Machine learning

interpretability

Mads Jensen, PhD

mads@cas.au.dk

AARHUS UNIVERSITY

INTERACTING MINDS CENTRE

CFIN

Contents

Interpretability

Interpretability

- what is interpretability?
- why care about interpretability?
- how do we get interpretability?

Interpretability

- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
- how do we get interpretability?

Interpretability

- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
  - understanding of the model behaviour
- how do we get interpretability?

Interpretability

- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
  - understanding of the model behaviour
  - get inside the black box
- how do we get interpretability?

Interpretability

- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
  - understanding of the model behaviour
  - get inside the black box
  - from the EU GDPR: [the data subject should have] the right . . . to obtain an explanation of the decision reached.
    From https://en.wikipedia.org/wiki/Right_to_explanation
- how do we get interpretability?

Interpretability

- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
  - understanding of the model behaviour
  - get inside the black box
  - from the EU GDPR: [the data subject should have] the right . . . to obtain an explanation of the decision reached.
    From https://en.wikipedia.org/wiki/Right_to_explanation
  - in cognitive neuroscience we want to know *why* something happened
- how do we get interpretability?

Interpretability

- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
  - understanding of the model behaviour
  - get inside the black box
  - from the EU GDPR: [the data subject should have] the right . . . to obtain an explanation of the decision reached.
    From https://en.wikipedia.org/wiki/Right_to_explanation
  - in cognitive neuroscience we want to know *why* something happened
    - e.g. what is the difference between seeing houses and faces?
- how do we get interpretability?

Interpretability

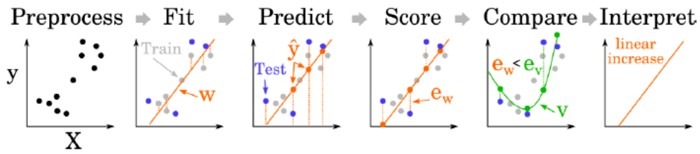
- what is interpretability?
  - "Interpretability is the degree to which a human can understand the cause of a decision." (Miller cited in Molnar, 2020, p. 18)
  - "Interpretability is the degree to which a human can consistently predict the model's result"(Kim et al. cited in Molnar, 2020, p. 18)
- why care about interpretability?
  - understanding of the model behaviour
  - get inside the black box
  - from the EU GDPR: [the data subject should have] the right . . . to obtain an explanation of the decision reached.
    From https://en.wikipedia.org/wiki/Right_to_explanation
  - in cognitive neuroscience we want to know *why* something happened
    - e.g. what is the difference between seeing houses and faces?
- how do we get interpretability?
  - the topic of today's lecture

Machine learning recap

- create features
- make cross-validation scheme
- fit model
- interpret model

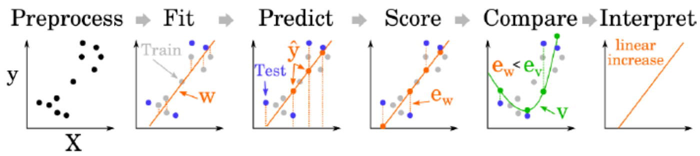
Machine learning recap

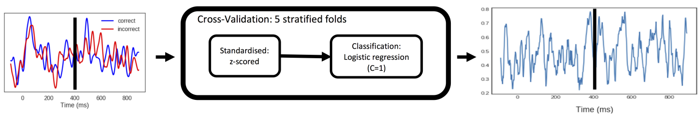
- create features
- make cross-validation scheme
- fit model
- interpret model



Machine learning recap

- create features
- make cross-validation scheme
- fit model
- interpret model





Top figure from King et al. (2018)

bottom figure mine.

Explanations

Why care about explanations?

Why care about explanations?

Science . . .

Why care about explanations?

Science . . .

Example:

Explainable machine learning (interpretable ML)

What is an explanation?

1. What is the aim of an explanation?
2. What is the structure of an explanation?

The aim of explanations

Explanation and understanding

- Knowledge of a fact[1]

[1] Fact is meant to include facts, statements, theories etc.

The aim of explanations

Explanation and understanding

- Knowledge of a fact[1]
- That the fact happened*

[1] Fact is meant to include facts, statements, theories etc.

The aim of explanations

Explanation and understanding

- Knowledge of a fact[1]
- That the fact happened*
- **Explanation**: understand why the fact happened

[1]Fact is meant to include facts, statements, theories etc.

The aim of explanations

Explanation and understanding

- Knowledge of a fact[1]
- That the fact happened*
- **Explanation**: understand why the fact happened

"What has to be added to knowledge to yield understanding".
(Lipton, 2004, p. 21)

[1]Fact is meant to include facts, statements, theories etc.

The structure of explanations

- **Explanandum**: the fact to be explained
- **Explanans**: the statements that explains

Types of explanations[1]

- Psychological explanation
- Functional explanation
- Mechanistic explanation
- Nomic explanation (also called nomological explanation)
- Casual explanation

[1] For more see e.g. Bird (2003), esp. chapter 2

Contrastive explanation

- Explaining why P happened rather than Q.

Contrastive explanation

- Explaining why P happened rather than Q.
- Fact and foil
  (P is the fact, Q the foil)

Contrastive explanation

- Explaining why P happened rather than Q.
- Fact and foil
  (P is the fact, Q the foil)

Examples:

Contrastive explanation

- Explaining why P happened rather than Q.
- Fact and foil
  (P is the fact, Q the foil)

Examples:

- Why did I go to London rather than Paris?

Contrastive explanation

- Explaining why P happened rather than Q.
- Fact and foil
  (P is the fact, Q the foil)

Examples:

- Why did I go to London rather than Paris?

- Why did Clara rather than Johanne sneeze?

Contrastive explanation

- Explaining why P happened rather than Q.
- Fact and foil
  (P is the fact, Q the foil)

Examples:

- Why did I go to London rather than Paris?
- Why did Clara rather than Johanne sneeze?
- Why did the model predict *cat* rather than *dog*?

Explanations in machine learning

Explanations in machine learning

"An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way." (Molnar, 2020, p. 31)

Explanations in machine learning

"An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way." (Molnar, 2020, p. 31)

Taxonomy of interpretability

- intrinsic interpretability
  - simple structures
- post-hoc interpretability
  - interpretation after training the model

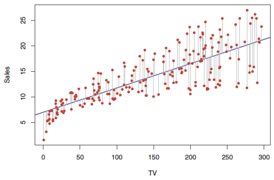
Linear models

Do linear models create good explanations?

"Linear models create truthful explanations, as long as the linear equation is an appropriate model for the relationship between features and outcome."

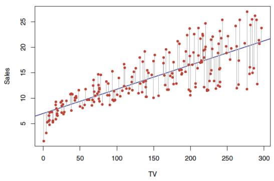
(Molnar, 2020, p. 63)

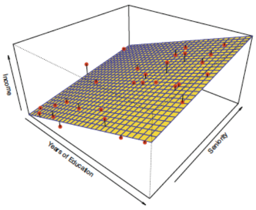
Linear regression



(Figure from James et al., 2013)

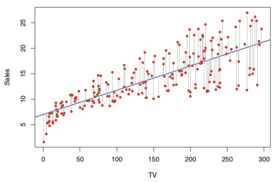
Linear regression

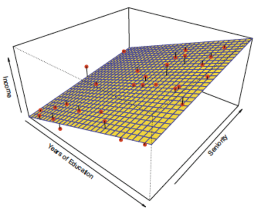




(Figure from James et al., 2013)

Linear regression



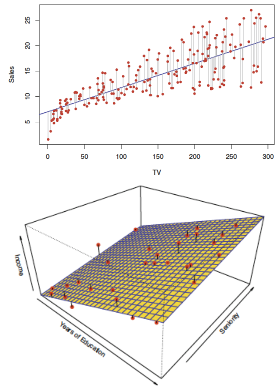


pros:

- weighted sum
- well known
- guarantee to find optimal weights

(Figure from James et al., 2013)

Linear regression



(Figure from James et al., 2013)

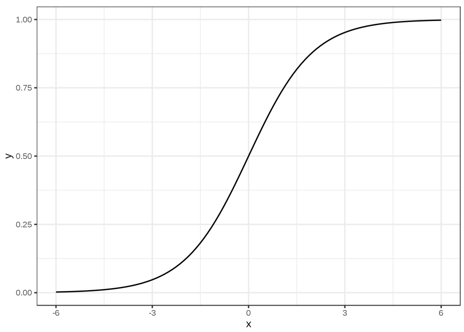
pros:

- weighted sum
- well known
- guarantee to find optimal weights

cons:

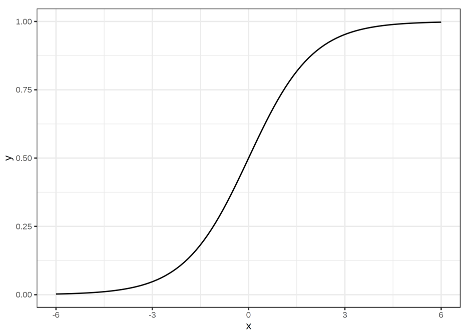
- can only represent linear relationships
- "interpretation of a weight can be unintuitive because it depends on all other features" (Molnar, 2020, p. 67)
- "Completely correlated features make it even impossible to find a unique solution" (Molnar, 2020, p. 68)
- interactions need to be handcrafted

Logistic regression



(Figure from Molnar, 2020)

Logistic regression

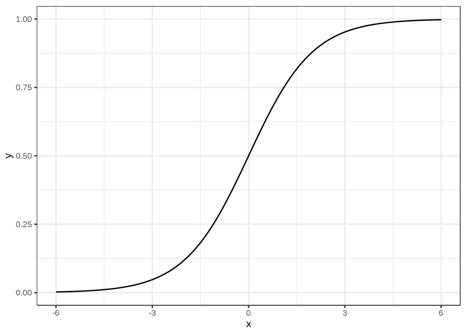


(Figure from Molnar, 2020)

pros:

- provide probabilities
- fast

Logistic regression



pros:

- provide probabilities
- fast

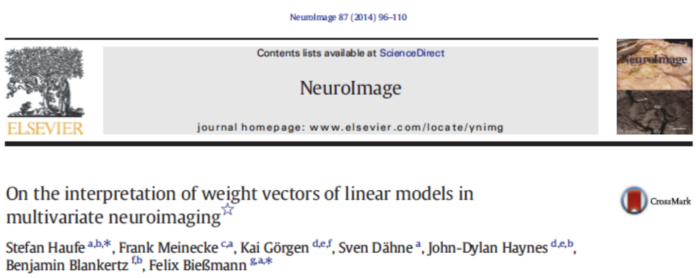
cons:

- "interpretation of the weights is *multiplicative* and not additive" (Molnar, 2020, p. 75, my italics)
- can only represent linear relationships
- interactions need to be handcrafted

(Figure from Molnar, 2020)

Filters and patterns

Haufe et al. 2014


NeuroImage 87 (2014) 96–110

Contents lists available at ScienceDirect

NeuroImage

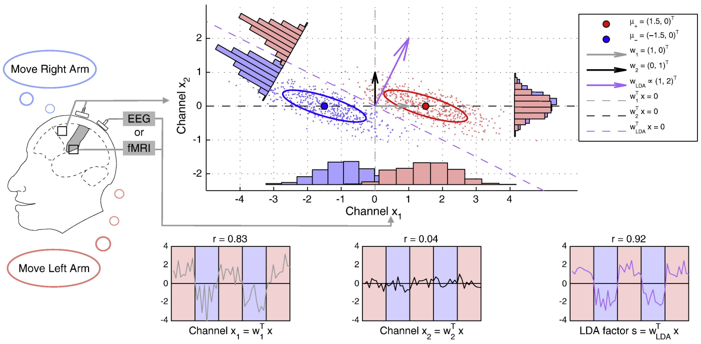
journal homepage: www.elsevier.com/locate/ynimg

ELSEVIER

On the interpretation of weight vectors of linear models in multivariate neuroimaging☆

Stefan Haufe [a,b,*], Frank Meinecke [c,a], Kai Görgen [d,e,f], Sven Dähne [a], John-Dylan Haynes [d,e,b], Benjamin Blankertz [f,b], Felix Bießmann [g,a,*]

On the interpretation of weight vectors . . .



(Figure from Haufe et al., 2014)

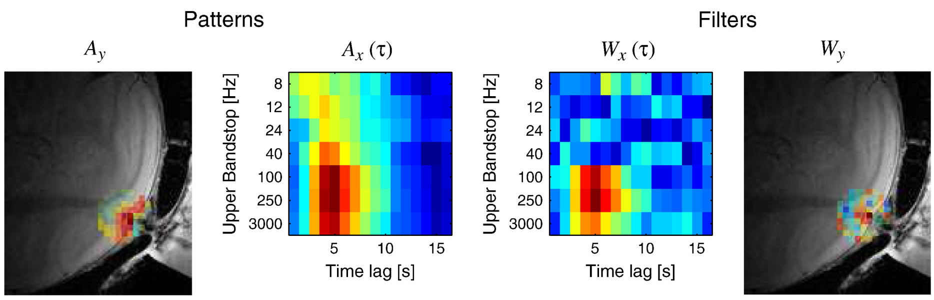
On the interpretation of weight vectors . . .

	Forward model	Backward model
Alternative name	Generative model	Discriminative model
Model (linear case)	$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \epsilon(n)$	$\mathbf{W}^{\top}\mathbf{x}(n) = \hat{\mathbf{s}}(n)$
Purpose	Factorize the data into *latent factors* $\mathbf{s}(n)$ and their corresponding *activation patterns* (columns of $\mathbf{A}$), plus noise $\epsilon\epsilon(n)$.	Extract *latent factors* $\hat{\mathbf{s}}(n)$ from the data by multiplying with *extraction filters* (columns of $\mathbf{W}$).
Interpretable	$\mathbf{A}$, $\mathbf{s}(n)$	$\hat{\mathbf{s}}(n)$
Supervised case	Encoding: Replace latent factors $\mathbf{s}(n)$ by known external target variables $\mathbf{y}(n)$ or pre-estimated factors $\hat{\mathbf{s}}(n)$. Thus, estimate how $\mathbf{y}(n)$ or $\hat{\mathbf{s}}(n)$ are *encoded* in the measurement.	Decoding: Seek latent factors $\hat{\mathbf{s}}(n)$ to approximate known external target variables $\mathbf{y}(n)$. Thus, estimate how $\mathbf{y}(n)$ can be *decoded* from the measurement.

(table from Haufe et al., 2014)

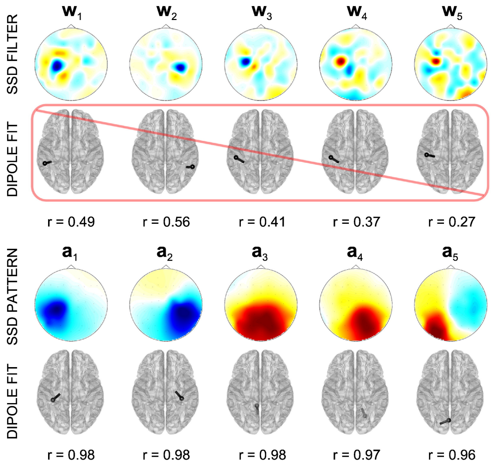
$x(n)$	M-dimensional vector of observed data
A	M $\times$ K matrix of patterns in forward models
W	M $\times$ K matrix of filters in backward model
$s(n), \hat{s}(n)$	K-dimensional vector of latent factors

On the interpretation of weight vectors . . .



(Figure from Haufe et al., 2014)

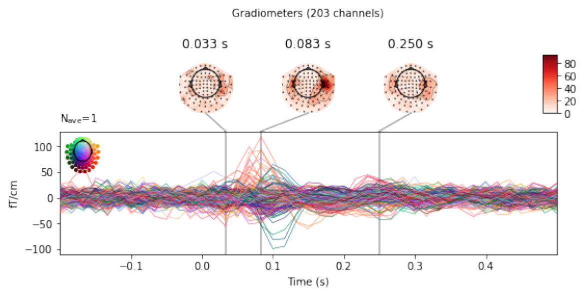
On the interpretation of weight vectors . . .



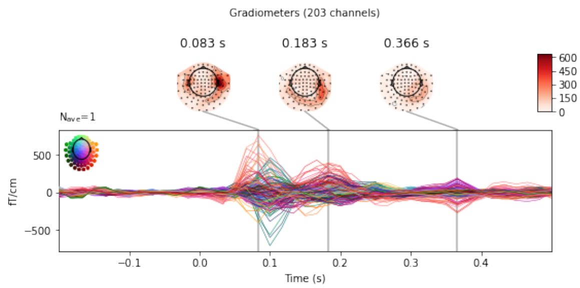
(Figure from Haufe et al., 2014)

On the interpretation of weight vectors . . .

filters (coefficients/weights)

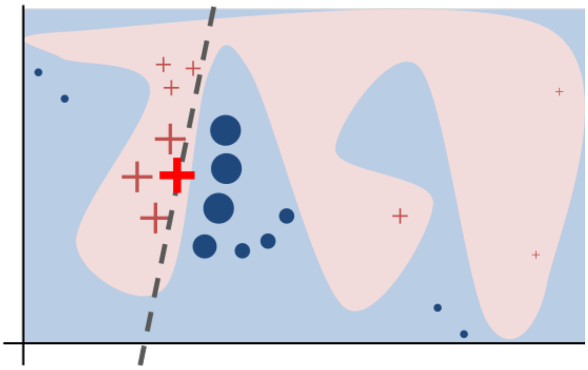


patterns



Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME)



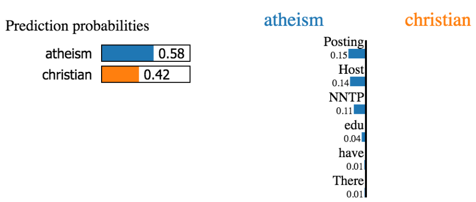
(Figure from Ribeiro et al., 2016)

Local Interpretable Model-agnostic Explanations (LIME)

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

(Figure from Ribeiro et al., 2016)

Local Interpretable Model-agnostic Explanations (LIME)



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

(Figure from https://github.com/marcotcr/lime)

Local Interpretable Model-agnostic Explanations (LIME)

Explaining prediction of 'Cat' in pros and cons

(Figure from https://github.com/marcotcr/lime)

Local Interpretable Model-agnostic Explanations (LIME)

(a) Husky classified as wolf

(Figure from Ribeiro et al., 2016)

Local Interpretable Model-agnostic Explanations (LIME)

(a) Husky classified as wolf

(b) Explanation

(Figure from Ribeiro et al., 2016)

Feature selection

Feature selection

Example: MNE sample data

sensor space:

- 102 magnetometers, 204 gradiometers
- downsampled to 60 Hz
- X = (123 * 306 * 43)
- X has 13.158 features in each row
  and 1,613,145 data points in total

source space:

- 5124 source space points
- downsampled to 60 Hz
- X = (123 * 5124 * 43)
- X has 220.332 in each row
  and 27,100,836 data points in total

Feature selection

before fitting

after fitting

Feature selection

before fitting

- variance thresholding
- univariate feature selection
  - select k best features
  - select percentile
  - χ^2, f-test

after fitting

Feature selection

before fitting

- variance thresholding
- univariate feature selection
  - select k best features
  - select percentile
  - χ^2, f-test

after fitting

- select based on weights/coefficients
- recursive feature elimination
- model based:
  - l1-based feature selection
  - feature importance from a tree based model

Questions?

1. Interpretability
2. Explanations
3. Explanations in machine learning
4. Linear models
   - Linear regression
   - Logistic regression
5. Filters and patterns
6. Local Interpretable Model-agnostic Explanations (LIME)
7. Feature selection

References I

Bird, A. (2003). *Philosophy of science* (Reprinted.). Routledge.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. faculty.marshall.usc.edu/gareth-james/ISL/

King, J. R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., Larson, E., & Gramfort, A. (2018). Encoding and Decoding Neuronal Dynamics: Methodological Framework to Uncover the Algorithms of Cognition. 19. https://hal.archives-ouvertes.fr/hal-01848442

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). Blackwell.

References II

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv: 1602.04938 [cs, stat]. Retrieved May 20, 2018, from http://arxiv.org/abs/1602.04938