I. Introduction

- Examples of Learning Tasks
- Specification of Learning Tasks
- Elements of Machine Learning
- Comparative Syntax Overview
- Functions Overview
- Algorithms Overview
- Classification Approaches Overview
Elements of Machine Learning

(1) Model Formation: Real World → Model World

- **Objects**: $O$
- **Classes**: $C$
- **Feature vectors**: $X$

$\alpha(o)$ → $\gamma(o)$ → $\gamma(x)$ → $c(x) \approx y(x)$

Related questions:
- From what kind of experience should be learned?
- Which level of fidelity is sufficient to solve a certain task?
Elements of Machine Learning
(2) Design Choices for Model Function Construction

- Optimization approach
- Optimization objective: Loss function [ + Regularization ]
- Model function \( \sim \) Hypothesis space
- Task
- Data
(2) Design Choices for Model Function Construction: LMS in a Nutshell

Optimization approach

Optimization objective
Loss function \[ + \text{ Regularization} \]

Model function \[ \sim \] Hypothesis space

Task

Data

Binary classification

\[
D = \{ (x_1, c(x_1)), \ldots, (x_n, c(x_n)) \} \subseteq X \times \{-1, 1\}
\]
Elements of Machine Learning

(2) Design Choices for Model Function Construction: LMS in a Nutshell (continued)

- Optimization approach
- Optimization objective
  - Loss function [ + Regularization ]
- Model function $\sim$ Hypothesis space

- Task
- Data

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

Binary classification

$D = \{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\} \subseteq X \times \{-1, 1\}$
Elements of Machine Learning

(2) Design Choices for Model Function Construction: LMS in a Nutshell (continued)

- Optimization approach
  - Optimization objective
    - Loss function [ + Regularization ]
  - Model function \( \sim \) Hypothesis space

- Objective: Minimize squared loss
- Regularization: None
- Loss: Sum of squared residuals

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

Binary classification

\[ D = \{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\} \subseteq X \times \{-1, 1\} \]
Optimization approach

- Optimization objective
  - Loss function \[ + \text{ Regularization} \]

Model function \( \sim \) Hypothesis space

Stochastic gradient descent

- Objective: Minimize squared loss
- Regularization: None
- Loss: Sum of squared residuals
- Hypothesis space: \( w \in \mathbb{R}^{p+1} \)
- Linear model: \( y(x) = w_0 + \sum_{i=1}^{p} w_i x_i \)

Task

Data

Binary classification

\[ D = \{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\} \subseteq X \times \{-1, 1\} \]
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?
The feature space is an inner product space.

- An inner product space (also called pre-Hilbert space) is a vector space with an additional structure called “inner product”.
- Example: Euclidean vector space equipped with the dot product.
- Enables algorithms such as gradient descent and support vector machines.
The feature space is an inner product space.

- An inner product space (also called pre-Hilbert space) is a vector space with an additional structure called “inner product”.
- Example: Euclidean vector space equipped with the dot product.
- Enables algorithms such as gradient descent and support vector machines.

The feature space is a $\sigma$-algebra.

- A $\sigma$-algebra on a set $X$ is a collection of subsets of $X$ that includes $X$ itself, is closed under complement, and is closed under countable unions.
- Enables probability spaces and statistical learning, such as naive Bayes.
The feature space is an inner product space.

- An inner product space (also called pre-Hilbert space) is a vector space with an additional structure called “inner product”.
- Example: Euclidean vector space equipped with the dot product.
- Enables algorithms such as gradient descent and support vector machines.

The feature space is a $\sigma$-algebra.

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

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

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

- The aforementioned examples of feature spaces are not meant to be complete. However, 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.
Discriminative classifiers (models) learn a boundary between classes.

Generative classifiers exploit the distributions underlying the classes.
Discriminative vs. Generative Approach to Classification (continued)

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

Generative classifiers exploit the distributions underlying the classes.
Remarks:

- When classifying a new example (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 \( P(x=x, y=y) \), or, similarly, the a-posteriori probability \( P(y=y | x=x) \), \( y \in \{ \ominus, \oplus \} \).

- 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 \( P(y=y | x=x) \) (without computing the joint probabilities \( P(x=x, y=y) \)) and those that do not. In the latter case, only a so-called “discriminant function” is computed.
Frequentism:

- There is a (hidden) mechanism that generates $D$.
- To model this mechanism you consider
  - a family of distributions, or
  - a model function, or
  - a combination of both,

parameterized by $\theta$. The possible values for $\theta$ form the hypothesis space $H$.
- Select a most probable hypothesis $h_{ML} \in H$ by estimating $\theta$ using a sample $D' \subset D$. $h_{ML}$ is called maximum likelihood hypothesis.
Frequentism:

- There is a (hidden) mechanism that generates $D$.
- To model this mechanism you consider
  - a family of distributions, or
  - a model function, or
  - a combination of both,

parameterized by $\theta$. The possible values for $\theta$ form the hypothesis space $H$.

- Select a most probable hypothesis $h_{ML} \in H$ by estimating $\theta$ using a sample $D' \subset D$. $h_{ML}$ is called maximum likelihood hypothesis.
Frequentist Approach to Parameter Estimation (continued)

Frequentism:

- There is a (hidden) mechanism that generates $D$.
- To model this mechanism you consider
  - a family of distributions, or
  - a model function, or
  - a combination of both,

parameterized by $\theta$. The possible values for $\theta$ form the hypothesis space $H$.

- Select a most probable hypothesis $h_{ML} \in H$ by estimating $\theta$ using a sample $D' \subset D$. $h_{ML}$ is called maximum likelihood hypothesis.
Frequentism:

- There is a (hidden) mechanism that generates $D$.

- To model this mechanism you consider
  - a family of distributions, or
  - a model function, or
  - a combination of both,

  parameterized by $\theta$. The possible values for $\theta$ form the hypothesis space $H$.

- Select a most probable hypothesis $h_{ML} \in H$ by estimating $\theta$ using a sample $D' \subset D$. $h_{ML}$ is called maximum likelihood hypothesis.

$$\theta \sim D', \quad h_{ML} = \underset{h \in H}{\text{argmax}} \ P(D' \mid h)$$
Remarks:

- $\theta$ is a parameter or a parameter vector that is considered as fixed (in particular: not as a random variable), but unknown.

- In the experiment of flipping a coin, one may suppose a Laplace experiment and consider the binomial distribution, $B(n,p)$.

- $P(D' \mid h)$ is the probability of observing $D'$ under $h$. I.e., it is the probability of observing $D'$ if the hidden mechanism that generates $D'$ behaves according to the considered model whose parameter $\theta$ is set to $h$. 
Subjectivism:

- Consider a model for the mechanism that has generated $D$.
- There are different beliefs about the parameter (vector) $\theta$ that characterizes the model. The possible values for $\theta$ form the hypothesis space $H$.
- Select a most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the ML estimates under $D$ with the priors. $h_{\text{MAP}}$ is called maximum a-posteriori hypothesis.

Belief/Prior 1: $P(p = 0.5) = 0.95$ for $\theta_1$

Belief/Prior 2: $P(p = 0.75) = 0.50$ for $\theta_2$
Subjectivism:

- Consider a model for the mechanism that has generated $D$.
- There are different beliefs about the parameter (vector) $\theta$ that characterizes the model. The possible values for $\theta$ form the hypothesis space $H$.
- Select a most probable hypothesis $h_{MAP} \in H$ by weighting the ML estimates under $D$ with the priors. $h_{MAP}$ is called maximum a-posteriori hypothesis.

Belief/Prior 1: $P(p = 0.5_{\theta_1}) = 0.95$

Belief/Prior 2: $P(p = 0.75_{\theta_2}) = 0.50$
Subjectivism:

- Consider a model for the mechanism that has generated $D$.
- There are different beliefs about the parameter (vector) $\theta$ that characterizes the model. The possible values for $\theta$ form the hypothesis space $H$.
- Select a most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the ML estimates under $D$ with the priors. $h_{\text{MAP}}$ is called maximum a-posteriori hypothesis.

\[
\begin{align*}
\text{Belief/Prior 1: } & P(p = 0.5) = 0.95 \\
\text{Belief/Prior 2: } & P(p = 0.75) = 0.50
\end{align*}
\]
Subjectivism:

- Consider a model for the mechanism that has generated $D$.
- There are **different beliefs** about the parameter (vector) $\theta$ that characterizes the model. The possible values for $\theta$ form the hypothesis space $H$.
- Select a most probable hypothesis $h_{\text{MAP}} \in H$ by weighting the ML estimates under $D$ with the priors. $h_{\text{MAP}}$ is called **maximum a-posteriori hypothesis**.

\[
\begin{align*}
\text{Belief/Prior 1: } & P(p = 0.5) = 0.95 \\
\text{Belief/Prior 2: } & P(p = 0.75) = 0.50
\end{align*}
\]

\[
\begin{align*}
\theta_1 + D & \rightarrow P(D \mid \theta_1) \\
\theta_2 + D & \rightarrow P(D \mid \theta_2)
\end{align*}
\]
Subjectivism:

- Consider a model for the mechanism that has generated $D$.
- There are different beliefs about the parameter (vector) $\theta$ that characterizes the model. The possible values for $\theta$ form the hypothesis space $H$.
- Select a most probable hypothesis $h_{MAP} \in H$ by weighting the ML estimates under $D$ with the priors. $h_{MAP}$ is called maximum a-posteriori hypothesis.

\[
\begin{array}{l}
\theta_1 + D \rightarrow P(D | \theta_1) \\
\theta_2 + D \rightarrow P(D | \theta_2)
\end{array}
\]

\[\begin{align*}
h_{MAP} &= \arg\max_{h \in \{\theta_1, \theta_2\}} P(h | D) \\
&= \arg\max_{h \in \{\theta_1, \theta_2\}} \frac{P(D | h) \cdot P(h)}{P(D)}
\end{align*}\]

Belief/Prior 1: $P(p = 0.5)_{\theta_1} = 0.95$

Belief/Prior 2: $P(p = 0.75)_{\theta_2} = 0.50$
Remarks:

- $\theta$ is considered as random variable. There is prior knowledge about the distribution of $\theta$.
- $p$ is a parameter of the binomial distribution and denotes the success probability for each trial.
  - Belief 1: With a probability of 0.95 the coin is fair (both sides are equally likely).
  - Belief 2: With a probability of 0.5 the odds of preferring a particular side is 3:1.

Given $D$ from a number of trials compute $P(D \mid \theta_1)$ and $P(D \mid \theta_2)$ and the respective values for $P(\theta_1 \mid D)$ and $P(\theta_2 \mid D)$.

Disclaimer. While only mild conditions are required for MAP estimation to be a limiting case of Bayes estimation, it is not very representative of Bayesian methods in general. This is because MAP estimates are point estimates, whereas Bayesian methods are characterized by the use of distributions to summarize data and draw inferences. [Wikipedia]

- The subjectivist approach is also called Bayesian interpretation of probability. The Bayesian interpretation of probability enables by design the integration of prior knowledge, background knowledge, and human expertise. [Wikipedia: probability interpretations, Bayes interpretations]
- Food for thought: Discuss the use of frequentist and subjectivist approaches to decision making if you had to develop an AI that plays poker.