Chapter ML:I

I. Introduction

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>$x, x_i, x_1, \ldots, x_p$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature vector</td>
<td>$x$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Functions Overview

<table>
<thead>
<tr>
<th>Function name</th>
<th>Function definition</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator function</td>
<td>$I_{\neq}(a, b) = \begin{cases} 0 &amp; a = b \ 1 &amp; a \neq b \end{cases}$</td>
<td>Part: Linear Models, Section: Loss Computation</td>
</tr>
<tr>
<td>function</td>
<td>$f(x) = \ldots$</td>
<td>Part: \ldots, Section: \ldots</td>
</tr>
</tbody>
</table>
# Algorithms Overview

<table>
<thead>
<tr>
<th>Algorithm name</th>
<th>Signature</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least Mean Squares</td>
<td>LMS($D, \eta$)</td>
<td>Part: Introduction, Section: Examples of Learning Tasks</td>
</tr>
<tr>
<td>algorithm</td>
<td>ALG(...)</td>
<td>Part: . . . , Section: . . .</td>
</tr>
</tbody>
</table>
## Classification Approaches Overview

<table>
<thead>
<tr>
<th>Taxonomy</th>
<th>Model function</th>
<th>Classification rule</th>
<th>Optimization principle</th>
<th>Optimization objective</th>
<th>Optimization approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generative approaches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Statistical approaches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unrestricted decision boundary (monothetic analysis)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Discriminative approaches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Linear decision boundary (linear feature space)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Polythetic approaches</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Linear decision boundary (inner product space)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Classification Approaches

- **Generative approaches**
  - Bayes rule for combined conditional events
    - $X \sim N(\mu, \sigma^2)$ (or other family)

- **Statistical approaches**

- **Unrestricted decision boundary (monothetic analysis)**
  - Nominal feature
    - $\bigwedge_i x_i = v_i$
      - $i = 1, \ldots, p$
  - Monothetic feature
    - $\bigvee_i \bigwedge_j x_{ij} = v_{ij}$
      - $i = 1, \ldots, \text{leaves}$
      - $j = 1, \ldots, \text{depth}(\mathcal{D}^{\alpha})$
  - Domain predictors
    - $\mathcal{D}^{\alpha}$ on domain predicates

- **Discriminative approaches**
  - Perceptron
    - $y(x) = \text{sign}(w^T \phi(x))$
  - Logistic function
    - $y(x) = \frac{1}{1 + e^{-w^T \phi(x)}}$
  - SVM with linear kernel
    - $y(x) = \text{sign}(w^T \phi(x))$
  - SVM with nonlinear kernel
    - $y(x) = \frac{1}{1 + e^{-w^T \phi(x)}}$
  - Multilayer perceptron
    - $y(x) = \sigma(\mathbf{W}^T \sigma(\mathbf{W} \theta(x)))$

### Model Functions

- Logistic regression
  - $y(x) = \frac{1}{1 + e^{-w^T \phi(x)}}$
- Linear regression
  - $y(x) = \mathbf{w}^T \mathbf{x}$

### Classification Rules

- Exploit misclassified examples individually: Hebbian learning
  - $w^T \mathbf{x} \geq 0$
- Empirical risk minimization
  - $w^T \mathbf{x} - b \geq 0$
- Decision tree
  - $(\text{greedy})$

### Optimization Principles

- Linear regression
  - $\mathbf{w}^T \mathbf{x}$
- Logistic regression
  - $\mathbf{w}^T \mathbf{x} = b$
- Empirical risk minimization
  - $w^T \phi(x) \geq 0$

### Optimization Objectives

- No misclassified example
  - $\text{argmax}_{c \in C} \{ y_c(x) \}$
- Maximize version space
  - $\mathbf{w} = \mathbf{c} \cdot \mathbf{x}$
- Maximum a-posteriori hypothesis
  - $\text{argmax}_{c \in C} \{ \mathcal{P}(x|\mu_c, \sigma_c) \}$

### Optimization Approaches

- Perceptron training algorithm
  - Gradient descent
    - $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$ (stochastic)
  - Newton-Raphson
    - BFGS
- Logistic regression
  - $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$ (stochastic)
  - Newton-Raphson
    - BFGS
- Regression
  - $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla J$ (stochastic)
  - Newton-Raphson
    - BFGS
- Backpropagation algorithm
  - Candidate elimination algorithm
  - Algorithms: ID3, C4.5, C5.0, CART

---

ML:1-74  Introduction © STEIN 2021