University of Amsterdam
Informatics Institute

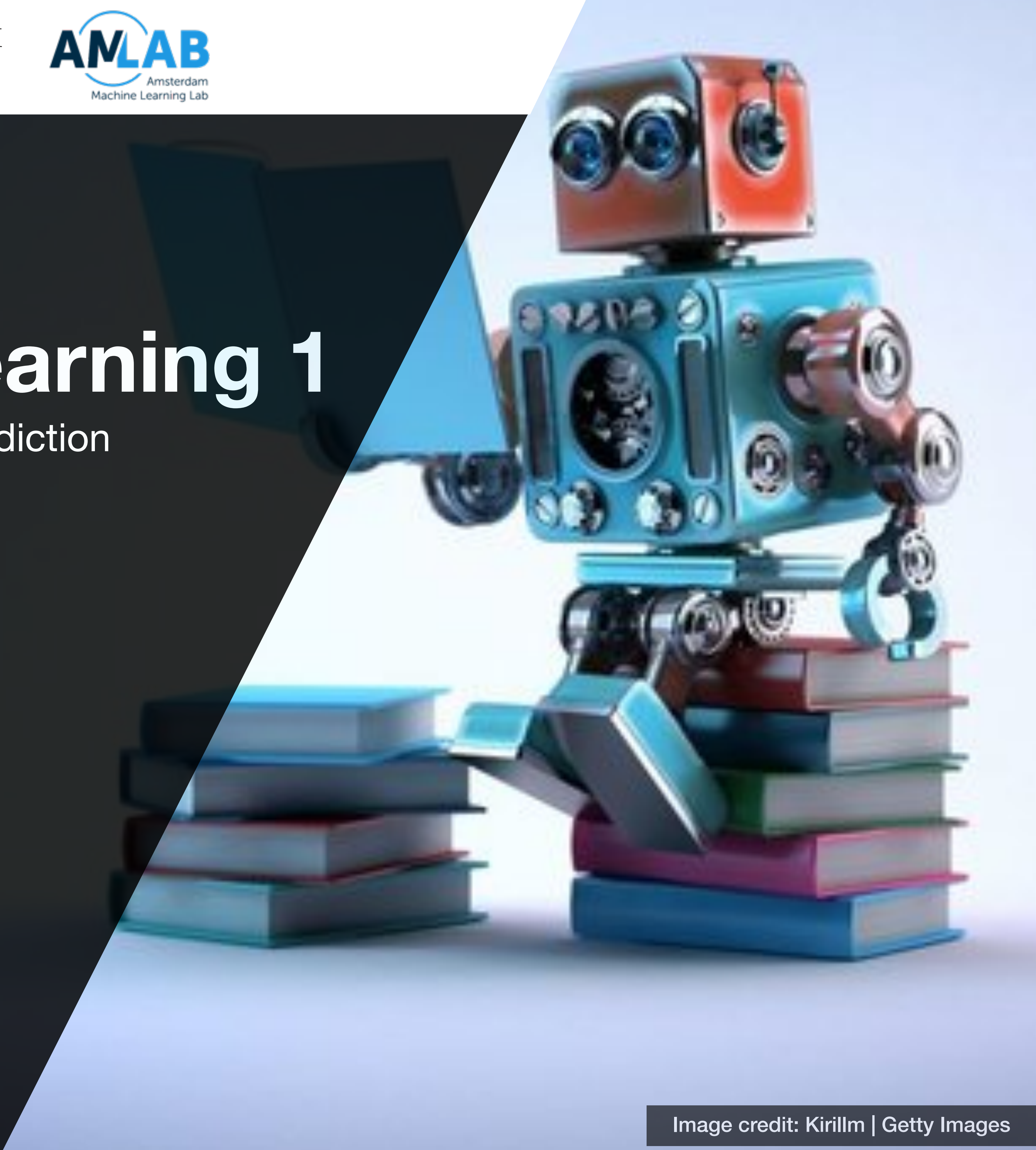
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 2.6 - Bayesian Prediction

Erik Bekkers

(Bishop 1.2.6)



Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Bayesian Approach

$p(t' \mid x', \underline{w}, \beta)$

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

  $(t_1, t_2, \ldots t_N)$

- Frequentist approach: search for one optimal estimate of $\boldsymbol{w}$

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{w} p(D \mid \underline{w})$$

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{w} p(\underline{w} \mid D)$$

- Bayesian approach: Given a prior belief over **w**, p(**w**), and our data D, we are interested in the posterior distribution

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \underline{w})\, p(\underline{w})}{p(D)}$$

- $p(\boldsymbol{w} \mid D)$ reflects the plausibility of different $\boldsymbol{w}$, given our prior knowledge and how likely our data is generated using **w**.

Bayesian Approach

- Prior distribution: $p(\mathbf{w})$, should represent some prior knowledge/belief of the plausibility of ***w***.

- After observing data $D = (x_1, x_2, \ldots, x_N)$, posterior distribution

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})p(\mathbf{w})}{p(D)}$$

- Predictive distribution:

$$p(x' \mid D) = \int p(x', \mathbf{w} \mid D)\, \mathrm{d}\mathbf{w} = \int p(x' \mid \cancel{D}, \underline{w})\, p(\underline{w} \mid D)\, dw$$

$$= \int p(x' \mid \underline{w})\, p(\underline{w} \mid D)\, d\underline{w}$$

Conditional independence

- **Note**: even if $$p(D \mid \mathbf{w}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{w})$$

$$p(D) = \int p(D, \mathbf{w})\, \mathrm{d}\mathbf{w} = \underbrace{\int p(D \mid \mathbf{w}) p(\mathbf{w})\, \mathrm{d}\mathbf{w}}_{} \neq \overbrace{\prod_{i=1}^{N} p(x_i)}^{}$$

$$\int p(x_1 \mid \underline{w})\, p(x_2 \mid \underline{w}) \cdots p(x_N \mid \underline{w})\, p(\underline{w})\, d\underline{w}$$

$$\neq$$

$$\int p(x_1 \mid w_1)\, p(w_1)\, dw_1 \int p(x_2 \mid w_2)\, p(w_2)\, dw_2 \cdots$$

Curve Fitting: Bayesian Approach

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Posterior distribution after observing data:

$$p(\mathbf{w}\,|\,\mathbf{x},\mathbf{t}) = \frac{p(\underline{t}|\underline{x},\underline{w})\,p(\mathbf{w})}{p(\underline{t}|\underline{x})} \quad \text{with} \quad p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x},\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

- Predictive distribution:

$$p(t'\,|\,x',\mathbf{x},\mathbf{t}) = \int p(t',\mathbf{w}\,|\,x',\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{w}$$

$$= \int p(t'|x', \cancel{\underline{x}}, \cancel{\underline{t}}, \underline{w}) \cdot p(\underline{w}|\underline{x}, \underline{t})\, d\underline{w}$$

$$= \int p(t'|x', \underline{w}) \cdot p(\underline{w}|\underline{x}, \underline{t})\, d\underline{w}$$

Curve Fitting: Bayesian Approach

- Predictive distribution: $p(t'|x', \mathbf{x}, \mathbf{t}) = \int p(t'|x', \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{w}$

$p(\mathbf{w}|\mathbf{x}, \mathbf{t}) = \dfrac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x})}$ with $p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}) \mathrm{d}\mathbf{w}$

Advantages:

- Inclusion of prior knowledge
- Represents uncertainty in *t'* both due to target noise, and uncertainty over ***w***.

Disadvantages:

- Posterior is hard to compute analytically → approximate!
- Prior is often chosen for mathematical convenience, not reflection of prior belief!