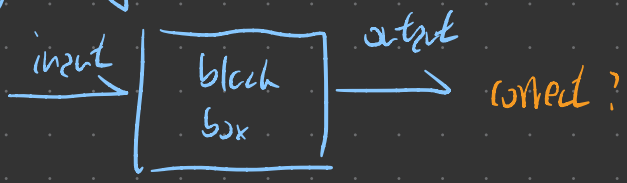
* Black Box ML.

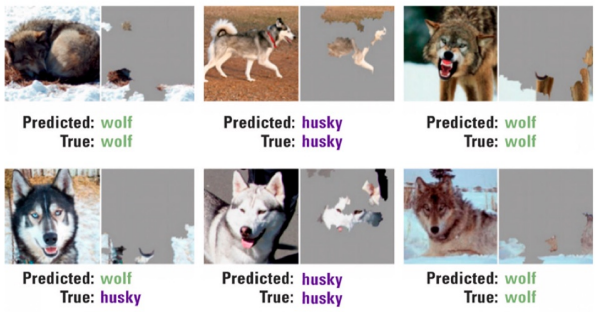
You can't confirm it's correct or actually wrong but with correct-like output in coincidence.

Ex:



"Husky vs. Wolf"

- Husky was classified as wolf in a image classification task. Why?
  - Because the training data of the wolves all have snowy background



https://www.compact.nl/articles/deep-learning-finding-that-perfect-fit/

(a) Husky classified as wolf (b) Explanation

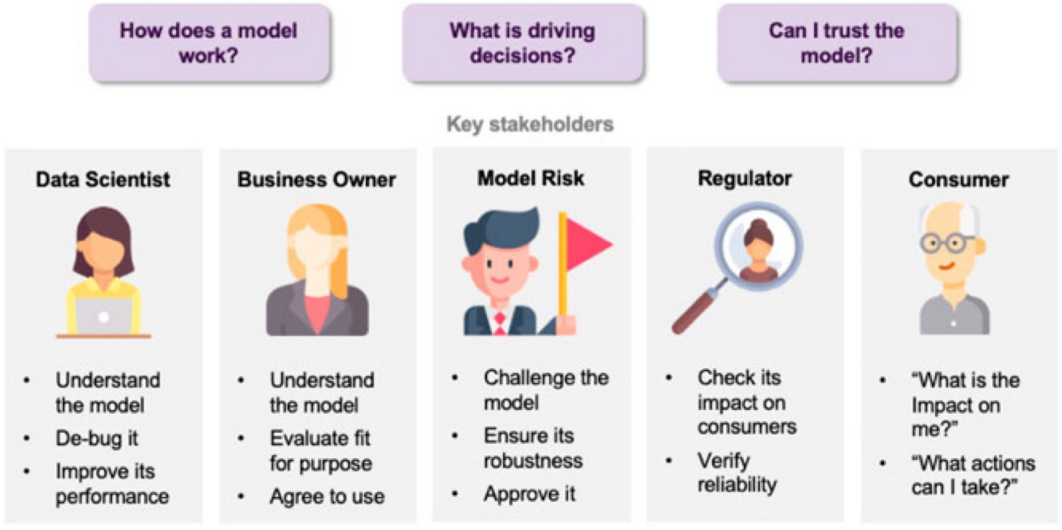
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

https://arxiv.org/pdf/1602.04938.pdf

↙ Model learning with wrong data.

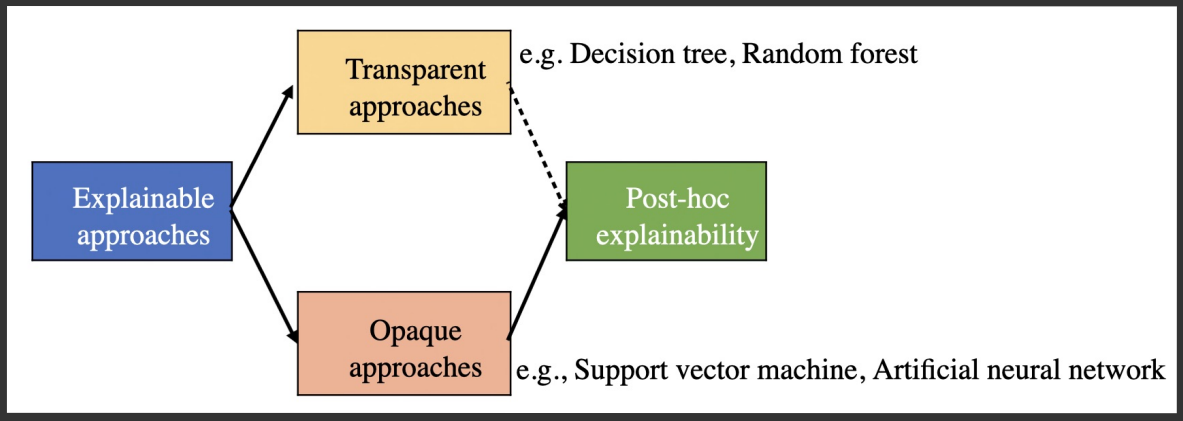
* The application requires explainable machine learning.
  - Medical dignosis
  - Self-driving car
  - Credit or loan

Benefits of explainable machine learning



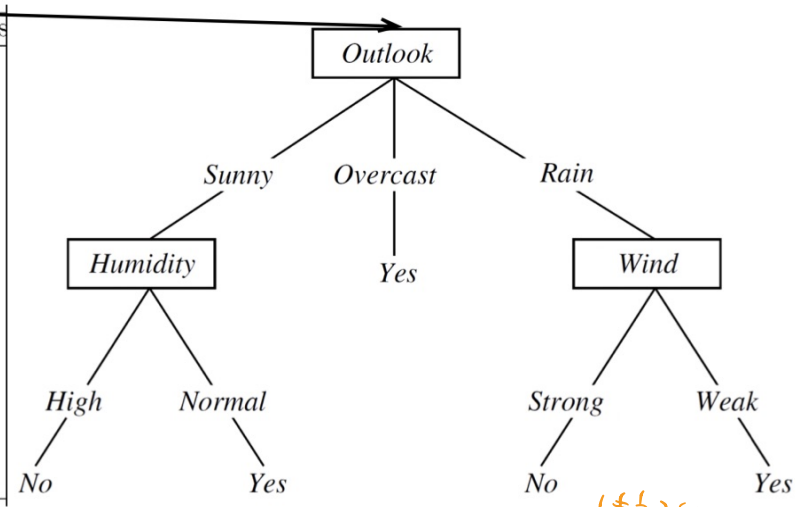
https://www.frontiersin.org/articles/10.3389/fdata.2021.688969/full

* Map of explainable approaches.



* Transparent model — Decision tree

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



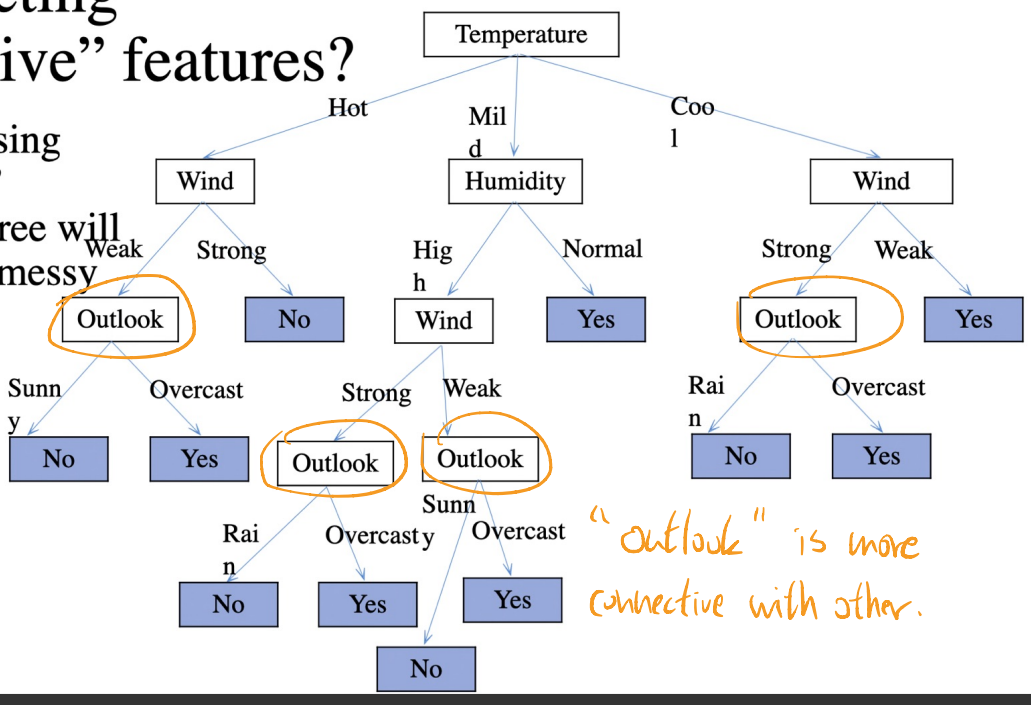
Mitchell, "Machine learning", 1997

informative adj. 見聞廣博的

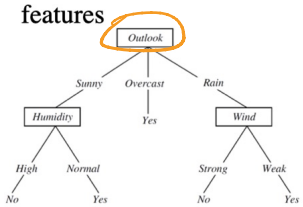
Tree is built by evaluating the informative degree of features.

Why selecting "informative" features?

- Without choosing "informative" features, the tree will become very messy



This tree is much simpler if constructed by the "informativeness" of features



"outlook" is more connective with other.

Def. The ID3 Algorithm

$Entropy(S) = -\sum_i p_i \log_2 p_i$

Ex $-\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \doteq 0.940$

- For the final classification results (that is, PlayTennis), there are 9 Yes and 5 No
  - The entropy is

$$Entropy(S) \equiv -\sum_i p_i \log_2 p_i$$
$$= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$
$$= 0.940$$

PlayTennis
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

Again, this represents how **pure** the classification results are

Def. Information Gain p_i : purity

The expected reduction of entropy caused by partitioning the examples.

$$InfoGain(S, A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{S} Entropy(S_v)$$

S_v : The entropy of subset combined by chosen values of A

Ex $InfoGain(S, Outlook) = Entropy(S) - (Entropy(Sunny) + Entropy(Overcast) + \cdots)$

$= 0.940 - \frac{5}{14}0.97 - \frac{4}{14}\cdot 0 - \frac{5}{14}\cdot 0.97$

* Feature importance for decision tree

By walking through the tree can see how decisions are made.

* Limitation of ID3 & Information Gain

- Instead of the values consisted by a lot of unique IDs,

ID3 favors features with a large num. of values.

Def. C4.5

To solve the limitation of ID3, like unique IDs, continous values, missing info.

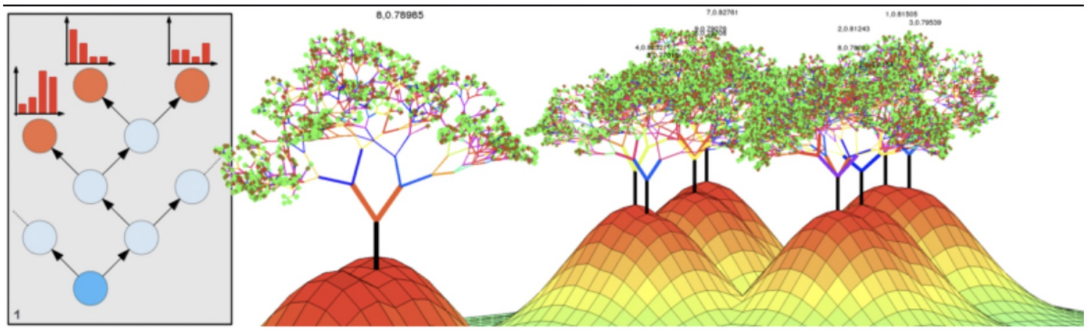
Def. Gain Ratio

$$GainRatio(S, T) = \frac{InfoGain(S, T)}{SplitInfo(S, T)}$$

where $SplitInfo(S, T) = \sum_{i=1}^{n} P'\left(\frac{i}{S}\right) \cdot \log\left(P'\left(\frac{i}{S}\right)\right)$

* Another transparent model — Random Forest.

- Random forest is another model that is used more frequently than decision tree
  - The advantage of random forest is that, by utilizing many decision trees, the "forest" can predict more accurately than a single tree



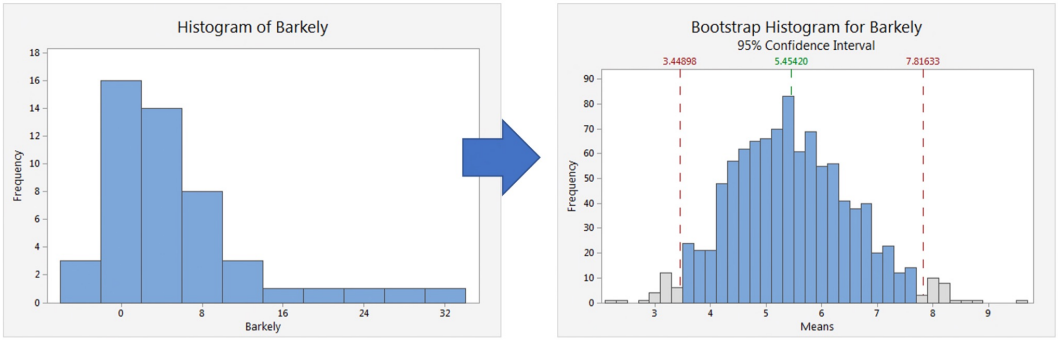
https://medium.com/@rdhawan201455/random-forest-and-how-it-works-67f408e43a43

* Bagging

To redistributing population by multiple times

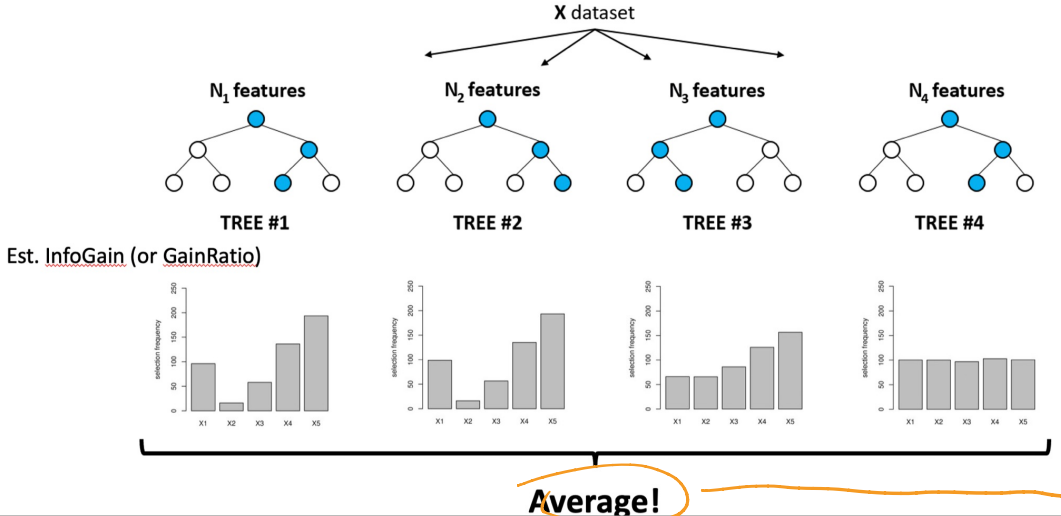
Why bootstrapping?

- Because sample distribution is sometimes (usually?) distributed in a non-normal way
- We can use **bootstrap** and the **Central Limit Theorem** to estimated the sample mean and variance



Via Central Limitation Thm. let result be more accurate.

Random forest feature importance



Result by multiple tree's analysis.

4 Features relevance / redundancy

Highly-correlated features

Features
Age
Age (Integer)
IQ
Mother's IQ
Father's IQ
Height
Partner's height

→ Income Prediction

They are too relative.
We just need one of them, so the other features are redundant

redundant adj. 冗余的

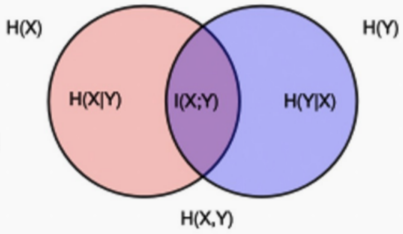
Def. Minimal Redundancy Maximal Relevance Algorithm. (mRMR)

- Score features based on their relevance to the labels but with minimum relevance to other features

$$f^{mRMR}(X_i) = I(Y, X_i) - \frac{1}{|S|}\sum_{X_s \in S} I(X_s, X_i)$$

Y: tgt
X_s: others.

Mutual information – mutual dependencies between the two variables

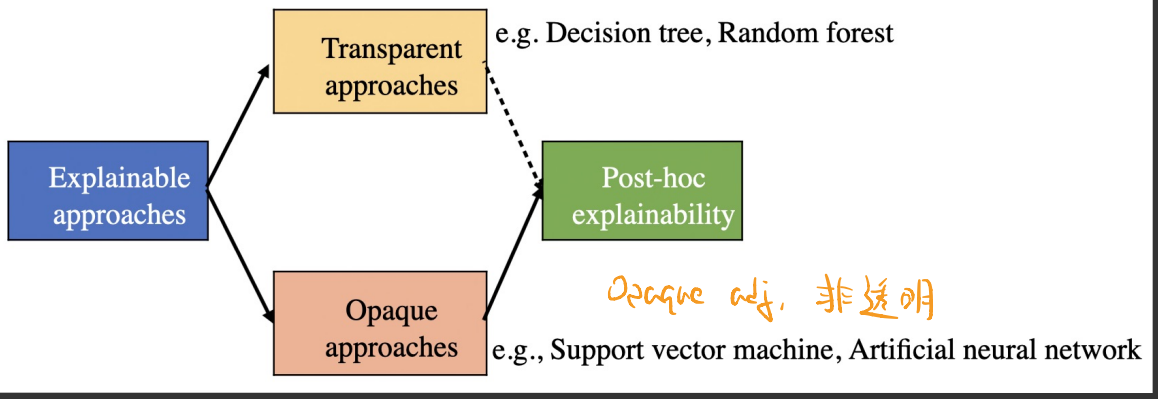


https://arxiv.org/pdf/1908.05376.pdf

Mutual adj. 互有的

AI Applied. / Sec 4. Explainable Machine Learning / week 02

* Map of explainable approaches.



Opaque adj. 非透明

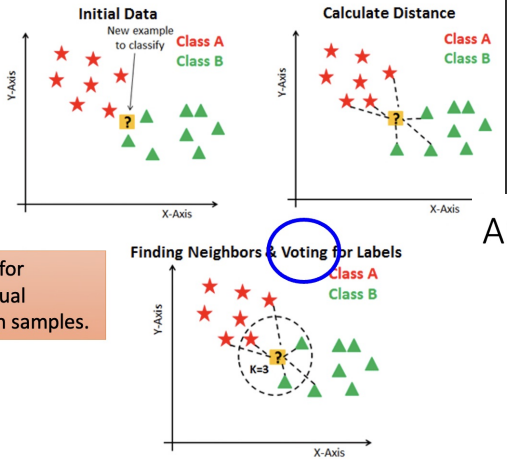
& Why approaches opaque ?

Based on sample relevance / redundant, nonlinearly distributed.

Ex k nearest neighbor (KNN), Support Vector Machine (SVM)

Artificial Neural Network (ANN).

Why kNN is "opaque?



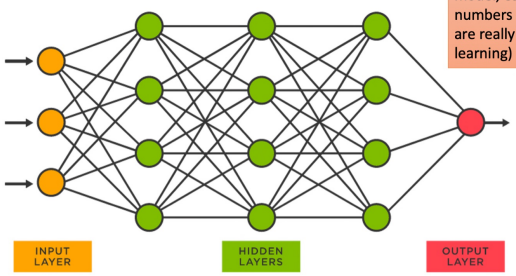
Because the features are used for calculating "distances"; the actual classification was conducted on samples.

https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

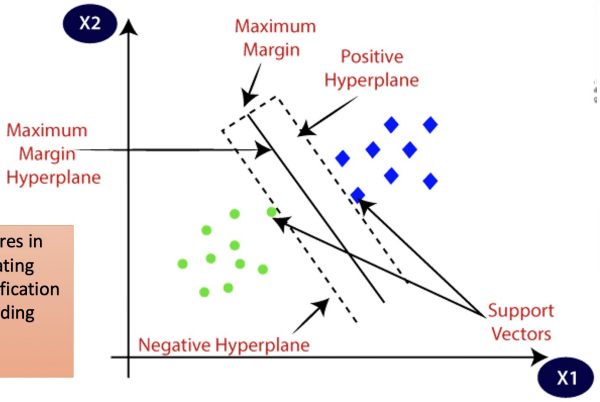
nonlinear.

Artificial neural network (ANN)

The reason why ANN is opaque is because of the highly non-linear feature weight combination in the model, especially when the numbers of nodes and layers are really high (such as deep learning)



Why SVM is "opaque"?



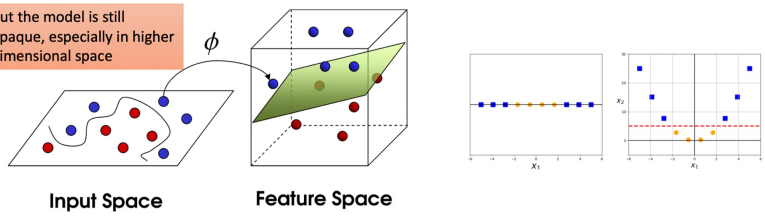
Similarly to kNN, features in SVM is used for calculating distances; actual classification was conducted on building the hyperplane with maximum margin

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

SVM kernel trick

- Very often one cannot find very good hyperplane for classification at the original feature space
- The "kernel trick" of SVM is to project the features to some higher degree space and attempt to find classification hyperplane at higher dimension

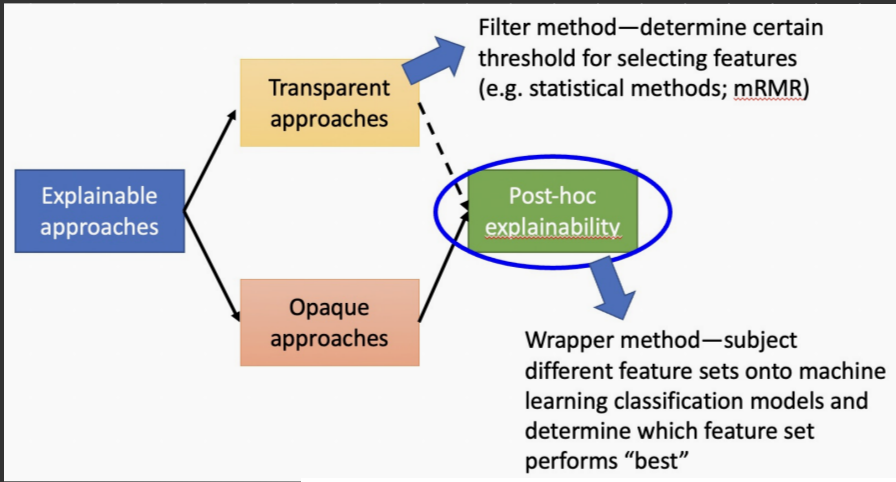
But the model is still opaque, especially in higher dimensional space



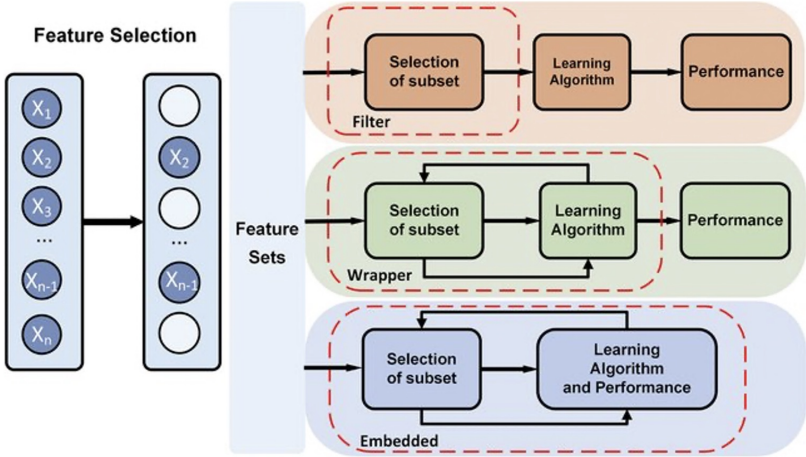
https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

AI Applied. / Sec 4. Explainable Machine Learning / week 02

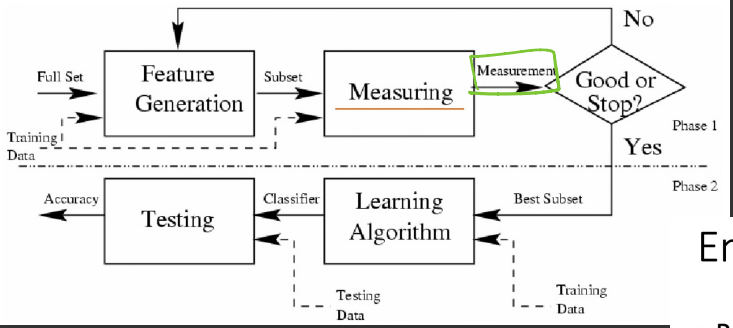
* Post-hoc approaches for features selection.



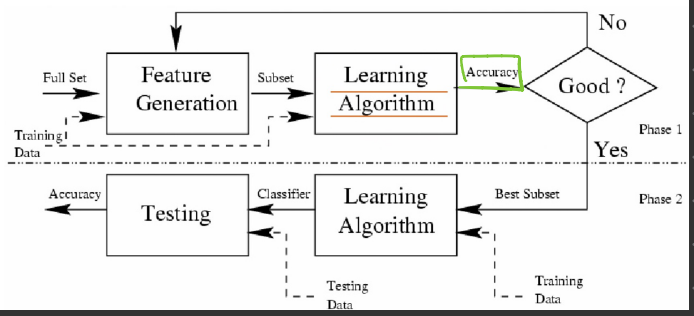
The feature selection classes again



Filter model



Wrapper model



Embedded method?

- Regularization methods such as Lasso, Ridge, or Elastic Net

Lasso L1 penalty	$argmin\left(\left(y-\beta_0-\sum\beta_iX_i\right)^2+\lambda\sum\lvert\beta_i\rvert\right)$
Ridge L2 penalty	$argmin\left(\left(y-\beta_0-\sum\beta_iX_i\right)^2+\lambda\sum\beta_i^2\right)$
Elastic Net L1+2 penalty	$argmin\left(\left(y-\beta_0-\sum\beta_iX_i\right)^2+\lambda\sum\lvert\beta_i\rvert+\lambda\sum\beta_i^2\right)$

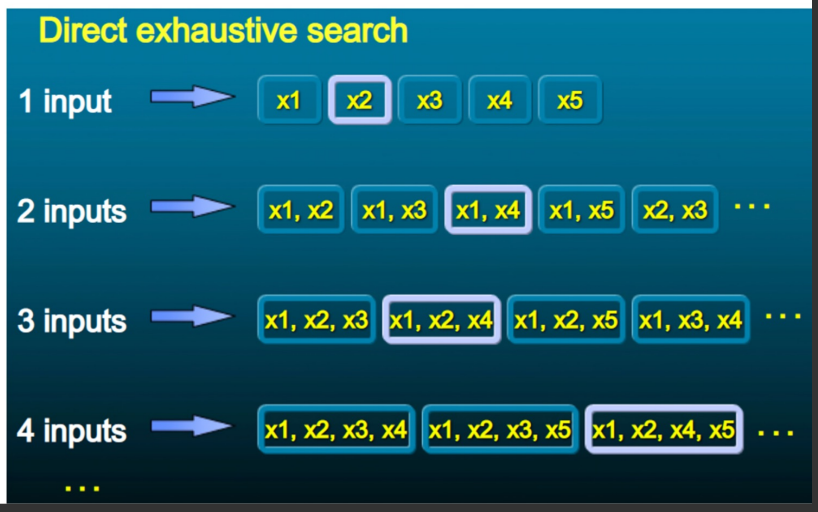
Using penalty, let performance affect result directly.

Wrapper Method types

- Exhaustive selection

Exhaustive selection

- Guaranteed to find the optimal feature set
- However the time spent on searching the optimal set is unrealistic

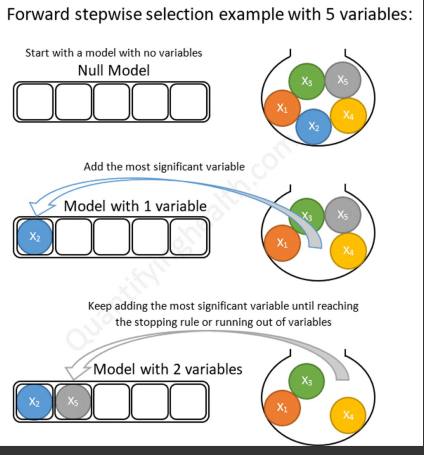


find optimal set violently.

- 正向表列 / 負向表列

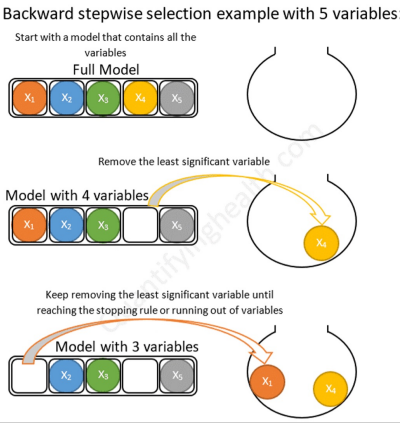
Forward selection

- Sort the features by importance or relevance values
- Add them into the model, one-by-one, and evaluate the performance of the ML model
- "Greedy"



Backward elimination

- Sort the features by importance or relevance values
- All all features to the model
- Eliminate the features, one-by-one, and evaluate the performance of the ML model
- Also "greedy"



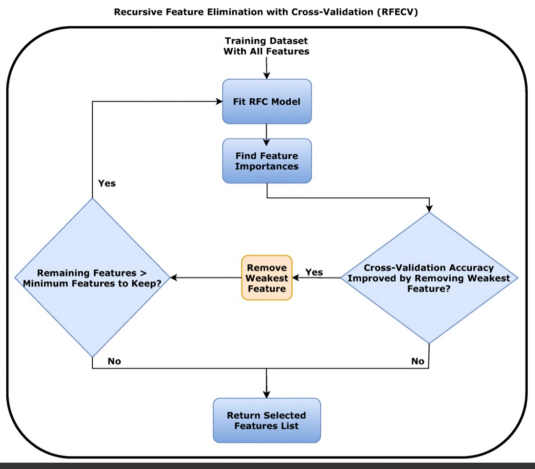
Def. Greedy Algorithm. (One of the types of algorithm)

Do the best choice for current iteration, usually not best sol. after all.

- RFECV

Recursive Feature Elimination with Cross Validation (RFECV)

- A backward selection methods with cross-validation for more robust evaluation



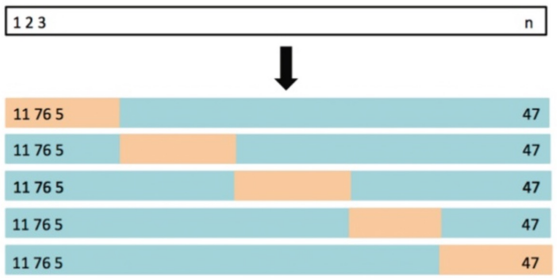
Cross - Validation

train set, test set ex: YOLOv5 requires train, test to learning.

- k fold cross validation

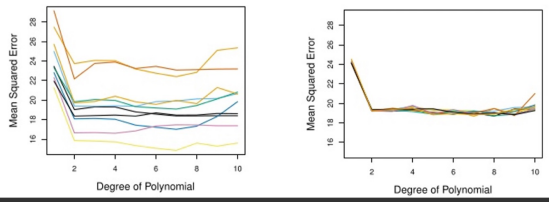
Instead of 2 sets likes a training & a test, Spliting dataset into k parts

- Split the data into k equal parts.
- In every training, take one part out and use the rest for training.
- The taken-out part can then be used for test purpose.



The k-fold is more stable

- Compared to the random-splitting of training and test datasets, the k-fold is much more stabilized.



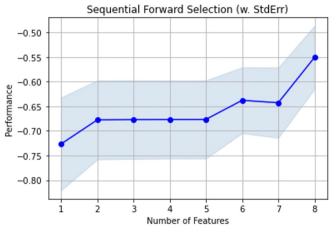
Wrapper Method types

- Bidirectional method.
  - Preparing 2 sets of feature evaluation model, one is forward & the other is backward.
  - Unrealistic since the logic of both method are same.
  - But implement an early-stop mechanism.

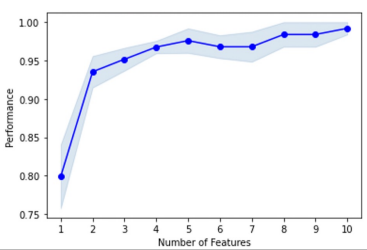
Early-stop mechanism

- Tricky to define (too arbitrary)

This one is easy

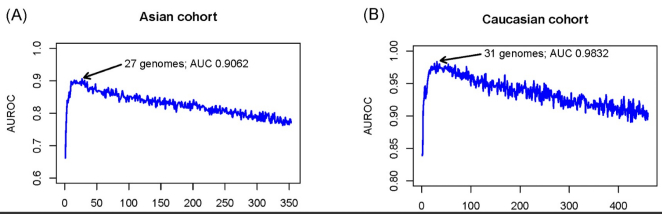


This one may cause some problem though if not "early-stopped" properly



A forward selection example

- I recently conducted an analysis on how bacteria living in our gut may be associated with colorectal cancer and may be used as predictor for potential diseases



- Model specific
  - Can only calculate feature importance using the specific model
  - E.g. decision tree, random forest
- Model agnostic
  - Can calculate feature importance score regardless of the classification model
  - e.g. Local Interpretable Model-Agnostic Explanations (LIME)

Local interpretable model-agnostic explanation

- The paper was published in KDD 2016
- A model-agnostic approach for evaluating feature importance

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

LIME intuition

- Local fidelity
- Sometimes it is difficult to evaluate the importance in terms of all features (that is, globally)
- The LIME method will generate local approximations and then evaluate whether the local features perform well in classifying instances

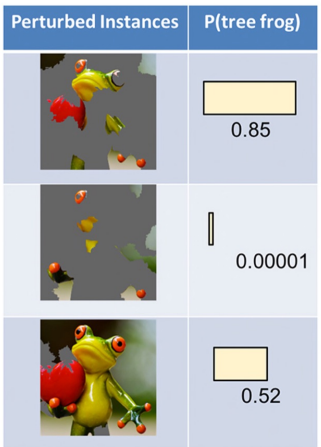
super-pixel

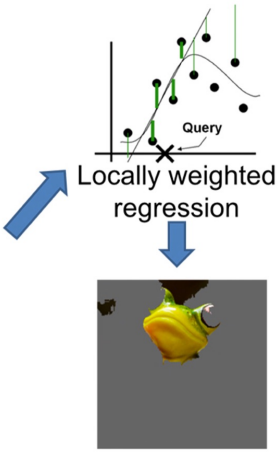
LIME intuition – an example of perturbation

LIME works by perturbing the very instances to check whether there is/are super-pixel combinations that works.

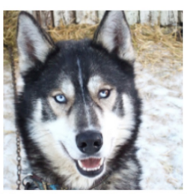
- Better if the probability of classification is better than the original one

Original Image
P(tree frog) = 0.54





https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Human trustworthiness before & after LIME

← Shapley Additive Explanations (SHAP)

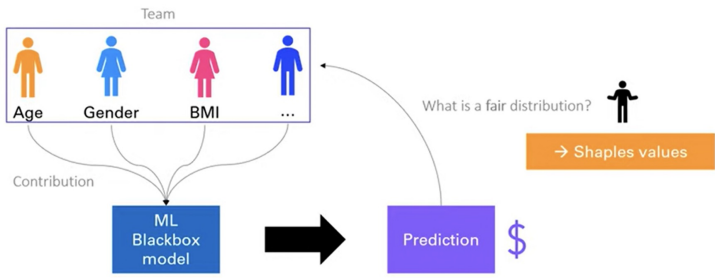
- The paper was published in NIPS 2017
- Also a model-agnostic approach for evaluating feature importance

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

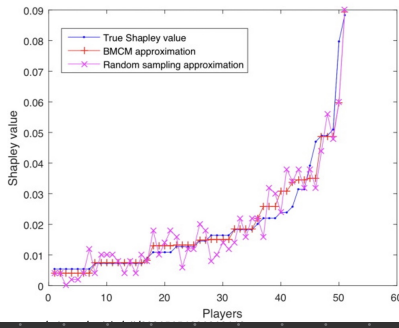
Features as players in SHAP design



Testing who distribute most.

Shapley sampling values

- The combinations can be approximated using sampling techniques such as random or Monte Carlo sampling



* Layer-wise Relevance Propagation (LRP)

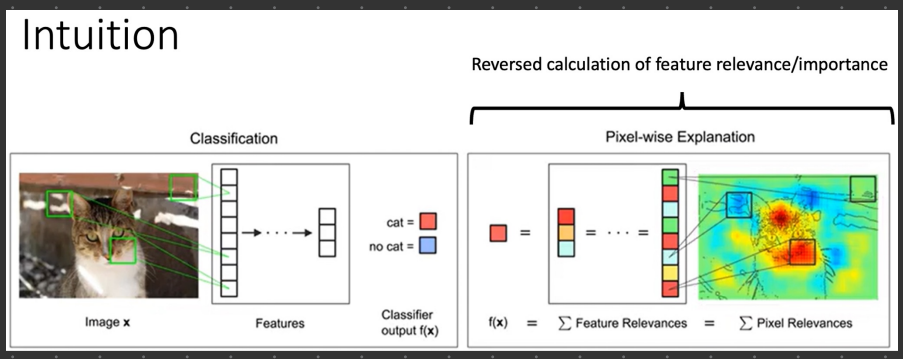
- Paper published at ICANN 2016
- A importance model designed specifically for neural network
  - Model-specific

Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers

Alexander Binder[1], Grégoire Montavon[2], Sebastian Bach[3], Klaus-Robert Müller[2,4], and Wojciech Samek[3]

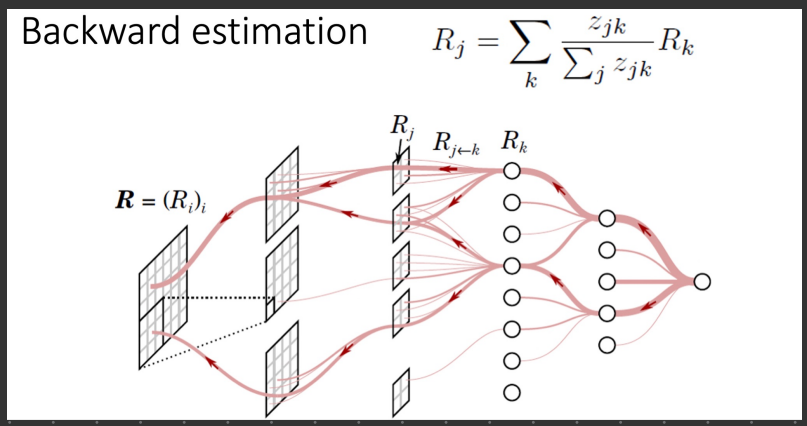
[1] ISTD Pillar, Singapore University of Technology and Design
[2] Machine Learning Group, Technische Universität Berlin
[3] Machine Learning Group, Fraunhofer Heinrich Hertz Institute
[4] Department of Brain and Cognitive Engineering, Korea University

Intuition



Backward estimation

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$$



4 Counterfactual Explanations

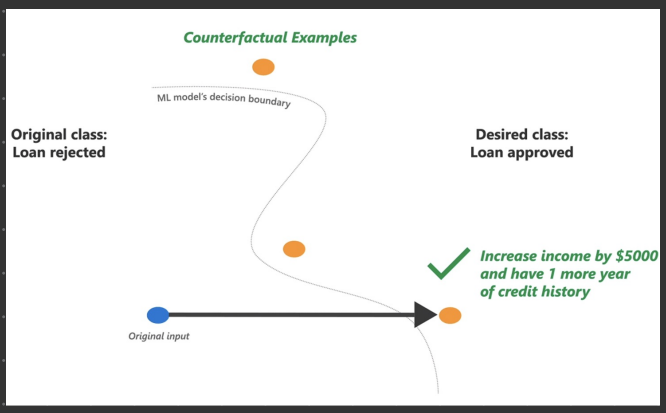
- Paper published on Harvard journal of Law and Technology, 2018

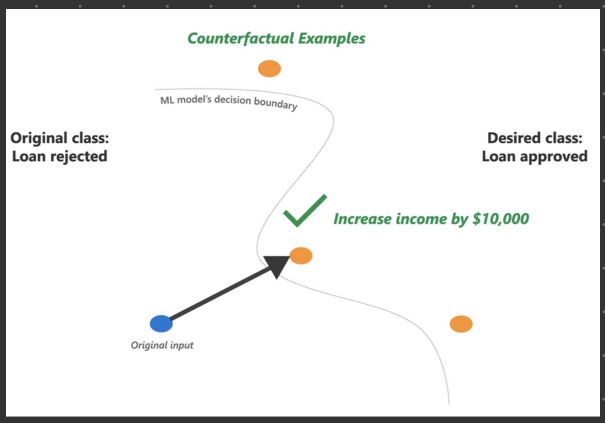
COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

Counterfactual?

- Definition: Contrary to fact – Merrian-Webster dictionary
- The "what if" statement
  - If I am in condition B, what can I do to "move" to condition A?





Find the bordary by searching.