Living Outside of “Black Box”

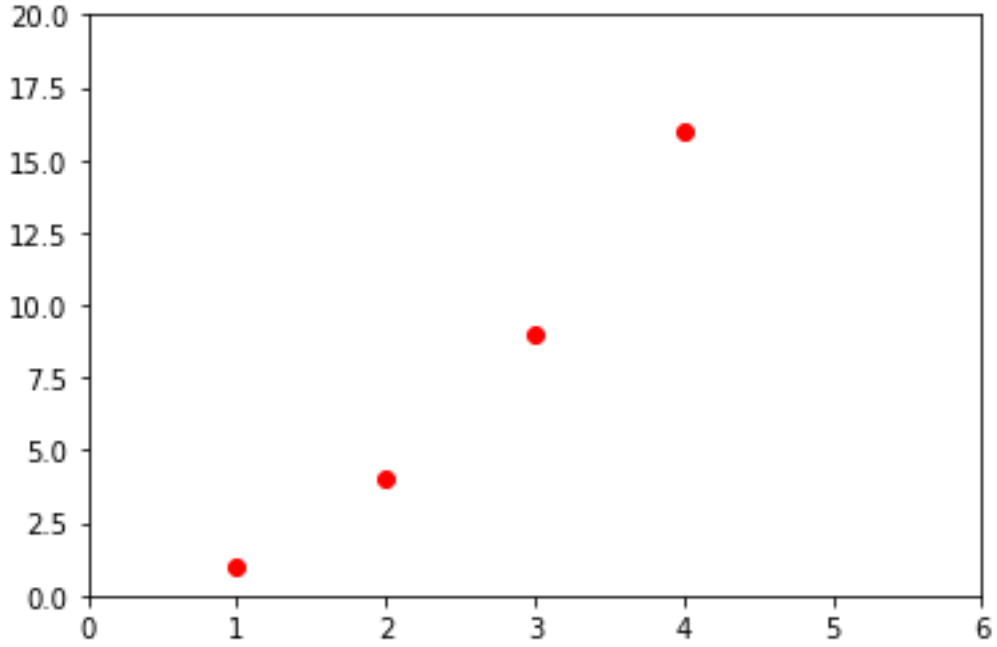
Survival Guide to Interpretable Models

Sergei Filatyev

Hasn’t deep learning solved all our problems?

Hasn’t deep learning solved all our problems?

- Small data



- Rules and regulations

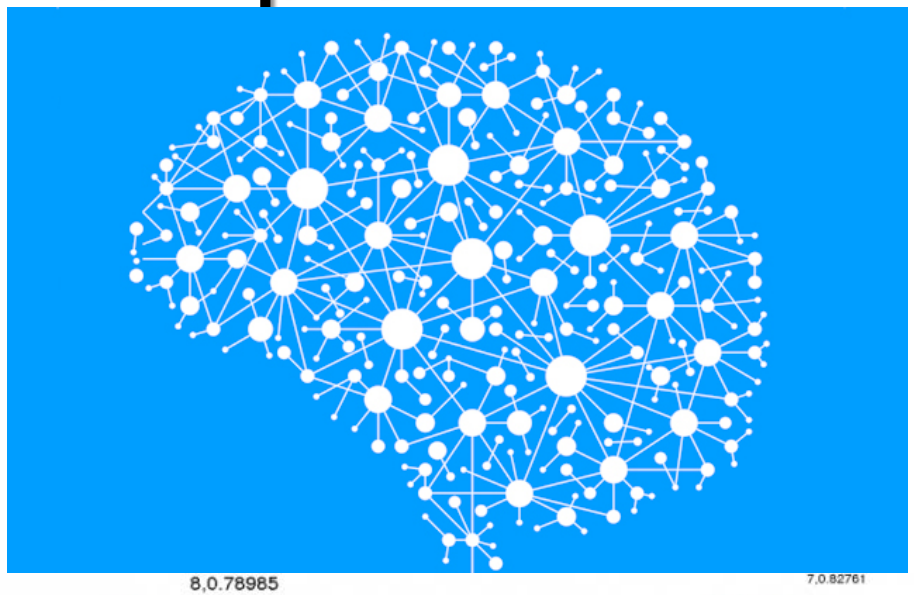
Hasn’t deep learning solved all our problems?

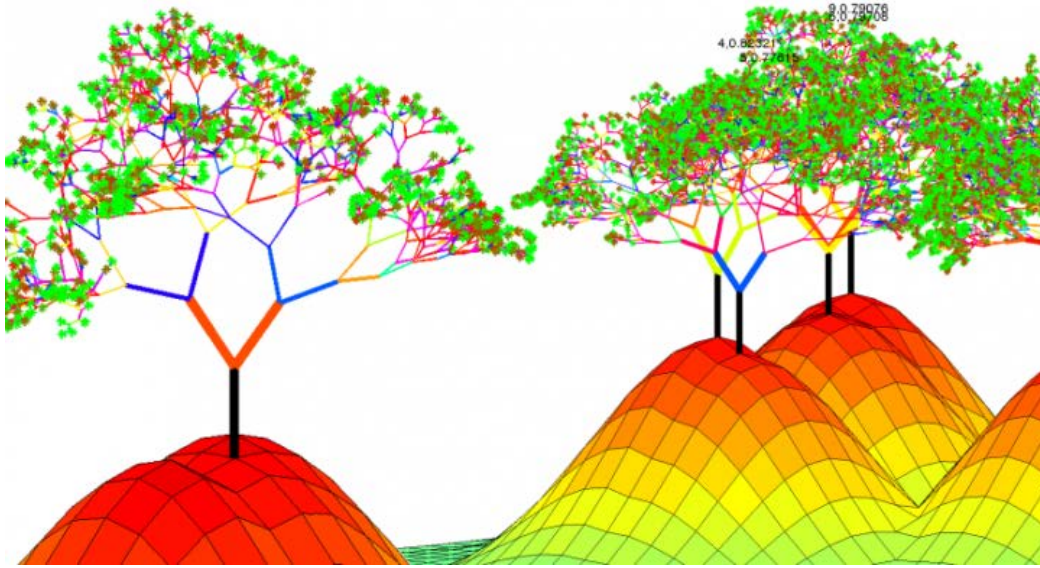
- Small data

- Rules and regulations

Hasn’t deep learning solved all our problems?

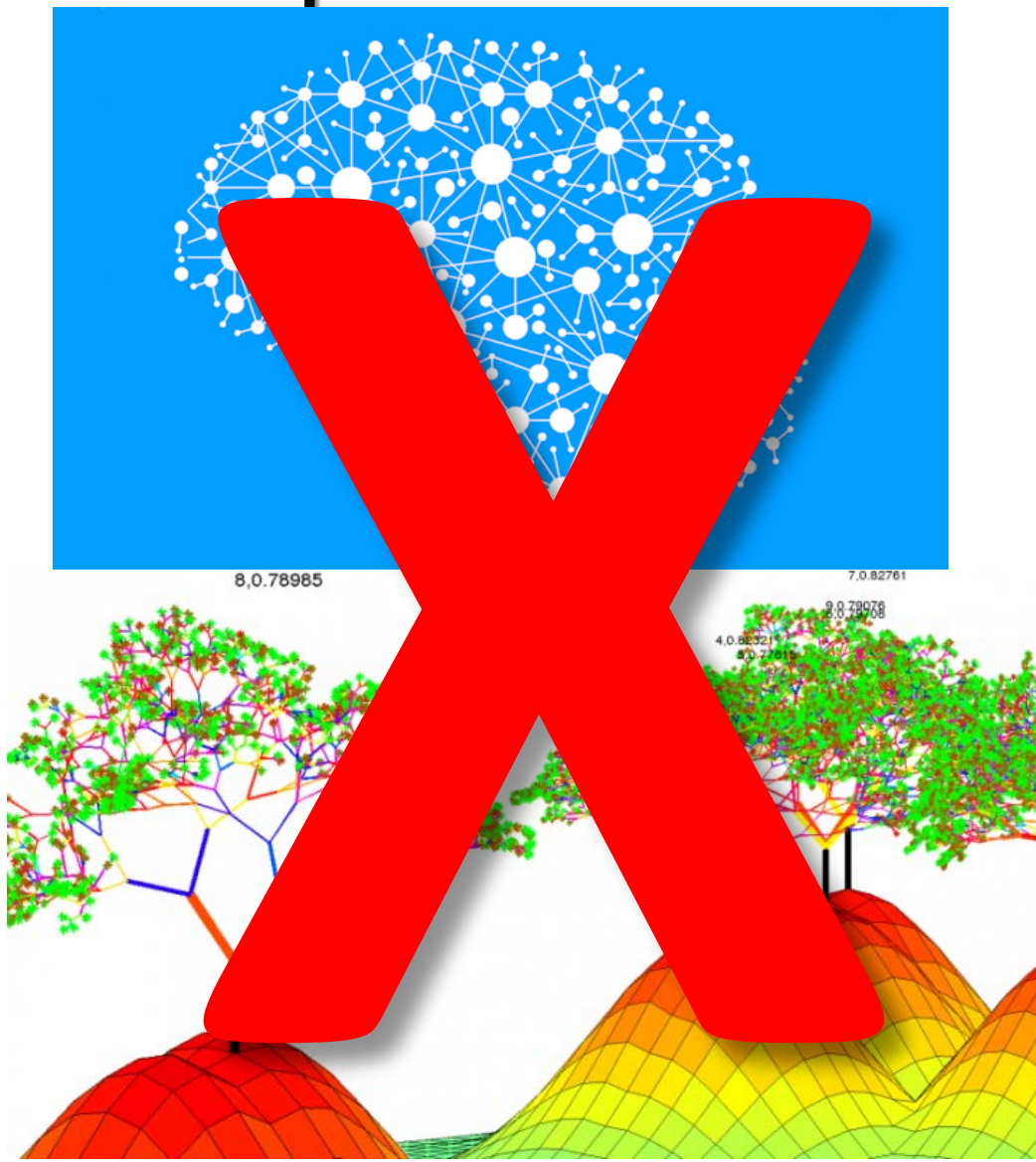
- Small data
- Rules and regulations





Hasn’t deep learning solved all our problems?

- Small data
- Rules and regulations

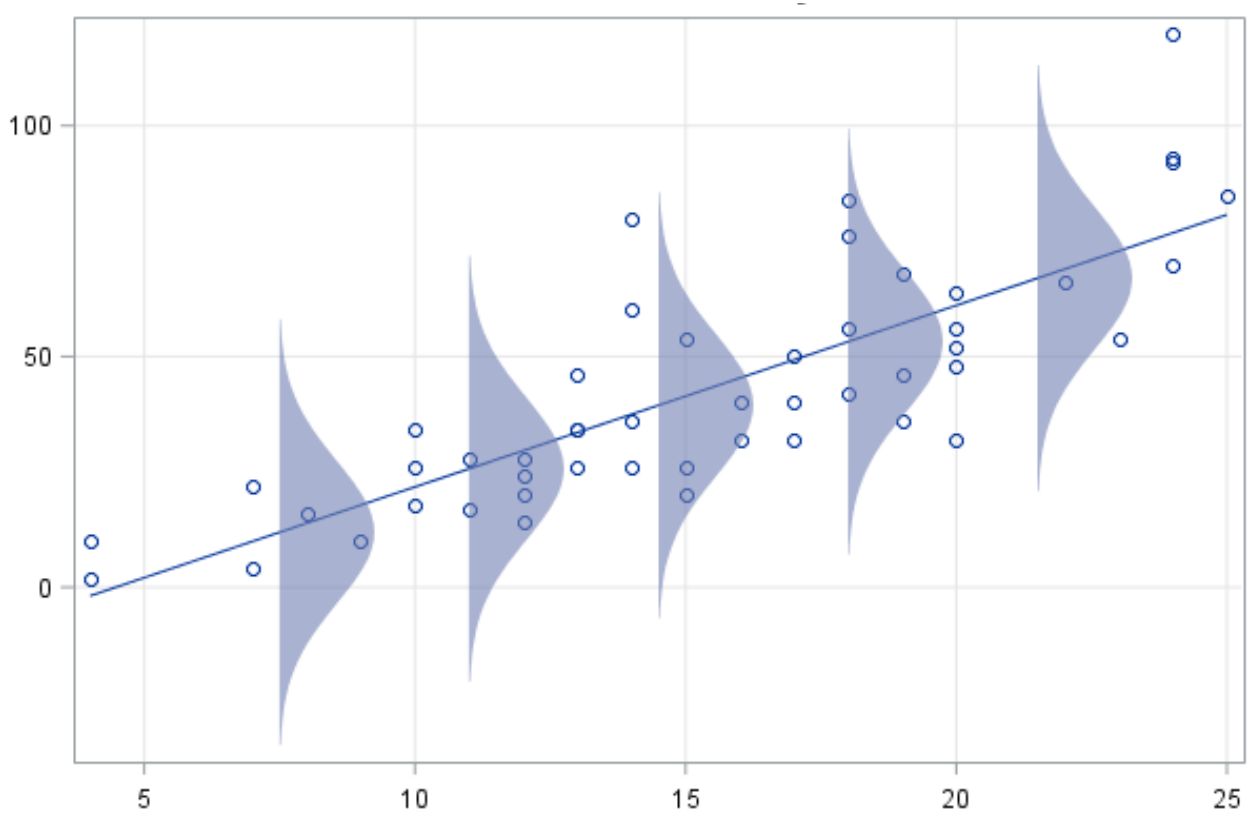


Linear Models

$$y_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \cdots + a_K X_{Ki} + \varepsilon_i$$

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

$$\mathbf{E} y_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \cdots + a_K X_{Ki}$$



lm

sklearn.linear_model.LinearRegression

PROC REG

Is it too restrictive?

Is it too restrictive?

Polynomial Regression

$X_1 = t, \qquad X_2 = t^2, \qquad X_3 = t^3, \ldots$

Is it too restrictive?

Polynomial Regression

$$X_1 = t, \qquad X_2 = t^2, \qquad X_3 = t^3, \ldots$$

$$X_1 = \sqrt{t}, \qquad X_2 = t^{2/5} \log t\,, \qquad X_3 = \int_0^t e^{-\tau^2} d\tau\,, \ldots$$

Is it too restrictive?

Polynomial Regression

$$X_1 = t, \qquad X_2 = t^2, \qquad X_3 = t^3, \ldots$$

$$X_1 = \sqrt{t}, \qquad X_2 = t^{2/5} \log t\,, \qquad X_3 = \int_0^t e^{-\tau^2} d\tau\,, \ldots$$

Interaction Terms

$$\mathrm{y} = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_1 X_2 + a_5 X_2 X_3 + a_6 X_1 X_3 + a_7 X_1 X_2 X_3 + \cdots$$

Generalized Linear Models: Binary Target

$$\mathbf{E}y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

Generalized Linear Models: Binary Target

$$\varphi(\mathbf{E}y) = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

Generalized Linear Models: Binary Target

$$\varphi(\mathbf{E}y) = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

Logit

$$\varphi(\mathbf{E}y) = \log \frac{\mathbf{E}y}{1 - \mathbf{E}y} = \log \frac{P(y = 1)}{P(y = 0)}$$

Generalized Linear Models: Binary Target

$$\varphi(\mathbf{E}y) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_KX_K$$

Logit

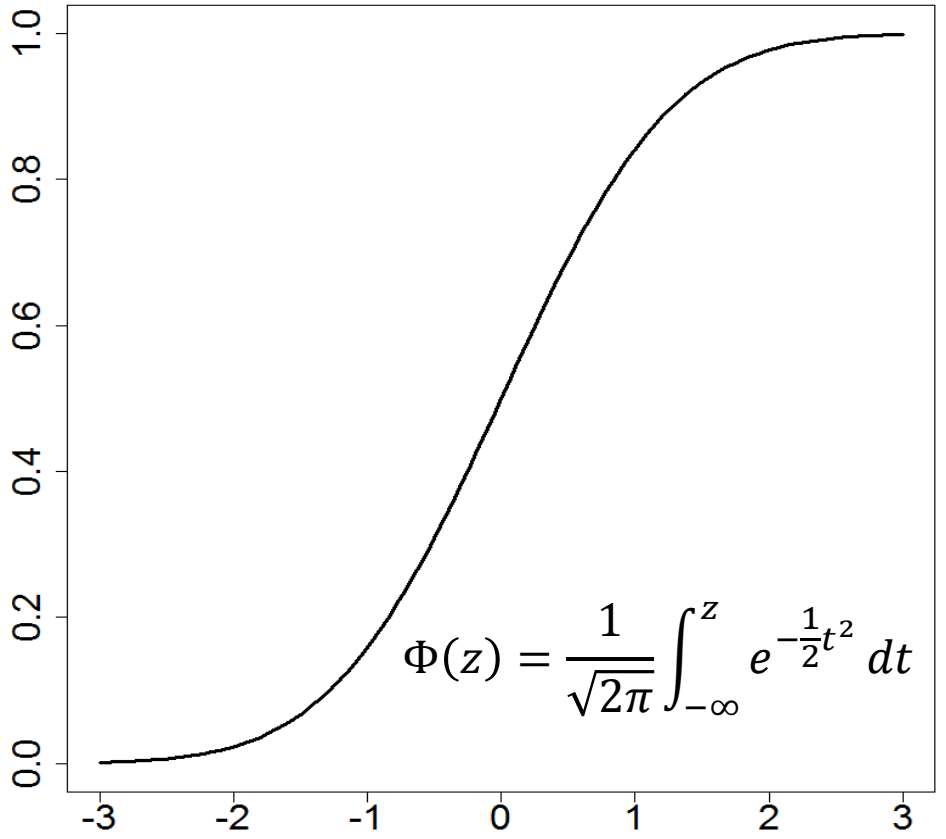
$$\varphi(\mathbf{E}y) = \log\frac{\mathbf{E}y}{1-\mathbf{E}y} = \log\frac{P(y=1)}{P(y=0)}$$

Probit

$$\mathbf{E}y = P(y=1) \;=\; \Phi(a_0 + a_1X_1 + a_2X_2 + \cdots + a_KX_K)$$

Generalized Linear Models: Binary Target

$$\varphi(\mathbf{E}y) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_KX_K$$



Logit

$$\varphi(\mathbf{E}y) = \log \frac{\mathbf{E}y}{1 - \mathbf{E}y} = \log \frac{P(y=1)}{P(y=0)}$$

Probit

$$\mathbf{E}y = P(y=1) \;=\; \Phi(a_0 + a_1X_1 + a_2X_2 + \cdots + a_KX_K)$$

Generalized Linear Models: Binary Target

glm
glmnet

sklearn.linear_model.LogisticRegression
statsmodels.discrete.discrete_model.Probit

PROC LOGISTIC
PROC PROBIT

Generalized Linear Models: Count $y = 0, 1, 2, 3, \ldots$

Generalized Linear Models: Count $y = 0, 1, 2, 3, \ldots$

Poisson

$$P(y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

$$\log \mu = \log \mathbf{E}y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

Generalized Linear Models: Count $y = 0, 1, 2, 3, \ldots$

Poisson

$$P(y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

$$\log \mu = \log \mathbf{E}y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

$$\mu = \mathbf{E}y = \mathbf{Var}\ \mathrm{y}$$

Generalized Linear Models: Count $y = 0, 1, 2, 3, \ldots$

Poisson $$P(y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

$$\log \mu = \log \mathbf{E}y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

$$\mu = \mathbf{E}y = \mathbf{Var}\ \mathrm{y}$$

Negative Binomial $\mathbf{E}y < \mathbf{Var}\ \mathrm{y}$

Generalized Poisson $\mathbf{E}y \neq \mathbf{Var}\ \mathrm{y}$

Generalized Linear Models: Count $y = 0, 1, 2, 3, \ldots$

glm
MASS::glm.nb
VGAM::vglm

statsmodels.discrete.discrete_model.Poisson
statsmodels.discrete.discrete_model.NegativeBinomial
PyMC3 package

PROC GENMOD
PROC NLMIXED
PROC GLIMMIX

Generalized Additive Models

$$\mathbf{E}y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

Generalized Additive Models

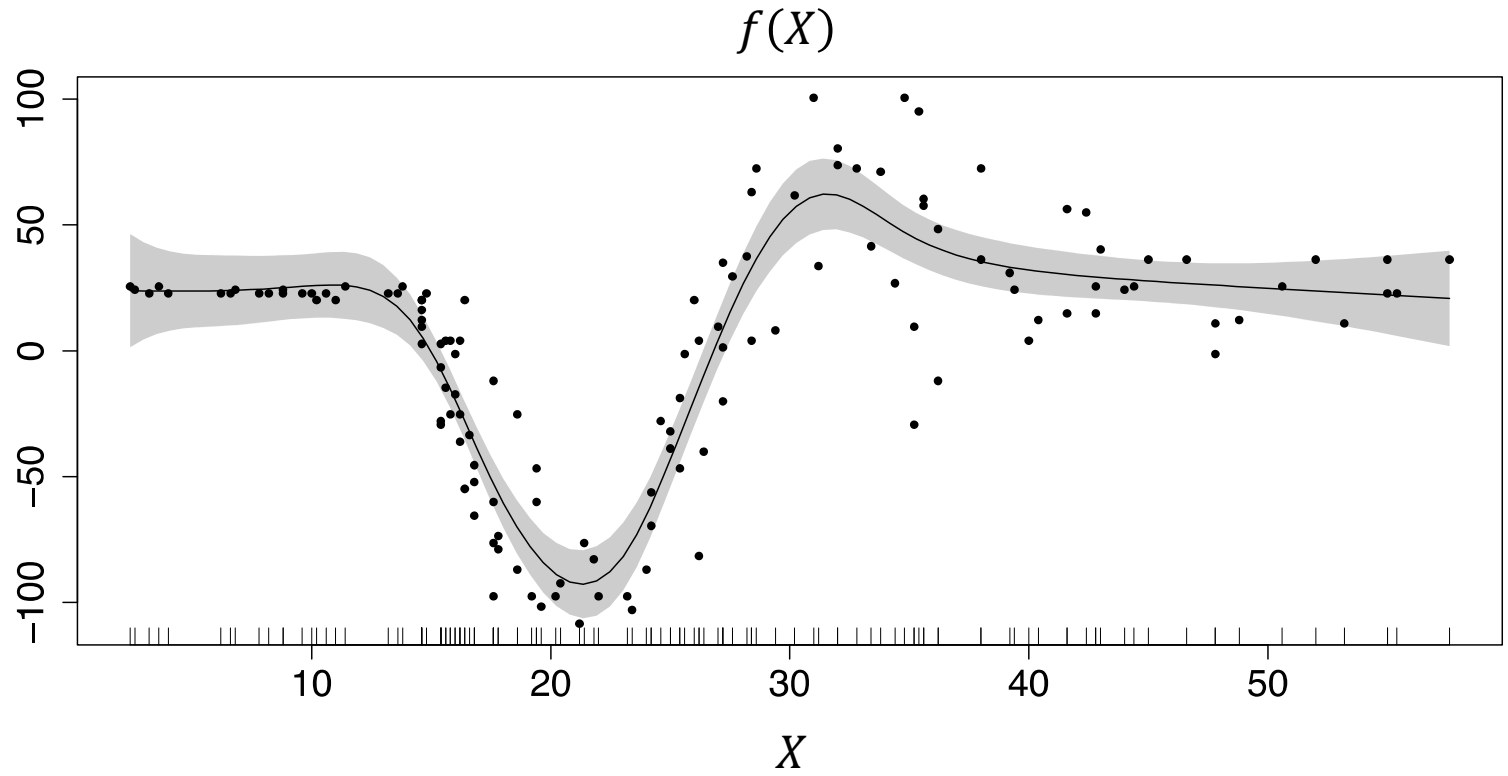
$$\varphi(\mathbf{E}y) = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_K X_K$$

Generalized Additive Models

$$\varphi(\mathbf{E}y) = f_1(X_1) + f_2(X_2) + \cdots + f_K(X_K)$$

Generalized Additive Models

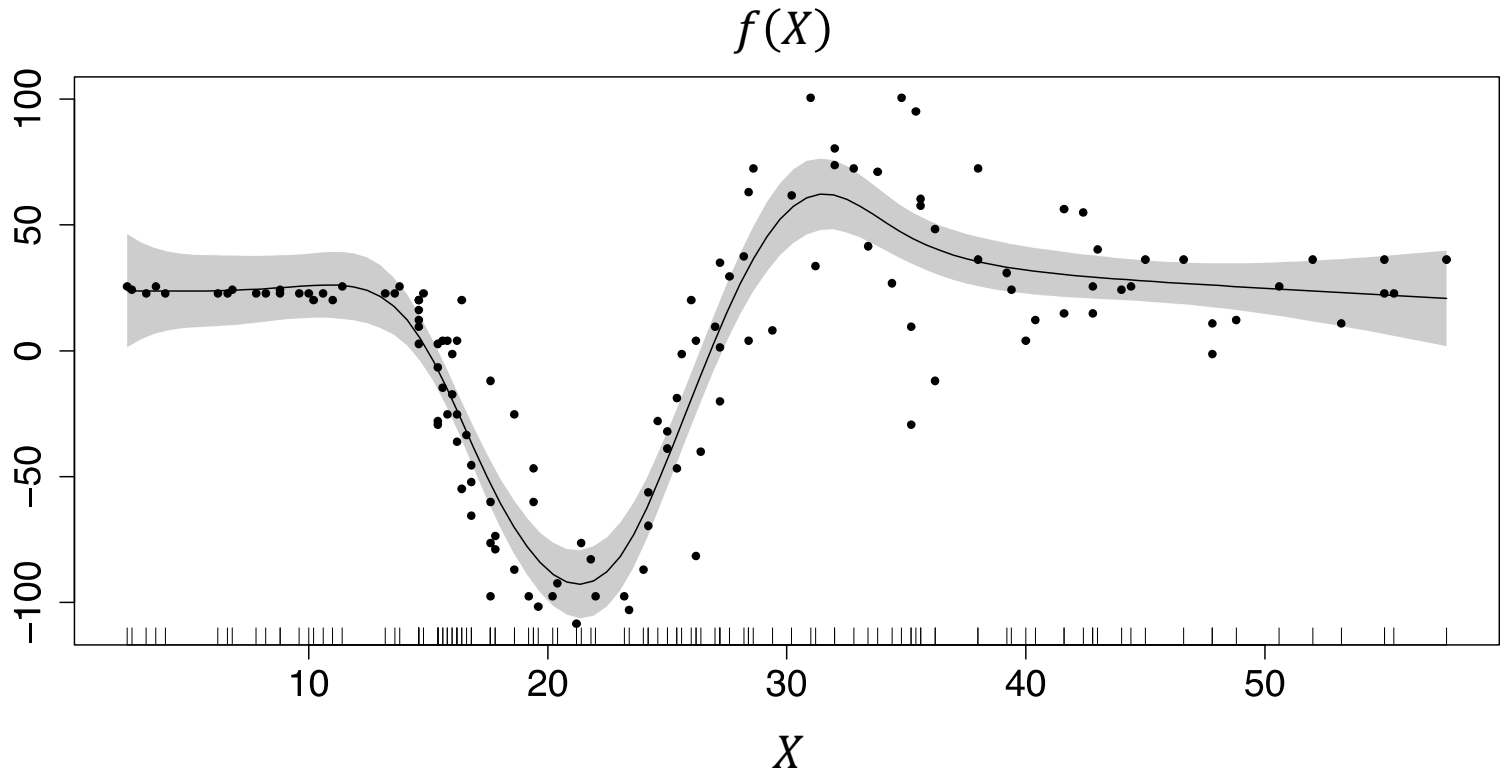
$$\varphi(\mathbf{E}y) = f_1(X_1) + f_2(X_2) + \cdots + f_K(X_K) \qquad f_i(X_i) = \text{spline}$$



Generalized Additive Models

$$\varphi(\mathbf{E}y) = f_1(X_1) + f_2(X_2) + \cdots + f_K(X_K) \qquad f_i(X_i) = \text{spline}$$

Extensions

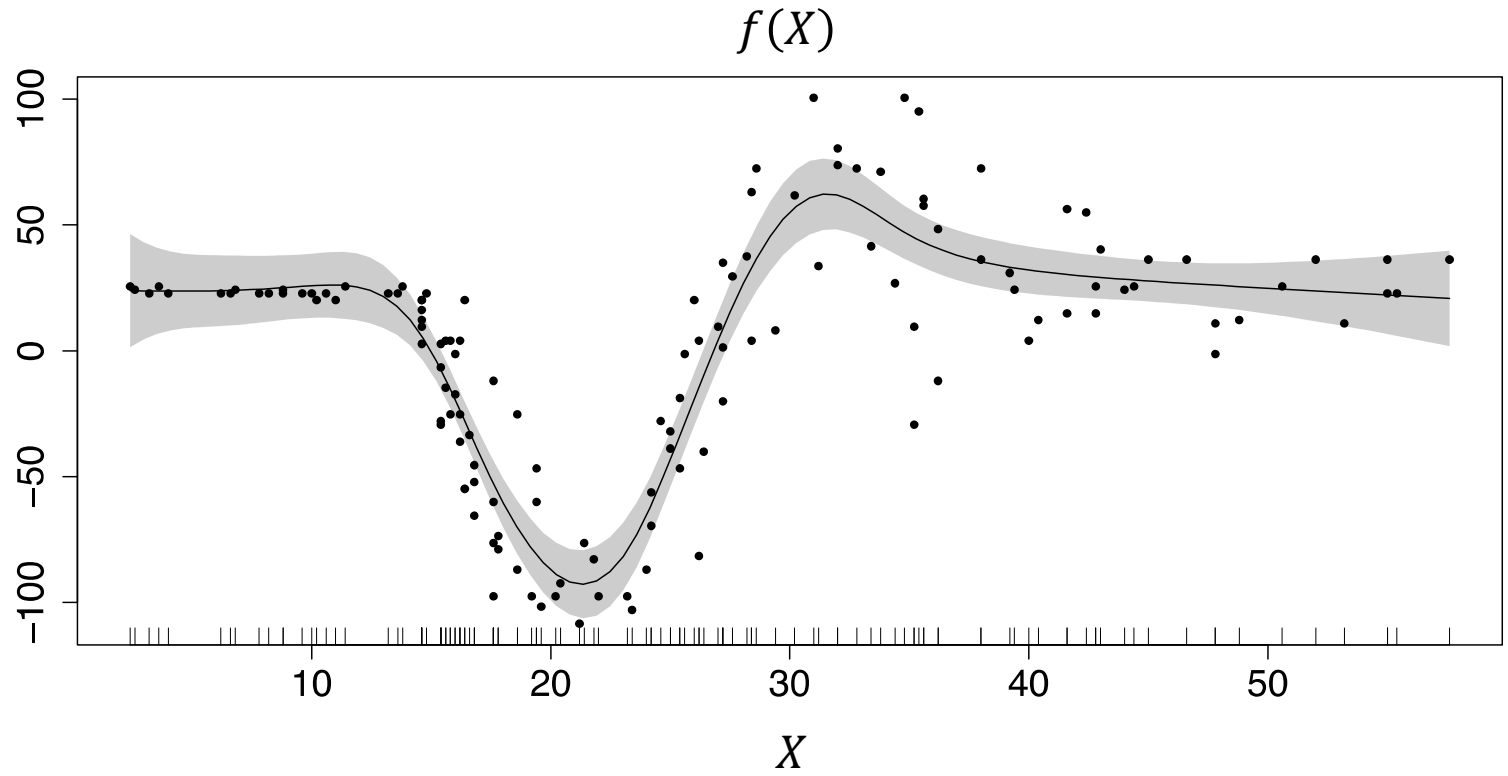


Generalized Additive Models

$$\varphi(\mathbf{E}y) = f_1(X_1) + f_2(X_2) + \cdots + f_K(X_K) \qquad f_i(X_i) = \text{spline}$$

Extensions

GAM + GLM



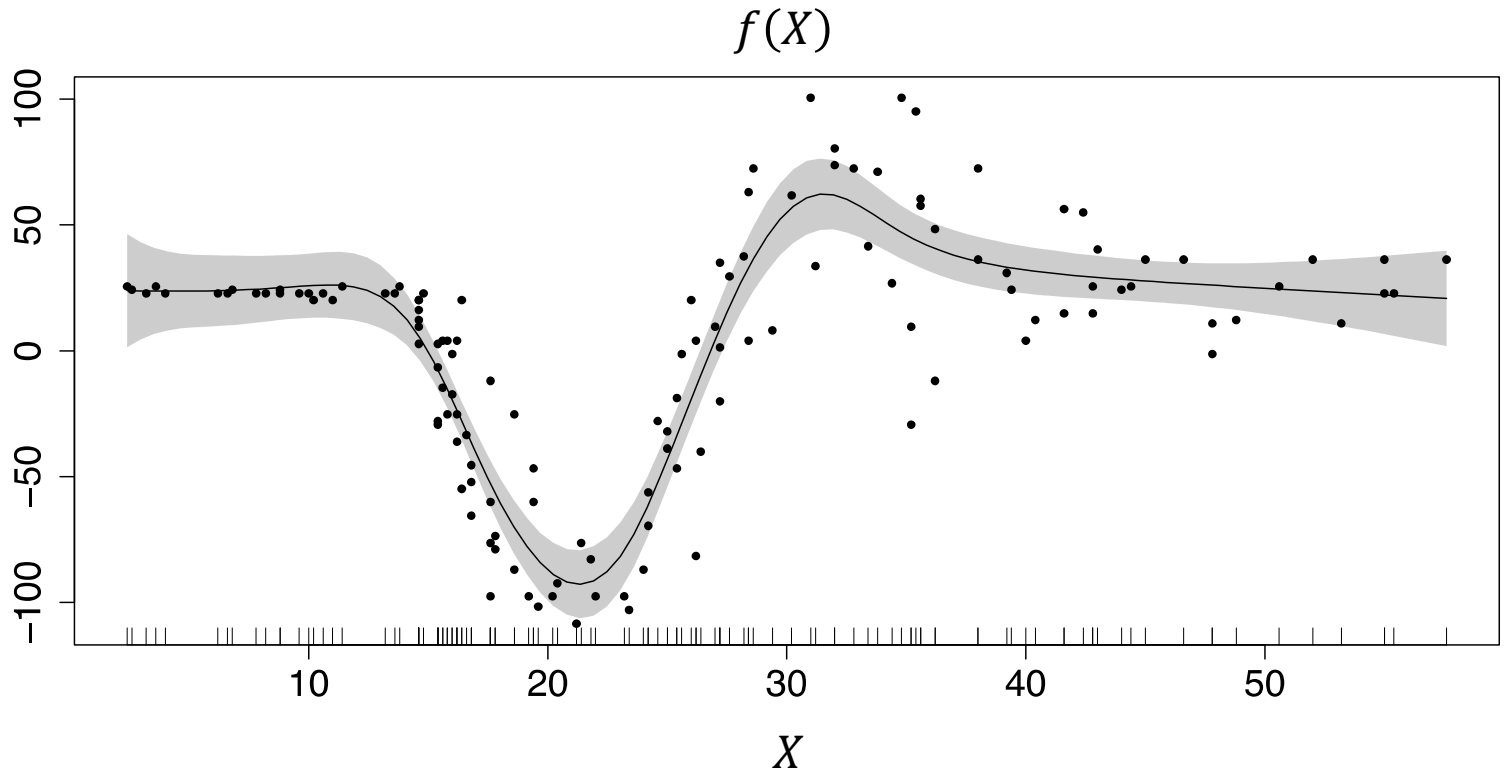
Generalized Additive Models

$$\varphi(\mathbf{E}y) = f_1(X_1) + f_2(X_2) + \cdots + f_K(X_K) \qquad f_i(X_i) = \text{spline}$$

Extensions

GAM + GLM

$f_1(X_1, X_2, \dots)$



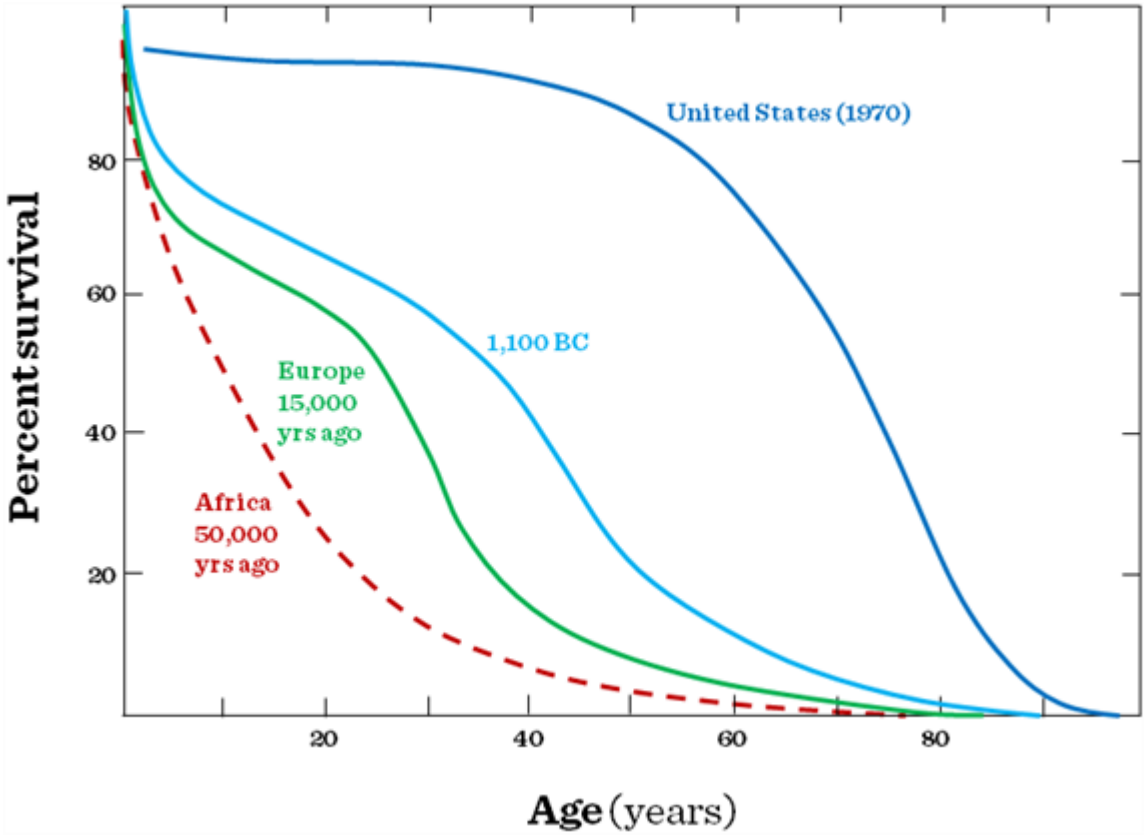
Generalized Additive Models

gam::gam
mgcv::gam

pyGAM package

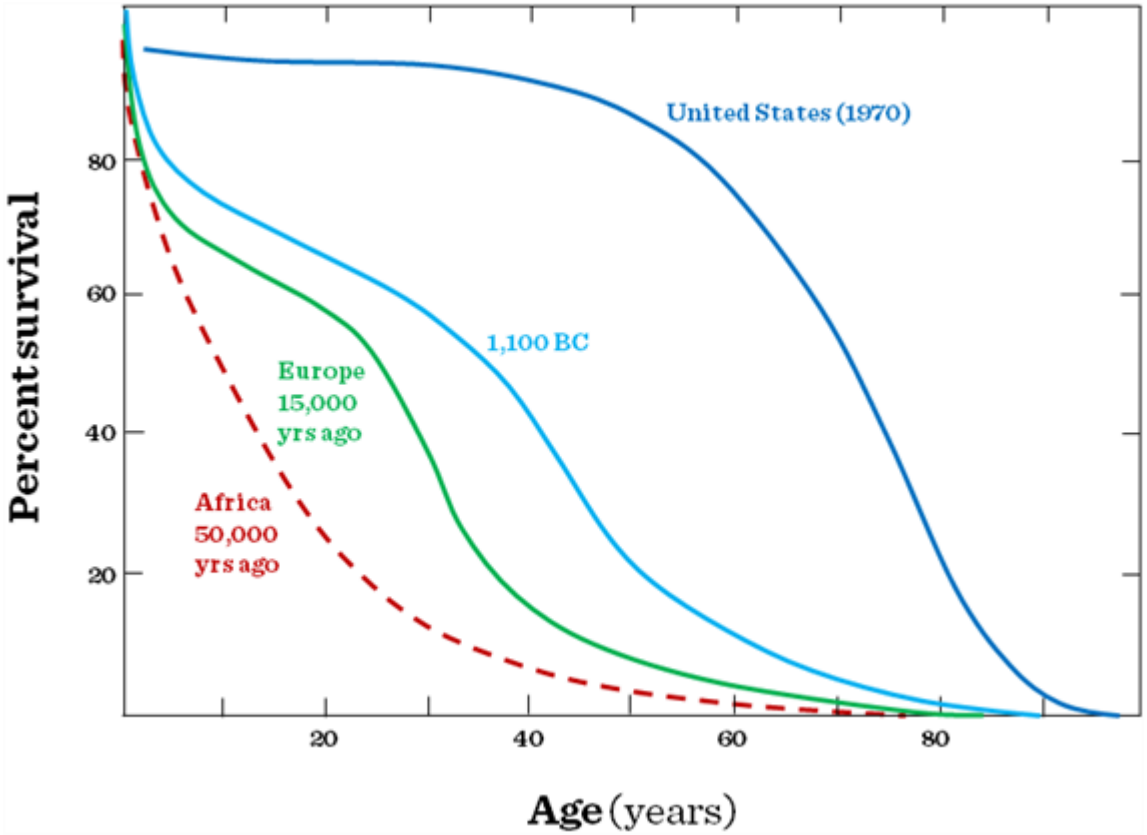
PROC GAM

Survival Analysis



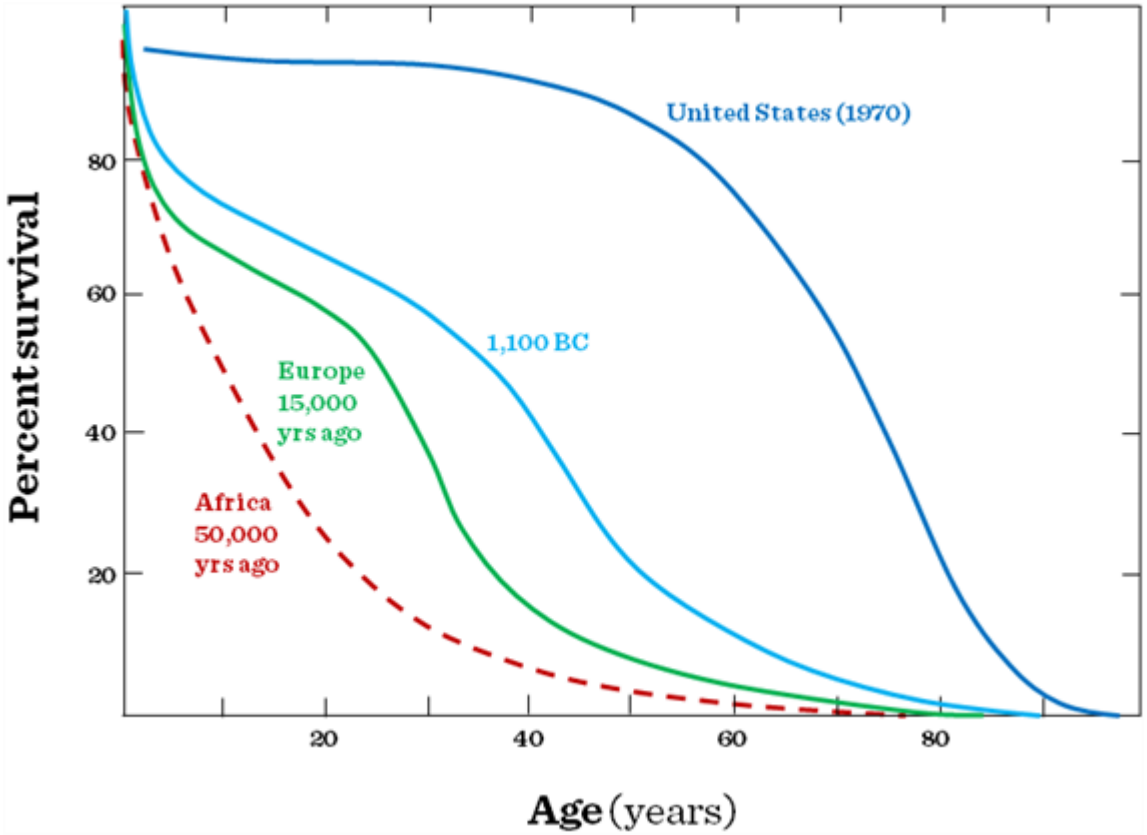
Survival Analysis

DURATION



Survival Analysis

DURATION



Time to event T

Survival function $S(t) = \Pr(T > t)$

Hazard function $\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = -\frac{d}{dt} \log S(t)$

Survival Analysis: Cox Proportional Hazard Model

$$\lambda_i(t) = \lambda_0(t)\exp(a_1 X_{1i} + a_2 X_{2i} + \cdots + a_K X_{Ki})$$

$$S_i(t) = S_0(t)^{\exp(a_1 X_{1i} + a_2 X_{2i} + \cdots + a_K X_{Ki})}$$

$$S_0(t) = \exp\{-at^p\}$$ Weibull

Survival Analysis: Cox Proportional Hazard Model

$$\lambda_i(t) = \lambda_0(t)\exp(a_1 X_{1i} + a_2 X_{2i} + \cdots + a_K X_{Ki})$$

$$S_i(t) = S_0(t)^{\exp(a_1 X_{1i} + a_2 X_{2i} + \cdots + a_K X_{Ki})}$$

$$S_0(t) = \exp\{-at^p\}$$ Weibull

Extensions: $X(t), a(t)$

Survival Analysis

survival package

lifelines package
scikit-survival package

PROC LIFETEST
PROC PHREG
PROC LIFEREG

Naïve Bayes

Naïve Bayes

$$P(y|X_1, X_2, \dots, X_K) = \frac{P(X_1, X_2, \dots, X_K|y)P(y)}{P(X_1, X_2, \dots, X_K)}$$

Naïve Bayes

$$P(y|X_1, X_2, \dots, X_K) = \frac{P(X_1, X_2, \dots, X_K|y)P(y)}{P(X_1, X_2, \dots, X_K)}$$

$$P(X_1, X_2, \dots, X_K|y) = P(X_1|y)P(X_2|y) \dots P(X_K|y)$$

Naïve Bayes

$$P(y|X_1, X_2, \dots, X_K) = \frac{P(X_1, X_2, \dots, X_K|y)P(y)}{P(X_1, X_2, \dots, X_K)}$$

$$P(X_1, X_2, \dots, X_K|y) = P(X_1|y)P(X_2|y) \dots P(X_K|y)$$

$$P(y|X_1, X_2, \dots, X_K) = \frac{P(X_1|y)P(X_2|y) \dots P(X_K|y)P(y)}{P(X_1, X_2, \dots, X_K)}$$

Naïve Bayes

$$P(y|X_1, X_2, \dots, X_K) = \frac{P(X_1, X_2, \dots, X_K|y)P(y)}{P(X_1, X_2, \dots, X_K)}$$

$$P(X_1, X_2, \dots, X_K|y) = P(X_1|y)P(X_2|y) \dots P(X_K|y)$$

$$P(y|X_1, X_2, \dots, X_K) = \frac{P(X_1|y)P(X_2|y) \dots P(X_K|y)P(y)}{P(X_1, X_2, \dots, X_K)}$$

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(X_1|y)P(X_2|y) \dots P(X_K|y)P(y)$$

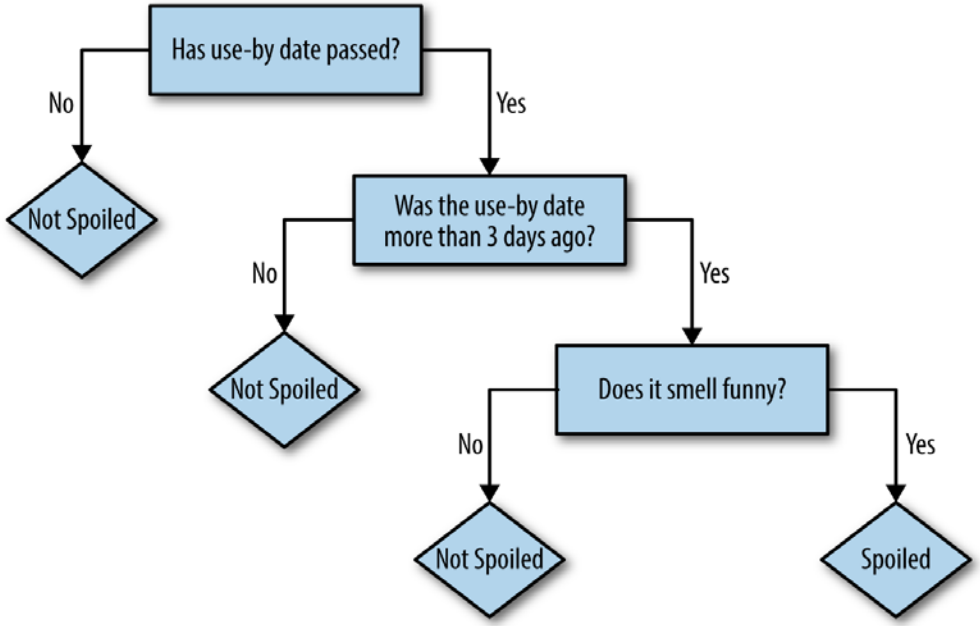
Naïve Bayes

e1071::naiveBayes
naivebayes package

sklearn.naive_bayes

PROC HPBNET
Enterprise Miner

Decision Trees



Problems: overfit, sensitive to data
Control: pruning, depth, number of samples on leaf

Selection Bias: variable with many splits, missing values
Control: Conditional Inference Trees

Decision Trees

rpart::rpart
party::ctree

sklearn.tree.DecisionTreeClassifier
sklearn.tree.DecisionTreeRegressor

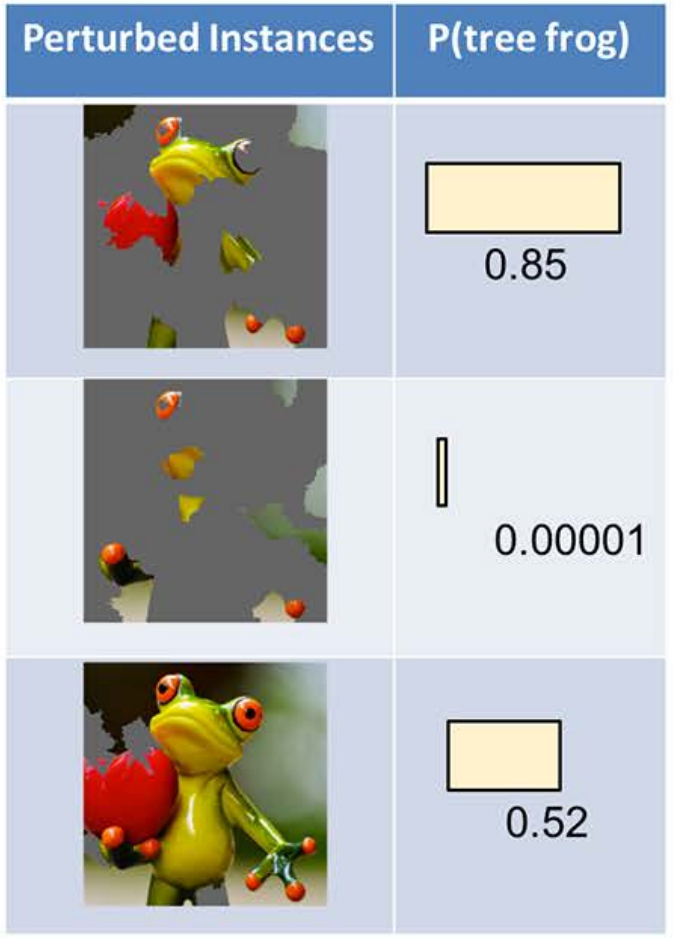
PROC HPSPLIT
PROC HPFOREST

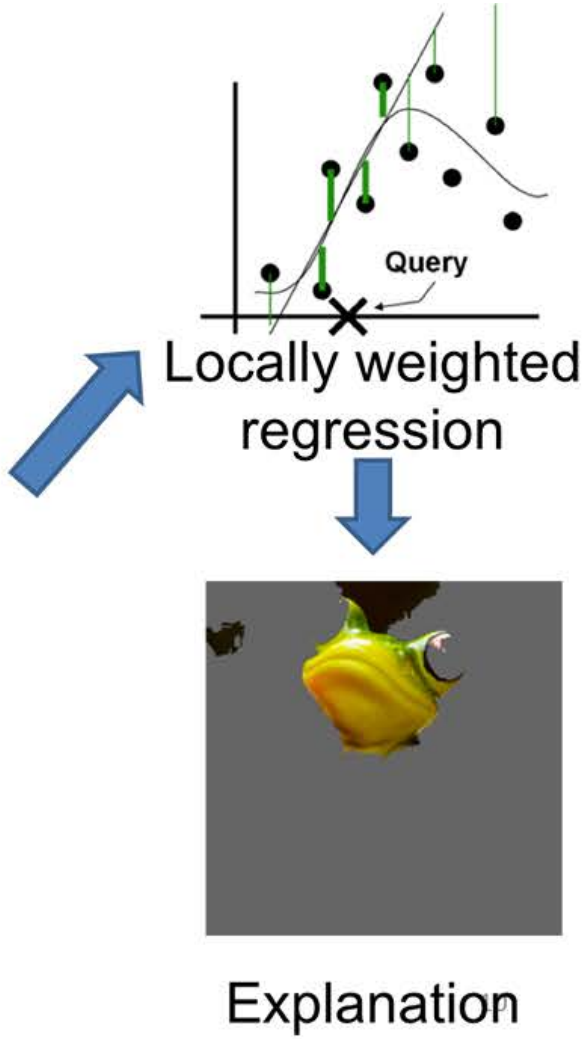
LIME: Locally Interpretable Model-Agnostic Explanations

Original Image

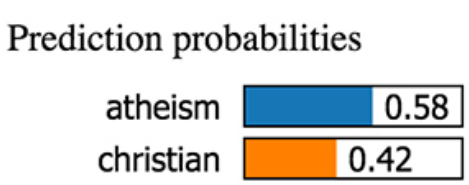
P(tree frog) = 0.54

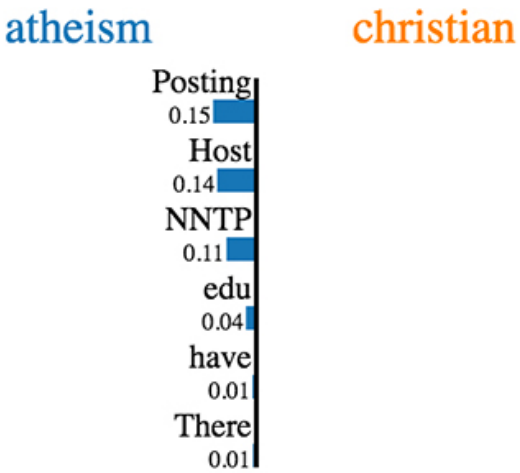
Interpretable Components





LIME: Locally Interpretable Model-Agnostic Explanations





Text with highlighted words

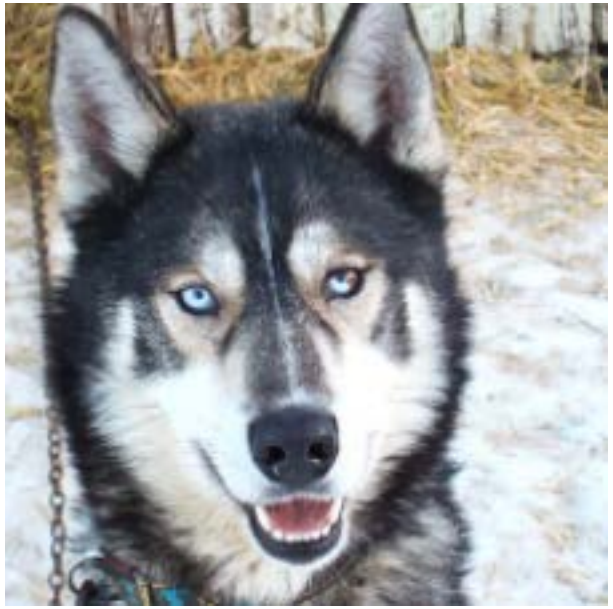
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

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.

Random Forest – 92.4% accuracy
Posting – 21.6% cases (2 cases – Christianity)

LIME: Locally Interpretable Model-Agnostic Explanations



Husky classified as wolf

Explanation

LIME: Locally Interpretable Model-Agnostic Explanations

lime package

lime package

Driverless AI

https://lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model.html#lime-local-interpretable-model-agnostic-explanations