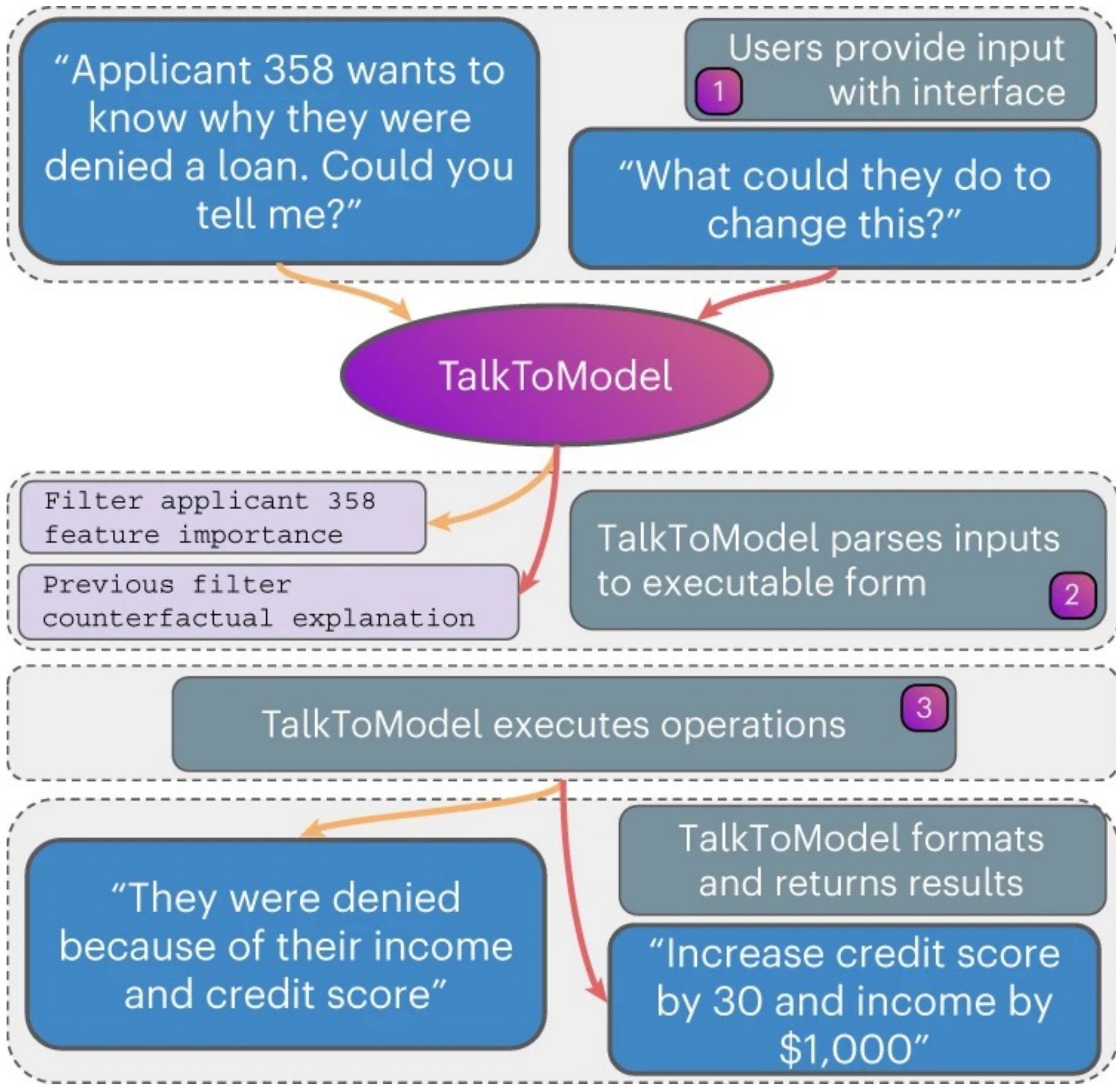
Lecture 15: Explainable AI

Mark Hasegawa-Johnson



Outline

- Rationale; Definitions of terms
- Regulations (GDPR etc)
- Post-hoc explanation
  - Layer-wise relevance propagation
- Explainability by design
  - Bayesian networks
- Intrinsic limitations on explainability

Explainable AI (XAI)

- An explainable AI (XAI) is an AI over which it is possible for humans to retain intellectual oversight.
- Rational = supported by reasons. XAI is rational if it can support its decisions by reasons.

Interpretability vs. Explainability

Term	Definition	Source
Interpretability	*"level of understanding how the underlying (AI) technology works"*	ISO/IEC TR 29119-11:2020(en), 3.1.42[35]
Explainability	*"level of understanding how the AI-based system ... came up with a given result"*	ISO/IEC TR 29119-11:2020(en), 3.1.31[35]

Source: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

Rationale: Should an AI explain itself?

- Should an AI explain why it reached any particular decision?
- Why? Why not?
- Can current AI explain its reasons?
- If not, what can be done to improve the situation?

Outline

- Rationale; Definitions of terms
- Regulations (GDPR etc)
- Post-hoc explanation
  - Layer-wise relevance propagation
- Explainability by design
  - Bayesian networks
- Intrinsic limitations on explainability

Regulatory frameworks

Regulations about explainability seek to avoid the harms of unexplained decisions by granting individuals a ”right to an explanation.”

- Europe: General Data Privacy Regulation (GDPR), 2018
  - Legally binding on everyone operating in EU
  - Penalties for violation can be extremely severe

GDPR Right to Explanation

The European Union's "General Data Protection Regulation" (GDPR) Article 15 specifies that:

> The data subject shall have the right to obtain … confirmation as to whether personal data concerning him or her are being processed, … access to the personal data, … the existence of automated decision-making, and … meaningful information about the logic involved.

Outline

- Rationale; Definitions of terms
- Regulations (GDPR etc)
- Post-hoc explanation
  - Layer-wise relevance propagation
- Explainability by design
  - Bayesian networks
- Intrinsic limitations on explainability

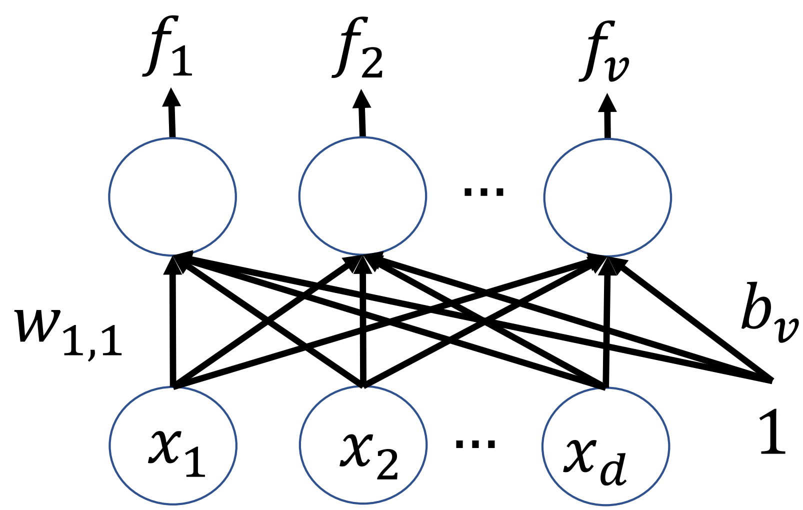
Post-hoc explanation

A “post-hoc explanation” is an algorithm that explains an AI’s decision after the decision has already been made.

Example: Logistic regression

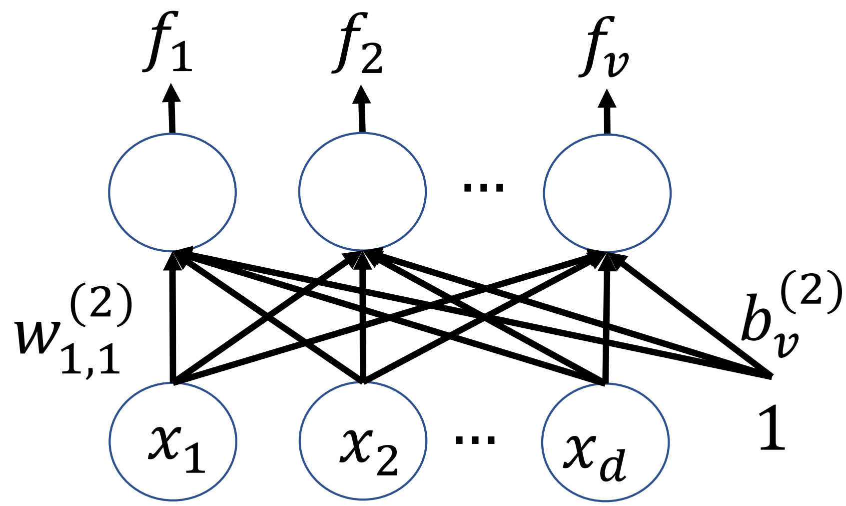
$$\boldsymbol{f} = \text{softmax}(\boldsymbol{z})$$

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$$



Explanations by analyzing the processing of a neural network

- $\boldsymbol{x}$ =binary indicator vector specifying the courses you've taken
- $\boldsymbol{f}$ =probability vector, f_k =probability that you should go into career k
- Suppose the neural net tells you that you should become a tiktok influencer. You might want to know why the neural net made that decision.

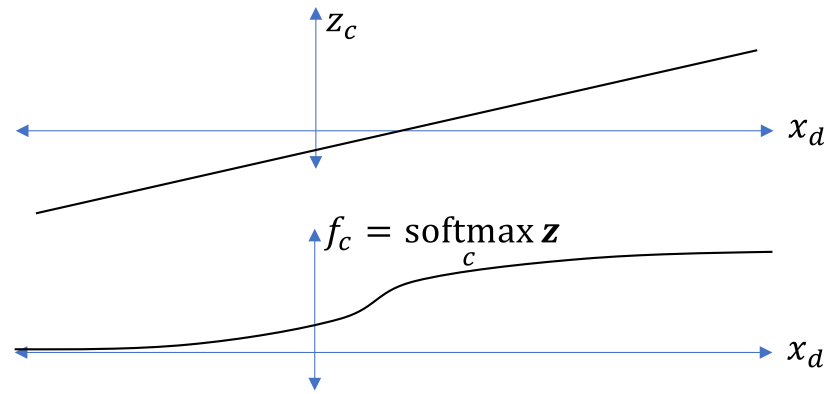


Gradient-based relevance

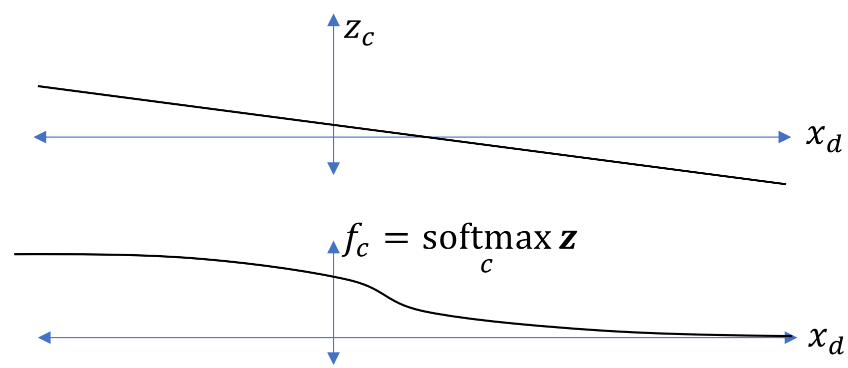
To what extent did feature x_d contribute to the network's decision?

- Is the slope positive or negative?
- Is x_d positive or negative?
  - More precisely: is x_d larger than or smaller than its expected value?

Example of an output f_c that gets larger in response to **<u>increases</u>** of the input x_d

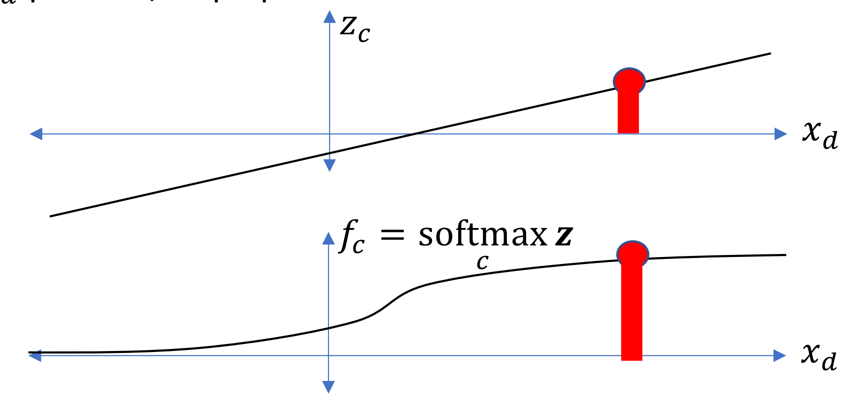


Example of an output f_c that gets larger in response to **<u>decreases</u>** of the input x_d

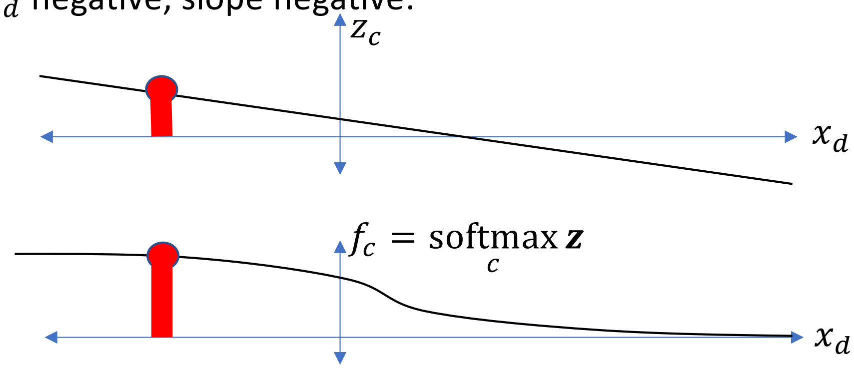


These are input features x_d that caused f_c to be **larger** than it would have been if $x_d = 0$:

- x_d positive, slope positive:

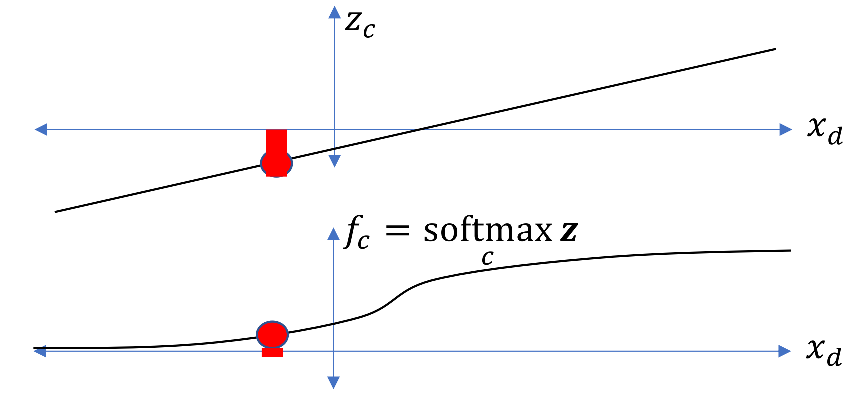


- x_d negative, slope negative:

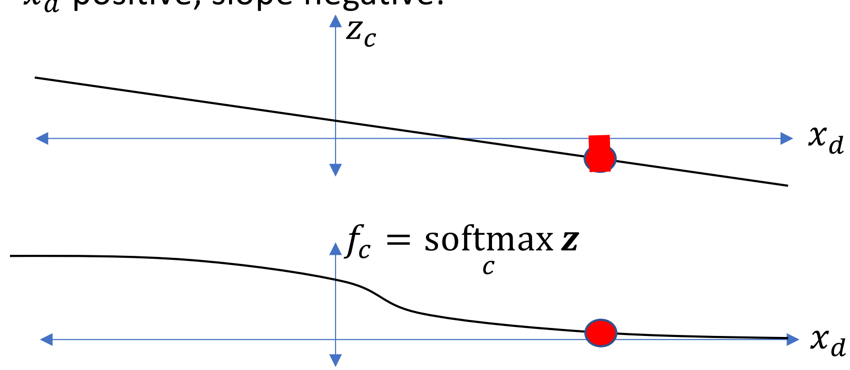


These are input features x_d that caused f_c to be **smaller** than it would have been if $x_d = 0$:

- x_d negative, slope positive:



- x_d positive, slope negative:



Relevance scoring in neural networks

- If $\operatorname{sign}\left(\frac{\partial z_c}{\partial x_d} \cdot x_d\right) = \operatorname{sign}(z_c)$, then feature x_d has **supporting** relevance to the neural net's output decision f_c
- If $\frac{\partial z_c}{\partial x_d} \cdot x_d = 0$, then feature x_d has **zero** relevance to the neural net's output decision f_c
- If $\operatorname{sign}\left(\frac{\partial z_c}{\partial x_d} \cdot x_d\right) = -\operatorname{sign}(z_c)$, then feature x_d has **opposing** relevance to the neural net's output decision f_c

Layer-wise relevance propagation

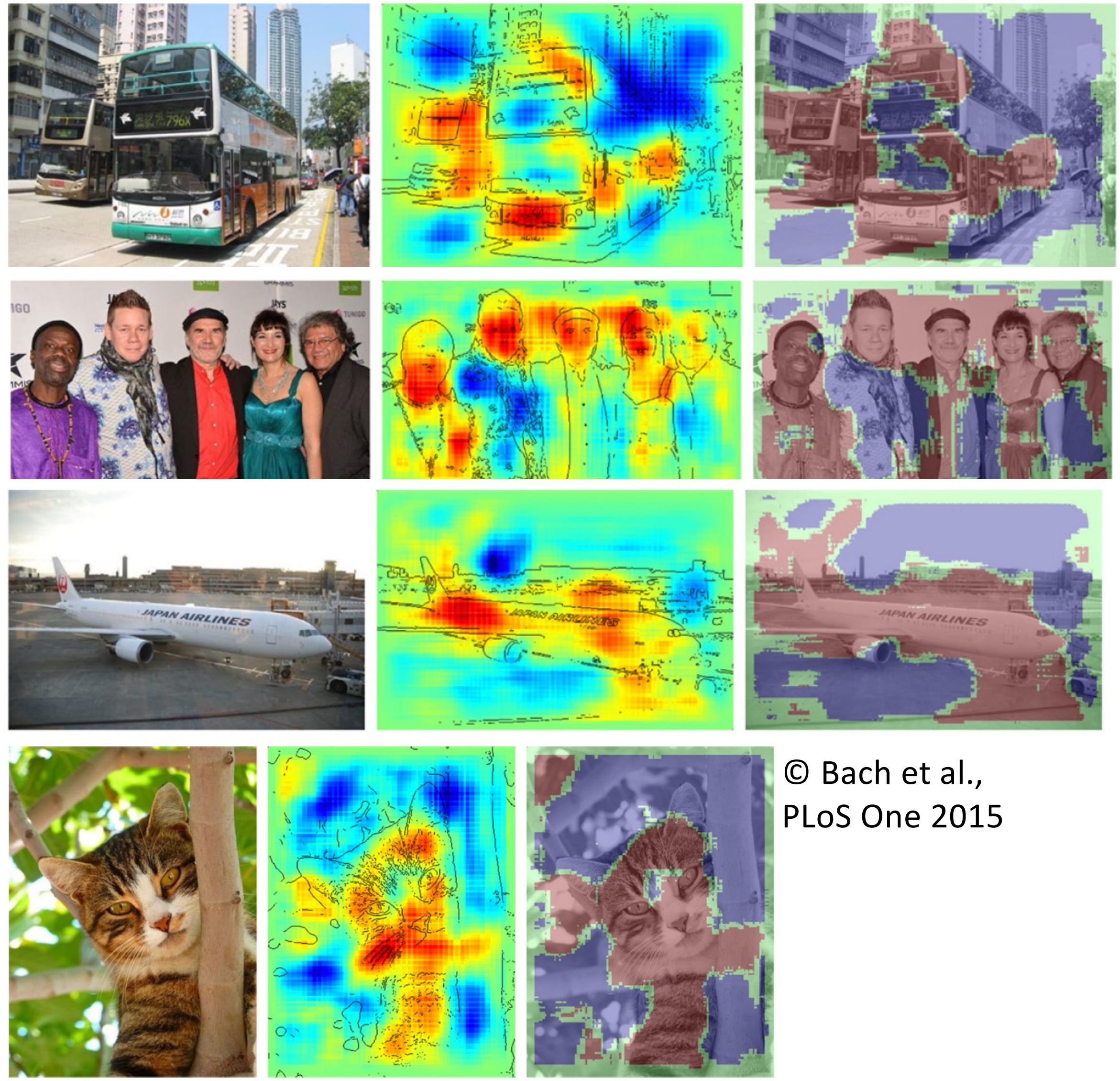
- Relevance of feature x_d to decision f_c:

$$R_{c,d} = \frac{\frac{\partial z_c}{\partial x_d} \cdot x_d}{\sum_{d'} \frac{\partial z_c}{\partial x_{d'}} \cdot x_{d'}} \cdot R_c$$

- Summation in the denominator ensures that $\sum_d R_{c,d} = R_c$, like "apportioning praise" or "apportioning blame."

Layer-Wise Relevance Propagation
(Bach et al., 2015)

- In LRP, relevance is normalized then back-propagated, layer by layer. This causes the smoothness you see here.
- Positive relevance: red, Negative: blue
- 2nd image: scaled,
- 3rd image: binary



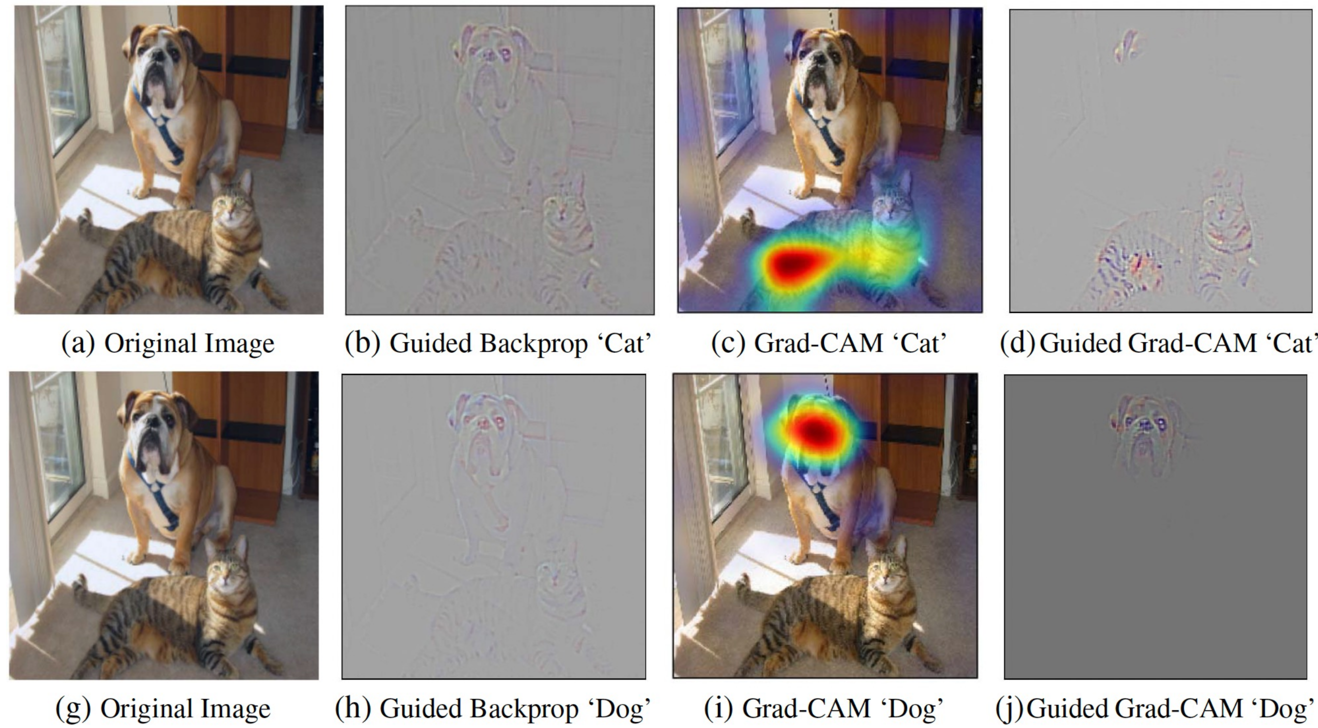
Quiz

Try the quiz!

Positive-only relevance scoring: Grad-CAM

Many relevance scoring systems keep only positive relevance, i.e.,

$$L_{c,d} = \frac{\max\left(0, \frac{\partial f_c}{\partial x_d} \cdot x_d\right)}{\sum_{d'} \max\left(0, \frac{\partial f_c}{\partial x_{d'}} \cdot x_{d'}\right)}$$



- Gradient-weighted Class Activation Mapping (Grad-CAM) pools these scores over collections of nodes
- Guided Grad-CAM multiplies the pooled scores times individual pixel scores

Advantages and disadvantages of relevance-based explainability

Advantage:

- Explains which input features caused the neural to make the decision it made

Disadvantage:

- Does not provide a logical reason why those particular features caused the neural net to make the decision it made
- There may not be any logical reason! The neural net is just a linear classifier of nonlinear combinations of features, it may not be logical.
- The “reasons” given by LRP may be true, but are almost always incomplete

Outline

- Rationale; Definitions of terms
- Regulations (GDPR etc)
- Post-hoc explanation
  - Layer-wise relevance propagation
- Explainability by design
  - Bayesian networks
- Intrinsic limitations on explainability

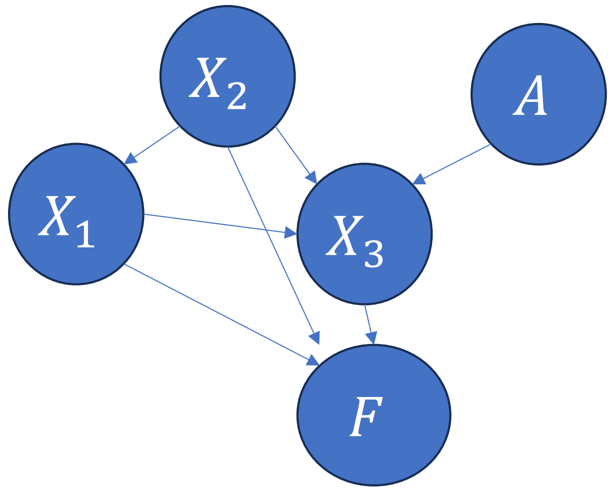
Decision-making algorithms that are explainable by design

Neural networks are not designed to be explainable. Other decision algorithms that are designed to be explainable include:

- Rule-based decision algorithms
- Decision trees
- Bayesian networks

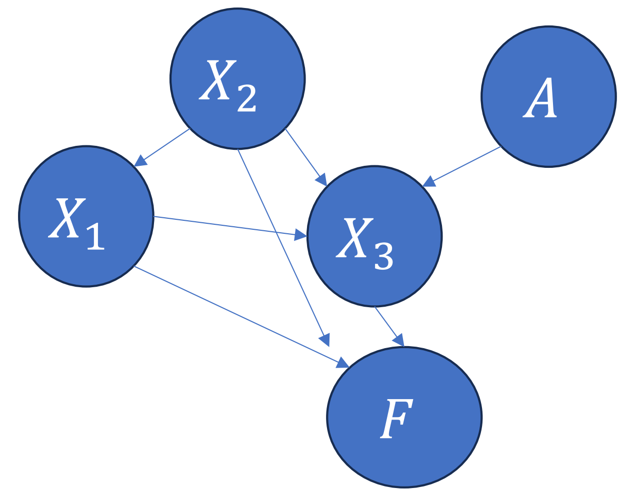
Bayesian networks and Counter-factual reasoning: Example

- X is a vector of relevant features including X_1 =GRE, X_2 = GPA, X_3 = letters of recommendation
- A is an irrelevant feature, for example, your height in centimeters.
- These variables have a complicated interdependence, shown here.
- For a given $X = \boldsymbol{x}, A = a, f(\boldsymbol{x}, a) = 0$, meaning you have not been admitted to grad school.
- The Bayes network shows the situation in which your height has no effect on your admission to grad school. How can we tell if the real world is correctly represented by this Bayes network?
- Counter-factual reasoning: Keep X_1 through X_3 the same, change the value of only A. Find out: does the outcome change?



Bayesian networks and Counter-factual reasoning: Equation

A fully trained Bayesian network explains the relationship between A and $f(X, A)$ by testing to see whether $\mathrm{P}(F, X, A = a)$ and $P(F, X, A = \neg a)$ differ:



$$\max_{F,X} |P(F, X, A = a) - P(F, X, A = \neg a)| > \epsilon?$$

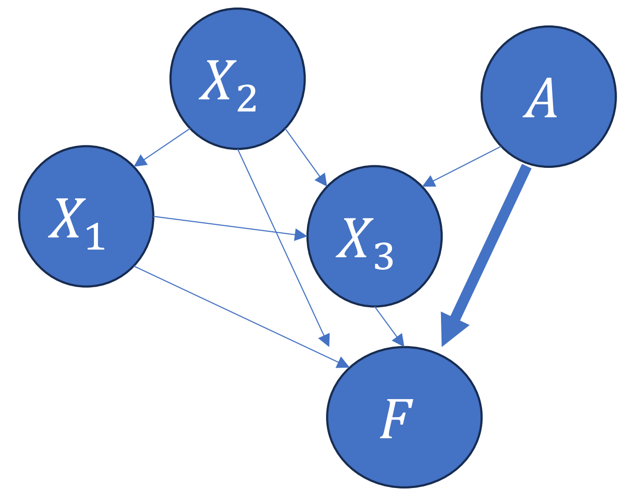
Bayesian networks and Counter-factual reasoning: Equation

If F is NOT conditionally independent of A given X, then you should add an edge to the Bayes network!

If

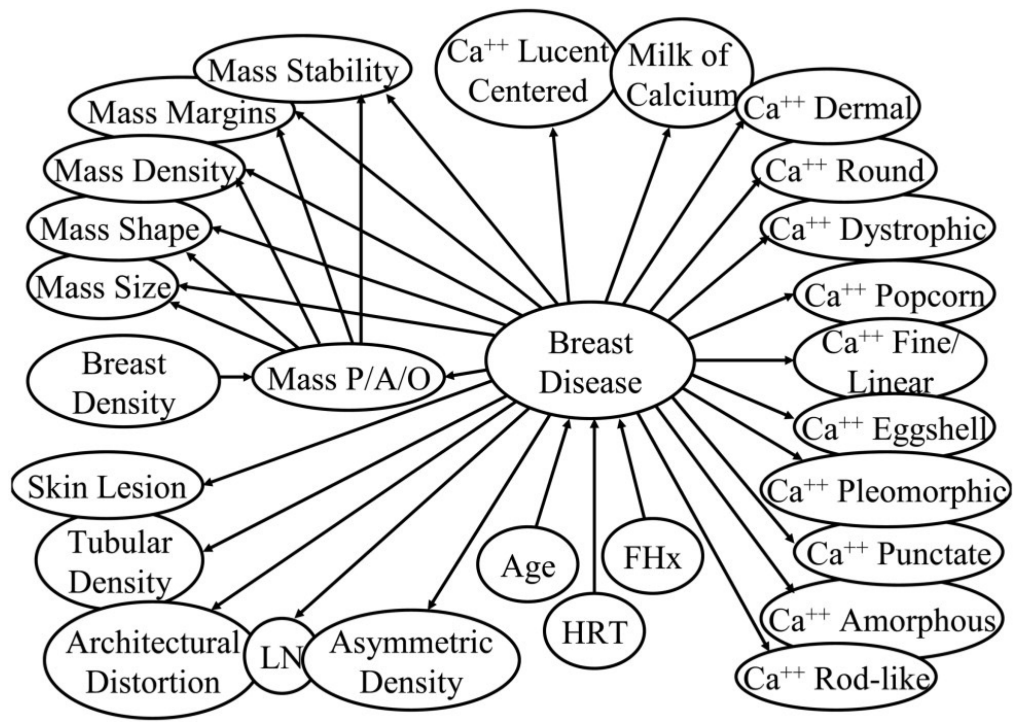
$$\max_{F,X} |P(F, X, A = a) - P(F, X, A = \neg a)| > \epsilon$$

…then add an edge.



Bayesian networks for explainable AI

- People who need to know why they are making a recommendation (e.g., doctors) are much more likely to accept this approach because:
- Each of the classifications made by a neural network (e.g., "Mass Stability") can be visually confirmed by the Radiologist
- The doctor can choose to ignore the final diagnosis ("Breast disease") if she disagrees with the reasons



Elizabeth Burnside, "Bayesian networks: Computer-assisted diagnosis support in radiology," 2005

Advantages and disadvantages of Bayes network explainability

Disadvantage:

- By forcing the network to depend on variables that make sense to a human being, we may reduce accuracy of the decision

Advantage:

- Decision is explainable by design
- Some end users will completely ignore an unexplainable decision (e.g., doctors). For such end users, an explainable decision is the only alternative.

Outline

- Rationale; Definitions of terms
- Regulations (GDPR etc)
- Post-hoc explanation
  - Layer-wise relevance propagation
- Explainability by design
  - Bayesian networks
- Intrinsic limitations on explainability

Intrinsic limits on explainability

- Complexity (e.g., Layer-Wise Relevance Propagation): A decision is reached by weighing partial contributions from many factors. If a particular combination of input factors has never occurred before, how do you decide if you have correctly weighted all the factors?
- Trust (e.g., Layer-Wise Relevance Propagation): An AI may reach a decision for reasons very different from those a human would use. How can you tell if the AI’s reasons are valid?
- Accuracy (e.g., Bayesian networks): By limiting the factors an AI can consider, you limit its potential accuracy.
- Hackers: If an AI is explainable, it may also be hackable.

Summary

- Relevance of input x_d to output z_c:

$$R_{c,d} = \frac{\partial z_c}{\partial x_d} \cdot x_d$$

  - Layer-wise relevance propagation LRP: Normalize so $\sum_d R_{c,d} = R_c$
  - Gradient-weighted class activation mapping Grad-CAM: Keep only positive relevances

- Counter-Factual Reasoning:

$$\max_{F,X} |P(F, X, A = a) - P(F, X, A = \neg a)| > \epsilon?$$

  - If the answer is yes, then you should add an edge between A and F