STATISTICAL LEARNING SYSTEMS

LECTURE 3: LOCAL AND NONPARAMETRIC REGRESSION REVISITED

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Parametric nonlinear regression

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $f(\cdot)$ is a known function and (ε_i) are i.i.d. random variables. The $\boldsymbol{\beta}$ are the sole unknown parameters of the model. Usually, ML or Nonlinear LS estimator is considered:

$$\hat{\boldsymbol{\beta}}^{NLS} = \operatorname{argmin}_{\boldsymbol{\beta}} \{ \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2 \}.$$

No explicit solution is usually known, hence iterative Newton-Raphson procedure is usually used to find a stationary point of a criterion above.

But how to determine $f(\cdot)$?

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Low dimensional problems and local smoothing - regression and smoothing splines

Until further notice, we stick to one-dimensional x.

Let $[t_k, t_{k+1}]$, $k = 1, \ldots, K$, be contiguous intervals whose sum is a range of x over which unknown $f(\cdot)$ is to be estimated. A polynomial spline of order q with knots t_k, $k = 1, \ldots, K$, is a function which is a polynomial over each interval $[t_k, t_{k+1}]$, and is $q - 1$ times continuously differentiable at each t_k, $k = 1, \ldots, K$.

We shall now briefly describe estimation of regression functions by regression splines. Most often, cubic splines are used in this context ($q = 3$).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Regression and smoothing splines

It is natural for a regression spline to write as a linear combination of the following basis functions:

$$1, \quad \{x^j\}_{j=1}^{q-1}, \quad \{(x-t_k)_+^q\}_{k=1}^K,$$

where $\{t_k\}_{k=1}^K$ are the knots. Accordingly, a spline can be written as

$$\tilde{f}(x) = \sum_{i=1}^{q} \alpha_i x^{i-1} + \sum_{k=1}^{K} \beta_k (x-t_k)_+^q.$$

Given data, we choose K and t_k, $k = 1, \ldots, K$, by crossvalidation, where, for each fixed K and $\{t_k\}_{k=1}^K$, parameters α i β are found by LS, i.e., by minimizing

$$\sum_{i=1}^{n} (y_i - \tilde{f}(x_i))^2.$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Regression and smoothing splines

We try regression splines when K is found to be much smaller than sample size n. Most interestingly, given a sample $\{x_i, Y_i\}_1^n$, it can be shown that minimization of

$$\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2 + \lambda \int_R [\tilde{f}''(x)]^2 dx$$

in the class of all twice continuously differentiable $\tilde{f}(\cdot)$ has (for $\lambda > 0$) an explicit and unique solution which is a natural cubic spline with knots $\{x_i\}_1^n$; a natural cubic spline is a cubic spline with additional constraint that it is linear beyond the boundary knots.

Notice that for $\lambda = 0$ we get a function which interpolates the data, while for $\lambda = \infty$ we get ordinary LS (OLS) linear regression fit.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Regression and smoothing splines

A natural cubic spline with K knots is represented by K basis functions. Starting from a basis for cubic splines we arrive at the basis:

$$N_1(x) = 1, \quad N_2(x) = x, \quad , N_{k+2}(x) = d_k(x) - d_{K-1}(x),$$

where

$$d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k},$$

with ξ_k denoting the knots. Hence, a natural spline with n knots (at $\{x_i\}_1^n$) can be written as

$$f(x) = \sum_{j=1}^{n} N_j(x)\theta_j.$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Regression and smoothing splines

The criterion

$$\mathrm{RSS}(f,\lambda) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_R [f''(x)]^2 dx$$

thus reduces to

$$\mathrm{RSS}(\theta,\lambda) = (\mathbf{y} - \mathbf{N}\theta)'(\mathbf{y} - \mathbf{N}\theta) + \lambda\theta'\mathbf{\Omega}_n\theta,$$

where $\{\mathbf{N}\}_{ij} = \{N_j(x_i)\}$ and $\{\mathbf{\Omega}_n\}_{jk} = \int N_j''(t)N_k''(t)dt$. The solution is easily seen to be

$$\hat{\theta} = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{\Omega}_n)^{-1}\mathbf{N}'\mathbf{y},$$

a generalized ridge regression. The fitted smoothing spline is given by

$$\hat{f}(x) = \sum_{j=1}^{n} N_j(x)\hat{\theta}_j.$$

Notice that the vector of fitted values, $\hat{\mathbf{f}} = [\hat{f}(x_1), \hat{f}(x_i 2), \ldots, \hat{f}(x_n)]'$, at the training predictors,

$$\hat{\mathbf{f}} = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\boldsymbol{\Omega}_n)^{-1}\mathbf{N}'\mathbf{y} = \mathbf{S}_\lambda \mathbf{y},$$

is linear in $\mathbf{y}$ and the smoother matrix $\mathbf{S}_\lambda$ depends only on the x_i and λ, and not on $\mathbf{y}$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Low dimensional and local smoothing - regression and smoothing splines

At least in principle, estimating regression functions by smoothing splines (as well as regression splines) can easily be extended to problems with multiple predictors. The criterion to be minimized assumes the form:

$$\hat{f}(\mathbf{x}) = \underset{\tilde{f}(\cdot)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \tilde{f}(\mathbf{x}_i))^2 + \lambda R(\tilde{f}) \right\},$$

where

$$R(\tilde{f}) = \sum_{k=1}^{p} \sum_{l=1}^{p} \int \left(\frac{\partial^2 \tilde{f}}{\partial x^{(k)} \partial x^{(l)}} \right)^2 d\mathbf{x}.$$

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund

Low dimensional and local smoothing - local linear regression

Let us return to regression functions defined on R.

The idea behind the local linear smoother is to approximate $f(x)$ locally by a linear function $\beta_0(x) + \beta_1(x)x$:

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \text{argmin}_{\beta_0(x),\beta_1(x)} \sum_{i=1}^{n} K(\frac{x - x_i}{\lambda})(Y_i - \beta_0(x) - \beta_1(x)x_i)^2,$$

where $K(\cdot)$ is a suitably defined kernel function and λ is a smoothing parameter.

The estimate (at x) is then $\hat{f}_\lambda(x) = \hat{\beta}_0(x) + \beta_1(x)x$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Local linear regression

Let $b(x)' = [1, x]$, $\mathbf{B}$ be the $n \times 2$ regression matrix with ith row $b(x_i)'$, and $\mathbf{W}(x)$ the $n \times n$ diagonal matrix with ith diagonal element $K_\lambda((x - x_i)/\lambda)$. It is easy to see that the estimate for $f(x)$ is then

$$\hat{f}(x) = b(x)'(\mathbf{B}'\mathbf{W}(x)\mathbf{B})^{-1}\mathbf{B}'\mathbf{W}(x)\mathbf{y}.$$

Local polynomial smoothers can be defined analogously, although no formula like the one above holds. One example of such smoothers is the locally weighted polynomial smoother (LOESS) of Cleveland (or its later modifications).

There can be good reasons to use local quadratic fits in the interior of the domain of a regression function, and local linear (or cubic) fits at the domain boundaries.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Low dimensional and local smoothing - purely kernel smoothing

Running mean:

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^{n} y_i I\{|x_i - x| \leqslant \lambda\}}{\#N_\lambda(x)},$$

where $N_\lambda(x) = \{i : |x_i - x| \leqslant \lambda\}$.

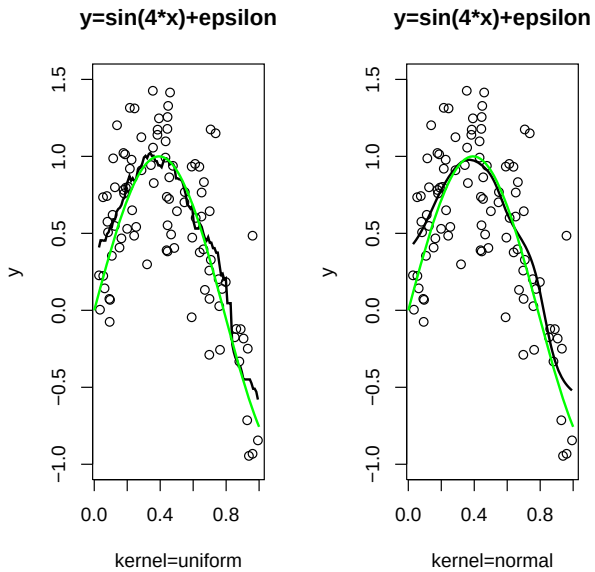
We simply take the average of the y_i corresponding to the x_i in a small neighborhood of x. Each such y_i is assigned the same positive weight.

As previously, λ specifies the size of the neighborhood and is responsible for the amount of smoothing.

Running median: mean is replaced by the median.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Low dimensional and local smoothing - purely kernel smoothing



Smooth weights: Nadaraya-Watson estimator:

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{\lambda}) y_i}{\sum_{j=1}^{n} K(\frac{x-x_j}{\lambda})} = \sum_{i=1}^{n} w_i(x) y_i,$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Choice of a kernel

In general, we assume $K(\cdot)$ to be a probability density. Anything smooth and compact is OK, but under some standard assumptions the optimal choice is the Epanechnikov kernel:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{for } |x| \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

Yet, it is the normal kernel which seems to be most often used.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Choice of the smoothing parameter λ

This is critical to the performance of an estimator.

- For λ too small, the estimator will be too erratic or wiggly as having large variance (due to averaging a small number of observations), but it will have small bias.
- For λ too large, important features will be smoothed out - due to small variance but large bias.
- One may choose λ interactively using the eyeball method: plot $\hat{f}_\lambda(x)$ for a range of different λ's and pick the one that looks best.
- Cross-validation may be used. The criterion is

$$CV(\lambda) = \sum_{i=1}^{n} (y_i - \hat{f}_{\lambda,-i}(x_i))^2,$$

where $-i$ indicates that point i is left out of the fit.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Additive models and projection pursuit regression

Additive models are a multivariate nonparametric modeling attempt to avoid curse of dimensionality. It postulates the following structure

$$Y_i = \alpha + \sum_{j=1}^{p} f_j(x_i^{(j)}) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where f_j are smooth arbitrary functions. We set (and keep this assumption for the additive estimator) $Ef_j(x^{(j)}) = 0$ for identifiability of α.

- More flexible than the linear model but still interpretable since the functions f_j may be plotted.
- Will do poorly when strong interactions exist. In this case one might consider adding, e.g., $f_{ij}(x^{(i)}x^{(j)})$.
- Categorical variables may be incorporated using the usual regression approach.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Additive models

The backfitting algorithm is used to estimate the f_j:

- Initialize: set $\alpha = \bar{Y}$ and initial estimates for f_j, $j = 1, \ldots, p$.
- Cycle $j = 1, \ldots, p, 1, \ldots, p, 1, \ldots$

$$f_j = S(x^{(j)}, y - \alpha - \sum_{k \neq j} f_k(x^{(k)})),$$

  where $S(x, y)$ is a given smoother.
- Repeat until convergence.

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund

Additive models

Remarks (on the algorithm and a bit more):

- $y - \alpha - \sum_{k \neq j} f_k(x^{(k)})$ is a partial residual - the current result of fitting all predictors except predictor x_j.
- The choice of S is left open to the user: could be splines or LOESS, say.
- The algorithm converges under some rather loose conditions.
- Component functions in the additive model can be defined on R^m for some $m < p$, i.e., can have m predictors as their domains.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Projection pursuit regression

Projection pursuit regression (PPR) is way to look adaptively for most promising one-dimensional projections in the space of predictors.

Notice, e.g., that

$$x_1 x_2 = \frac{1}{4}((x_1 + x_2)^2 - (x_1 - x_2)^2)$$

and $x_1 + x_2 = [x_1, x_2][1, 1]'$, $x_1 - x_2 = [x_1, x_2][1, -1]'$. This is a partial case of a general property saying that functions of the form

$$\alpha_0 + \sum_{j=1}^{J} f_j(\alpha_j^{\mathsf{T}} x)$$

approximate arbitrarily well continuous functions on hypercubes. PPR uses this property to approximate unknown regression function by functions of the form given above.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Projection pursuit regression

Curse of dimensionality is avoided (to some extent) as only one dimensional projections are considered.

The f_j's are estimated nonparametrically and J is chosen in an adaptive way. Projection pursuit regression is an iterative algorithm which looks for 'interesting' directions in the data, which can explain the largest part of the remaining variability of Y.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Projection pursuit regression

Projection pursuit regression algorithm:
Let $\hat{f}_{\alpha_j}(\alpha_j^{\mathbf{T}} x) = \hat{f}_j(\alpha_j^{\mathbf{T}} x)$.

- $J := 0$, $\alpha_0 = \bar{Y}$, $Y_i := Y_i - \bar{Y}$, $R_j := Y_j$.
- For any linear combination $z = \alpha^{\mathbf{T}} x$ and sample $(Z_i, R_i), i = 1, 2, \ldots, n$ with $Z_i = \alpha^{\mathbf{T}} X_i$, estimate $\hat{f}_\alpha(z)$ and

$$I(\alpha) = 1 - \frac{\sum_{i=1}^{n} (R_i - \hat{f}_\alpha(Z_i))^2}{\sum_{i=1}^{n} R_i^2} \ (= 1 - SSE/SST).$$

  Find $\alpha_{J+1} = \operatorname{argmax} I(\alpha)$ and store $\hat{f}_{\alpha_{J+1}}(\cdot)$.
- Stop if $I(\alpha) \geqslant$ given threshold,
  otherwise
  $J := J + 1$ and $R_i := R_i - \hat{f}_{\alpha_j}(\alpha_j^{\mathbf{T}} X_i)$ and go to step 2.

Optimization in step 2 is not trivial !

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Regression trees and related methods

Faced with truly multivariate problems we can resort, e.g., to:

- neural networks (note that feed-forward neural networks with one hidden layer are a nonadaptive version of the PPR which, in principle, retains the property of being a universal approximator; cf. Jaroszewicz's Intro to ML and DM)
- support vector machines for regression (to be discussed later)
- regression trees and their improvements.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Regression trees and related methods

Regarding regression trees, let us present them briefly here (and in greater detail during the lecture):

- $\hat{f}(\mathbf{x}) = \bar{y}_{\text{leaf}_k}$ for $\mathbf{x} \in N_k$, where $N_k \subset R^p$ is a hyperectangle specified by the conditions in the k^{th} leaf and $\bar{y}_{\text{leaf}_k}$ is an average of the response for elements of training set falling into this leaf.
- Partition criterion: choose a predictor and a threshold which yields the maximal decrease of Sum of Squared Errors (SSE):

$$\mathcal{R}_L(j,s) = \{\mathbf{x} : x_j \leqslant s\} \qquad \mathcal{R}_R(j,s) = \{\mathbf{x} : x_j > s\},$$

$$\min_{j,s} \min_{c_1,c_2} \{ \sum_{\mathcal{R}_L} (y_i - c_1)^2 + \sum_{\mathcal{R}_R} (y_i - c_2)^2 \},$$

$j = 1, \ldots, p$. The inner minimization is solved by $\hat{c}_i$ equal to averaged response over respective child.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

- trees are grown using cost complexity criterion

$$R_\alpha(T) = SSE + \alpha|T|$$

and then pruned using crossvalidation estimator of SSE.

Drawbacks: difficulty with adopting to a linear structure and introducing high level interaction effects; regression tree fit is discontinuous because step functions are discontinuous.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

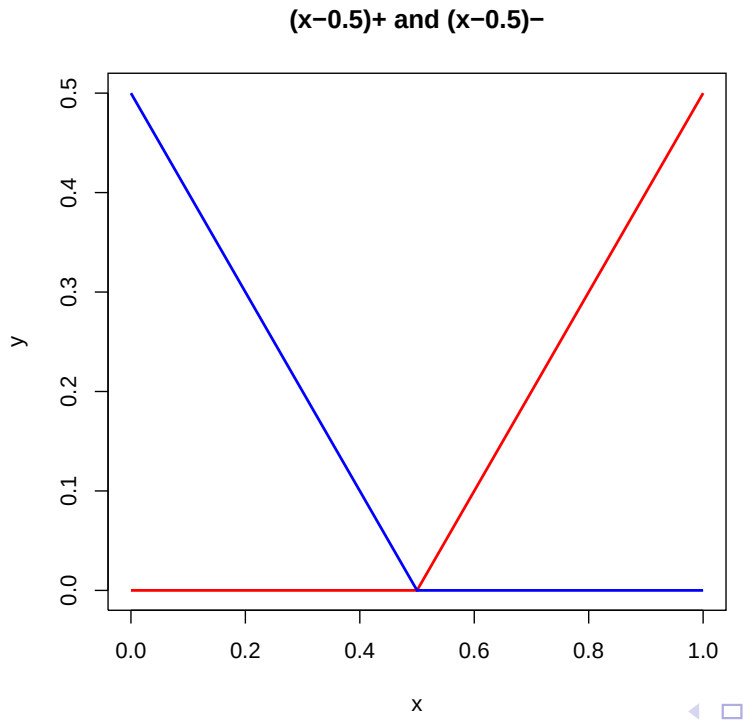
Multivariate Adaptive Regression Splines

Idea behind the Multivariate Adaptive Regression Splines (MARS): Replace step functions with something smoother, actually with

$$(x-t)_+ = \begin{cases} x-t, & \text{if } x > t \\ 0 & \text{otherwise} \end{cases}$$

and

$$(x-t)_- = \begin{cases} t-x, & \text{if } t > x \\ 0 & \text{otherwise} \end{cases}$$



Multivariate Adaptive Regression Splines

The collection of basis functions is

$$\mathcal{C} = \{(X^{(j)} - t)_+, (X^{(j)} - t)_-\}_{t \in \{x_1^{(j)}, \ldots, x_n^{(j)}\}} \quad j = 1, \ldots, p$$

Model building strategy is similar to forward stepwise linear regression, but with one essential difference: at each step we are allowed to use the base functions from the set $\mathcal{C}$ and their products. Thus the model has the form

$$f(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\mathbf{x}),$$

where each $h_m(\mathbf{x})$ is a function from $\mathcal{C}$ or a product of two or more such functions.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Multivariate Adaptive Regression Splines

- At stage 0 we fit a constant, i.e., we start with the model that includes only the constant function $h_0(\mathbf{x}) = 1$.
- At stage 1 we add to the model a function of the form $\beta_1(x^{(j)} - t)_+ + \beta_2(x^{(j)} - t)_-$, where $j \in \{1, 2, \ldots, p\}$ and $t \in \{x_i^{(j)}\}_1^n$. We choose the function which gives the best LS fit for the current residual. The pair of the functions chosen is added to the set of functions $h_m(\mathbf{x})$ present in the earlier model. (Suppose that the best function is $\hat{\beta}_1(x^{(2)} - x_7^{(2)})_+ + \hat{\beta}_2(x^{(2)} - x_7^{(2)})$; hence, in our example, we add functions $h_1(\mathbf{x}) = (x^{(2)} - x_7^{(2)})_+$ and $h_2(\mathbf{x}) = (x^{(2)} - x_7^{(2)})_-$ to the set which contains only $h_0(\mathbf{x})$.)

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Multivariate Adaptive Regression Splines

- We continue in this manner, namely, at each next stage we consider including in the model a pair of products of the form

$$h_m(\mathbf{x})(x^{(j)} - t)_+ \text{ and } h_m(\mathbf{x})(x^{(j)} - t)_-$$

that is, we use the LS fit to add to the model two new summands of the form

$$\beta_1 h_m(\mathbf{x})(x^{(j)} - x_i^{(j)})_+ + \beta_2 h_m(\mathbf{x})(x^{(j)} - x_i^{(j)})_-$$

for some j and i; the pair of the functions chosen,

$$h_m(\mathbf{x})(x^{(j)} - x_i^{(j)})_+ \text{ and } h_m(\mathbf{x})(x^{(j)} - x_i^{(j)})_-$$

is added to the set of functions $h_m(\mathbf{x})$ present in the earlier model.

- Usually, a large model is constructed which is then pruned using similar ideas as in regression trees.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

One more remark on MARS, and one concluding remark

One more remark on MARS: An important property of products of functions from $\mathcal{C}$ is their ability to operate locally. They are 0 over a part of the feature space, unlike, e.g., polynomials. This property allows to fit a parsimonious model.

And an obvious general remark: It should be crystal clear that our exposition of extending regression analysis beyond its linear setup is lacking in both detail and scope. Of the most immediate questions which call for getting answers after having gone through these slides are the following ones: How to look at regularization methods within a more general (mathematical) framework? How to find ML solutions when they cease to be next to trivial (e.g., for mixed effects multilevel models)? How to deal with generalized additive models (which are a natural extension of generalized linear models)? And these are just some of the most immediate questions.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund