(\mathbf{x}, y)$) and those that do not. In the latter case, only a so-called "discriminant function" is computed.

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(5) Frequentist versus Subjectivist Paradigm to Learning

Frequentist:

- There is a hidden, unique mechanism that generated the data D.
- Consider a model for this mechanism, such as a family of distributions or a model function, parameterized by θ, $\boldsymbol{\theta}$, $\mathbf{w}$, or similar. The considered parameter values (or vectors) form the hypothesis space H.
- Select for the unknown parameter that element from H such that the observed data D becomes most probable. The chosen element (our hypothesis), h_{ML} (here: θ_{ML}), is called maximum likelihood hypothesis. $p(D; h)$, also written as $L(h \mid D)$, is called likelihood.





©STEIN/VÖLSKE 2026

Elements of Machine Learning

(5) Frequentist versus Subjectivist Paradigm to Learning

Frequentist:

- There is a hidden, unique mechanism that generated the data D.
- Consider a model for this mechanism, such as a family of distributions or a model function, parameterized by θ, $\boldsymbol{\theta}$, $\mathbf{w}$, or similar. The considered parameter values (or vectors) form the hypothesis space H.
- Select for the unknown parameter that element from H such that the observed data D becomes most probable. The chosen element (our hypothesis), h_{ML} (here: θ_{ML}), is called maximum likelihood hypothesis. $p(D; h)$, also written as $L(h \mid D)$, is called likelihood.





$$\theta^*, \boldsymbol{\theta}^* \text{ or } \mathbf{w}^* \rightsquigarrow D, \qquad h_{\text{ML}} = \underset{h \in H}{\text{argmax}}\ p(D; h)$$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

(5) Frequentist versus Subjectivist Paradigm to Learning

Frequentist:

- There is a hidden, unique mechanism that generated the data D.
- Consider a model for this mechanism, such as a family of distributions or a model function, parameterized by θ, $\boldsymbol{\theta}$, $\mathbf{w}$, or similar. The considered parameter values (or vectors) form the hypothesis space H.
- Select for the unknown parameter that element from H such that the observed data D becomes most probable. The chosen element (our hypothesis), h_{ML} (here: θ_{ML}), is called maximum likelihood hypothesis. $p(D; h)$, also written as $L(h \mid D)$, is called likelihood.





$$\theta^*, \boldsymbol{\theta}^* \text{ or } \mathbf{w}^* \rightsquigarrow D, \qquad h_{\text{ML}} = \underset{h \in H}{\text{argmax}}\ p(D; h)$$

$$\theta_{\text{ML}} = \underset{\theta \in [0;1]}{\text{argmax}}\ p(D; \theta) = \underset{\theta \in [0;1]}{\text{argmax}}\ \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$$

©STEIN/VÖLSKE 2026

Remarks:

- Frequentist = Let the (large amount of) data speak.
- Likelihood is a function of a parameter θ that measures how plausible different parameter values are, given the observed data. In the example: The likelihood of a specific value for θ is proportional to the probability of observing a sequence D of heads and tails in the coin flip experiment under the assumption that the value for θ were the true parameter.

  Among others, the concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes. I.e., $p(D;\theta)$ is called likelihood (often noted as $L(\theta \mid D)$) since we reason about a past coin flip experiment parameterized by θ. [Mathworld]
- The unknown parameter value (vector) of the data generation mechanism, θ^*, $\boldsymbol{\theta}^*$, $\mathbf{w}^*$, etc., has some value from H. In particular, θ, $\boldsymbol{\theta}$, or h in the argmax-expression is not the realization of a random variable or random vector—which would come along with a distribution and an expected value—but an *exogenous parameter (vector), which we vary* to find the maximum of $p(D;\theta)$, $p(D;\boldsymbol{\theta})$, $p(D;\mathbf{w})$, or, in general, $p(D;h)$.

  The fact that h is a given, unique parameter (though it needs to be searched) and *not* a random variable is reflected by the notation, which uses a »;« instead of a »|« in $p()$.
- In the experiment of flipping a coin, we assume a Laplace experiment and apply the binomial distribution, $B(n,p)$, with exactly k successes in n independent Bernoulli trials.
- A method for finding the maximum likelihood estimates of the parameters of an underlying distribution from a given dataset D (even if the data are incomplete) is the Expectation-Maximization (EM) algorithm. [Bilmes 1998]

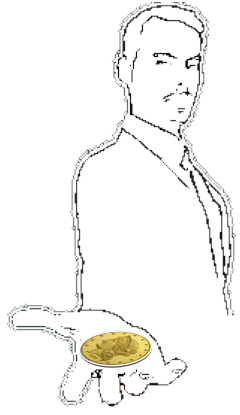
©STEIN/VÖLSKE 2026

Elements of Machine Learning

(5) Frequentist versus Subjectivist Paradigm to Learning (continued)

Subjectivist:

- We consider alternative mechanisms that could have generated the data D.
- As before, consider a model for the mechanisms. Moreover, we have beliefs (prior probabilities) $p(h)$ for the elements (alternative mechanisms, here: θ_1, θ_2) in H.
- Select the most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the likelihoods $p(D \mid h), h \in H$, with the corresponding priors. h_{MAP} (here: θ_{MAP}) is called maximum posterior hypothesis.



©STEIN/VÖLSKE 2026

Elements of Machine Learning

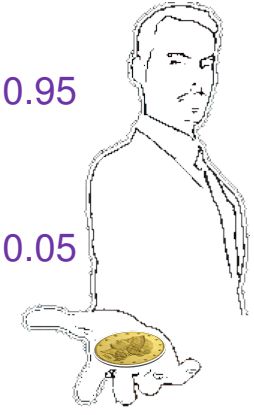
(5) Frequentist versus Subjectivist Paradigm to Learning (continued)

Subjectivist:

- We consider alternative mechanisms that could have generated the data D.
- As before, consider a model for the mechanisms. Moreover, we have beliefs (prior probabilities) $p(h)$ for the elements (alternative mechanisms, here: θ_1, θ_2) in H.
- Select the most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the likelihoods $p(D \mid h), h \in H$, with the corresponding priors. h_{MAP} (here: θ_{MAP}) is called maximum posterior hypothesis.

Belief / Prior 1: $P(\Theta{=}\underbrace{0.50}_{\theta_1}) = 0.95$

Belief / Prior 2: $P(\Theta{=}\underbrace{0.75}_{\theta_2}) = 0.05$



©STEIN/VÖLSKE 2026

Elements of Machine Learning

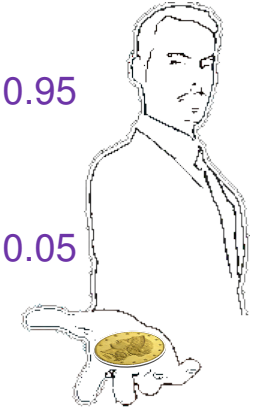
(5) Frequentist versus Subjectivist Paradigm to Learning (continued)

Subjectivist:

- We consider alternative mechanisms that could have generated the data D.
- As before, consider a model for the mechanisms. Moreover, we have beliefs (prior probabilities) $p(h)$ for the elements (alternative mechanisms, here: θ_1, θ_2) in H.
- Select the most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the likelihoods $p(D \mid h), h \in H$, with the corresponding priors. h_{MAP} (here: θ_{MAP}) is called maximum posterior hypothesis.

Belief / Prior 1: $P(\Theta=\underbrace{0.50}_{\theta_1}) = 0.95$

Belief / Prior 2: $P(\Theta=\underbrace{0.75}_{\theta_2}) = 0.05$



$$\theta_1 + D \rightarrow p(D \mid \theta_1)$$

$$\theta_2 + D \rightarrow p(D \mid \theta_2)$$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

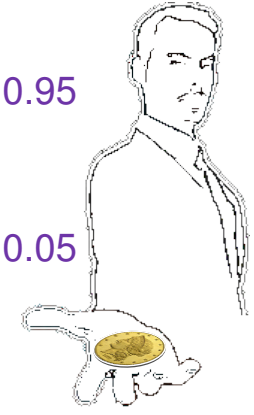
(5) Frequentist versus Subjectivist Paradigm to Learning (continued)

Subjectivist:

- We consider alternative mechanisms that could have generated the data D.
- As before, consider a model for the mechanisms. Moreover, we have beliefs (prior probabilities) $p(h)$ for the elements (alternative mechanisms, here: θ_1, θ_2) in H.
- Select the most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the likelihoods $p(D \mid h), h \in H$, with the corresponding priors. h_{MAP} (here: θ_{MAP}) is called maximum posterior hypothesis.

Belief / Prior 1: $P(\Theta{=}\underbrace{0.50}_{\theta_1}) = 0.95$

Belief / Prior 2: $P(\Theta{=}\underbrace{0.75}_{\theta_2}) = 0.05$



$$\theta_1 + D \;\rightarrow\; p(D \mid \theta_1)$$

$$\theta_2 + D \;\rightarrow\; p(D \mid \theta_2)$$

$$\theta_{\text{MAP}} = \underset{\theta \in \{\theta_1, \theta_2\}}{\text{argmax}}\; p(\theta \mid D) = \underset{\theta \in \{\theta_1, \theta_2\}}{\text{argmax}}\; \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

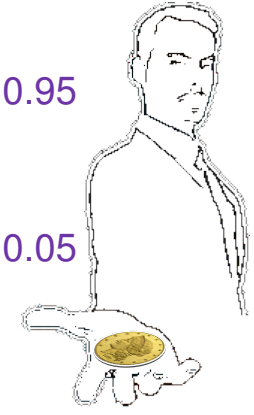
(5) Frequentist versus Subjectivist Paradigm to Learning (continued)

Subjectivist:

- We consider alternative mechanisms that could have generated the data D.
- As before, consider a model for the mechanisms. Moreover, we have beliefs (prior probabilities) $p(h)$ for the elements (alternative mechanisms, here: θ_1, θ_2) in H.
- Select the most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the likelihoods $p(D \mid h), h \in H$, with the corresponding priors. h_{MAP} (here: θ_{MAP}) is called maximum posterior hypothesis.

Belief / Prior 1: $P(\Theta{=}\underbrace{0.50}_{\theta_1}) = 0.95$

Belief / Prior 2: $P(\Theta{=}\underbrace{0.75}_{\theta_2}) = 0.05$



$$\theta_1 + D \;\rightarrow\; p(D \mid \theta_1)$$

$$\theta_2 + D \;\rightarrow\; p(D \mid \theta_2)$$

$$\theta_{\text{MAP}} = \underset{\theta \in \{\theta_1, \theta_2\}}{\text{argmax}}\; p(\theta \mid D) = \underset{\theta \in \{\theta_1, \theta_2\}}{\text{argmax}}\; \frac{\binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k} \cdot p(\theta)}{p(D)}$$

©STEIN/VÖLSKE 2026

Elements of Machine Learning

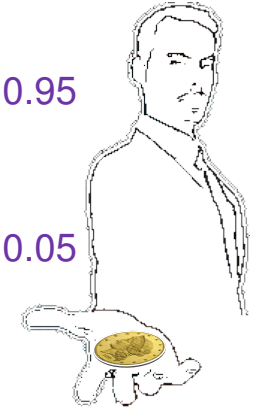
(5) Frequentist versus Subjectivist Paradigm to Learning (continued)

Subjectivist:

- We consider alternative mechanisms that could have generated the data D.
- As before, consider a model for the mechanisms. Moreover, we have beliefs (prior probabilities) $p(h)$ for the elements (alternative mechanisms, here: θ_1, θ_2) in H.
- Select the most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the likelihoods $p(D \mid h), h \in H$, with the corresponding priors. h_{MAP} (here: θ_{MAP}) is called maximum posterior hypothesis.

Belief / Prior 1: $P(\Theta{=}\underbrace{0.50}_{\theta_1}) = 0.95$

Belief / Prior 2: $P(\Theta{=}\underbrace{0.75}_{\theta_2}) = 0.05$



$$\theta_1 + D \;\rightarrow\; p(D \mid \theta_1)$$

$$\theta_2 + D \;\rightarrow\; p(D \mid \theta_2)$$

$$\theta_{\text{MAP}} = \underset{\theta \in \{\theta_1, \theta_2\}}{\text{argmax}}\; p(\theta \mid D) = \underset{\theta \in \{\theta_1, \theta_2\}}{\text{argmax}}\; \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k} \cdot p(\theta)$$

©STEIN/VÖLSKE 2026

Remarks:

- Subjectivist = Let (also) the knowledge (about priors, prevalences) speak.
- The elements in H (here: θ_1, θ_2) are realizations of a random variable Θ, and there is (subjective) knowledge about the distribution of Θ (= the prior probability model). Θ models the parameter p of the binomial distribution and defines the success probability for each trial.
  - Belief in θ_1 ($\Theta{=}0.5$): With probability 0.95 the coin is fair, i.e., sides are equally likely.
  - Belief in θ_2 ($\Theta{=}0.75$): With probability 0.05 the odds of preferring one side is 3:1.

  We compute for each element in H the conditional probability of the observed data D, i.e., $p(D \mid \theta_1)$ and $p(D \mid \theta_2)$ under the binomial distribution. We then compute the respective values for $p(\theta_1 \mid D)$ and $p(\theta_2 \mid D)$ with Bayes's rule, and finally select θ_{MAP}.

  The fact that h is the realization of a random variable (and not an exogenous parameter) is reflected by the standard probability notation, which uses a »|« in $p()$ (and not a »;«).
- The subjectivist paradigm is powerful, if we can consider knowledge about H in combination with data D. The subjectivist paradigm is necessary, if we have no data D at all to optimize, e.g., if we reason about "one time events". If all hypotheses are equally likely (a uniform prior), ML optimization and MAP optimization yield the same result.
- The subjectivist paradigm is also called Bayesian interpretation of probability. It enables by design the integration of prior knowledge or human expertise about alternative mechanisms one of which generated D.
  [Wikipedia: Bayesian interpretation, probability interpretations]

©STEIN/VÖLSKE 2026