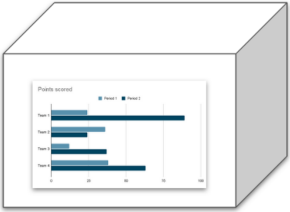
Interpretable Machine Learning

Interpretable Models



Learning goals

- What characteristics does an interpretable model have
- Why we should use interpretable models
- Examples for interpretable models: linear and polynomial regression models, generalized linear models, generalized additive models, model-based boosting, rule-based learning

MOTIVATION

- Achieving interpretability by using interpretable models is the most straightforward approach
- Classes of models deemed to be interpretable:
  - Linear regression
  - Logistic regression ($\rightsquigarrow$ Classification)
  - Decision trees
  - k-NN
  - Naive Bayes
  - ...

ADVANTAGES

- No further technique for interpretability required
  - reduces risk of bringing in another source of failure
- Since the models are often rather simple, training time is also fairly small
- Some of them fulfill the monotonicity constraint

  $\rightsquigarrow$ Larger feature values always lead to higher (or smaller) outcomes (e.g., regression values)
- Some models can also explain interaction effects

DISADVANTAGES

- Too simple models
  $\rightsquigarrow$ poor accuracy $\rightsquigarrow$ unusable in practice in the first place
- If too complex interactions are modelled, interpretability could suffer

FURTHER COMMENTS

- Some argue that one should always use interpretable models in the first place https://www.nature.com/articles/s42256-019-0048-x
  - …and not try to explain uninterpretable models posthoc
  - Can sometimes work out by spending enough time and energy
    on feature engineering and data cleaning

$\rightsquigarrow$ Drawback: Hard to achieve for data for which end-to-end learning is crucial
(e.g., images and text)

LINEAR REGRESSION

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \epsilon$$

- y output
- w_i weight of input feature x_i
- ϵ remaining error (e.g., because of noise)

$\rightsquigarrow$ model consists of $p + 1$ weights w_i

- Properties and assumptions:
  - linear
  - normality assumption of the target
  - homoscedastic (i.e., constant variance)
  - independence of features
  - fixed features (i.e., free of noise)
  - no strong correlation of features

INTERPRETATION OF LINEAR REGRESSION

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \epsilon$$

Let's consider different feature types:

- Numerical features: Increase of numerical value will lead to w_i times increased output
- Binary feature: Either weight w_i is active (1) or not (0).
- Categorical feature: One-hot-encoding of $L - 1$ new features for L categories
- Intercept w_0: reflects expected features values if features were standardised (0-mean, 1-stdev)

Feature importance:

- t-statistic by the estimated weight scaled with its standard error (i.e., less certain about the correct value)

$$t_{w_i} = \frac{w_i}{SE(w_i)}$$

LOGISTIC REGRESSION

$$P(y=1) = \frac{1}{1+\exp(-(w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p))}$$

- Probabilistic classification model
- Typically, we set the threshold to 0.5 to predict
  - Class 1 if $P(y=1) > 0.5$
  - Class 0 if $P(y=1) \leq 0.5$

INTERPRETATION OF LOGISTIC REGRESSION

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p$$

- weights relate to log odds-ratio
- Again linear in log odds-ratio

$\leadsto$ change by one unit changes the odds ratio by a factor of $\exp(w_i)$.

- Interpretation for different feature types is the same as for linear regression

GLM AND INTERACTIONS

- Linear models are often too restrictive for many applications

Non-Gaussian outputs via Generalized Linear Models (GLMs):

$$g(\mathrm{E}_Y(y \mid x)) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p$$

- link function g – can be freely chosen
- exponential family defining E_Y – can be freely chosen
- weighted sum $X^\top W$

Interaction effects via feature engineering:

- E.g., feature expansion: $w_{x_i,x_j} x_i \cdot x_j$

GENERALIZED ADDITIVE MODELS (GAMS)

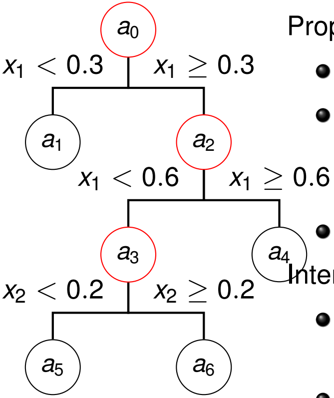
Non-Linear relations can be addressed by:

- feature transformations (e.g., exp or log)
- Categorization of features (i.e., intervals / buckets of feature values)
- GAMs:

$$g(\mathrm{E}_Y(y \mid x)) = w_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

- instead of $w_i x_i$ use flexible functions $f_i(x_i) \rightsquigarrow$ splines

DECISION TREES



Properties:

- able to model non-linear effects
- terminal nodes (aka leaf nodes) can have several observations and predicts the mean outcome over these
- Applicable to regression and classification

Interpretation:

- directly by following the tree (i.e., sequence of rules)
- Feature importance by (scaled) score of much the splitting criterion was reduced compared to the parent

DECISION RULES

`IF` $\texttt{COND}_1$ `AND` $\texttt{COND}_2$ `AND ... THEN value`

- $\texttt{COND}_i$ can be of the form `feature <op> value` where `<op>` can be for example $\{=, <, >\}$

Properties:

Support Fraction of observations to support appliance of rule

Accuracy for predicting the correct class under the condition(s)

$\rightsquigarrow$ often trade-off between these two

$\rightsquigarrow$ many different ways to learn a set of rules (incl. a default rule if none of the rules are met)

OTHER INTERPRETABLE MODELS

RuleFit https://arxiv.org/abs/0811.1679

- Combination of linear models and decision trees
- Allows for feature interactions and non-linearities

NaiveBayes

$$P(C_k \mid x) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

- product of probabilities for a class on the value of each feature
- strong independence assumption

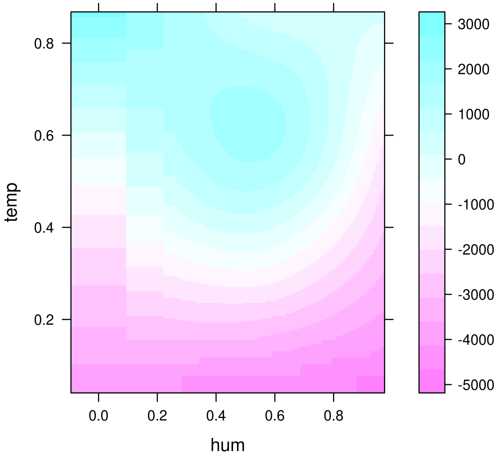
k-Nearest Neighbor

- (closely related to case-based reasoning)
- Average of of the outcome of neighbors – local explanation

MODEL-BASED BOOSTING

```
 Model-based Boosting

Call:
mboost(formula = cnt ~ bols(hum) + bols(temp) + bspatial(hum,     temp), data = data_bike)


 Squared Error (Regression)

Loss function: (y - f)^2


Number of boosting iterations: mstop = 100
Step size:  0.1
Offset:  4504.349
Number of baselearners:  3

Selection frequencies:
bspatial(hum, temp)
                  1
```

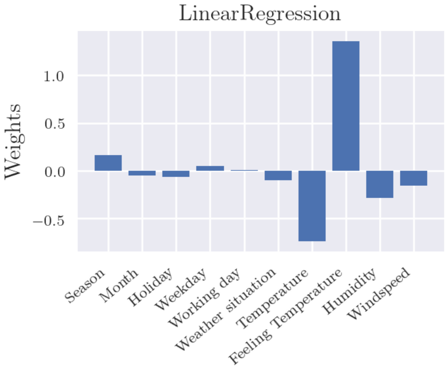
MODEL-BASED BOOSTING



BIKE RENTALS (REGRESSION)

- Target: number of rented bikes
- Source: bicycle rental company Capital-Bikeshare in Washington D.C.,
- Reference: https://link.springer.com/article/10.1007/s13748-013-0040-3
- Exemplary features:
  - season: spring, summer, fall or winter.
  - holiday or not.
  - working day or weekend
  - weather situation on that day
  - temperature

REGRESSION ON BIKE RENTALS

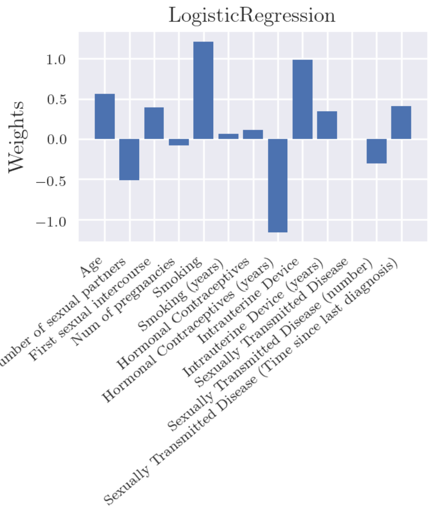


DECISION TREE ON BIKE RENTALS

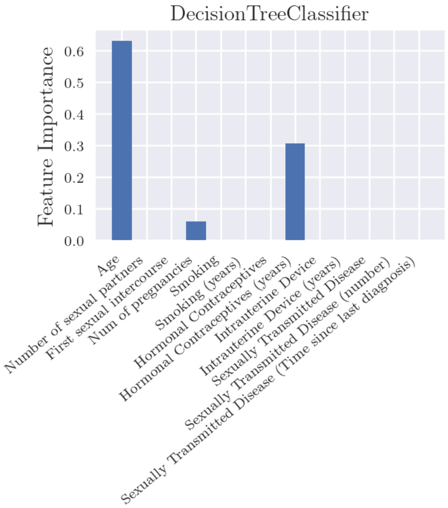
RISK FACTORS FOR CERVICAL CANCER (CLASSIFICATION)

- Target: patient will get cancer or not?
- Reference:
- Exemplary features:
  - Age in years
  - Number of sexual partners
  - First sexual intercourse (age in years)
  - Number of pregnancies
  - Smoking yes or no
  - Smoking (in years)
  - Hormonal contraceptives yes or no
  - Hormonal contraceptives (in years)
  - Intrauterine device yes or no (IUD)

LOGISTIC REGRESSION ON CANCER (CLASSIFICATION)



DECISION TREE ON CANCER (CLASSIFICATION)



PREDICTIVE PERFORMANCE

- Bike Rental (normalized MSE):
  - Linear Regression: 0.0276
  - Decision Tree: 0.0328
  - GLM: 0.0244

- Cancer (Accuracy)
  - Logistic Regression: 0.58
  - Decision Tree: 0.62

$\rightsquigarrow$ although easy to interpret, not really well-performing

  - Boosting: 0.79