SPARROW

SPARse appROximation Wighted regression

Pardis Noorzad

Department of Computer Engineering and IT
Amirkabir University of Technology

Université de Montréal – March 12, 2012
Outline

Introduction
  Motivation
  Local Methods

Sparrow
  SPARROW is a Local Method
  Defining the Effective Weights
  Defining the Observation Weights

Evaluation
Problem setting

- Given $\mathcal{D} := \{(x_i, y_i) : i = 1, \ldots, N\}$
  - $y_i \in \mathbb{R}$ is the output
  - at the input $x_i := [x_{i1}, \ldots, x_{iM}]^T \in \mathbb{R}^M$
- our task is to estimate the regression function

$$f : \mathbb{R}^M \mapsto \mathbb{R}$$

such that

$$y_i = f(x_i) + \epsilon_i$$

the $\epsilon_i$’s are independent with zero mean.
Global methods

- In parametric approaches, the regression function is known
- for e.g., in **multiple linear regression** (MLR) we assume

\[
f(x_0) = \sum_{j=1}^{M} \beta_j x_{0j} + \epsilon
\]

- we can also add higher order terms but still have a model that is linear in the parameters \( \beta_j, \gamma_j \)

\[
f(x_0) = \sum_{j=1}^{M} (\beta_j x_{0j} + \gamma_j x_{0j}^2) + \epsilon
\]
Global methods

Continued

- Example of a global nonparametric approach:

- \( \epsilon \)-support vector regression (\( \epsilon \)-SVR) (Smola and Schölkopf, 2004)

\[
f(x_0) = \sum_{i=1}^{N} \beta_j K(x_0, x_i) + \epsilon
\]
Local methods

- A successful nonparametric approach to regression: local estimation
  (Hastie and Loader, 1993; Härdle and Linton, 1994; Ruppert and Wand, 1994)
- In local methods:

  \[ f(x_0) = \sum_{i=1}^{N} l_i(x_0)y_i + \epsilon \]
Local methods

- For e.g. in \textit{k-nearest neighbor regression} (\textit{k-NNR})

\[
f(x_0) = \sum_{i=1}^{N} \frac{\alpha_i(x_0)}{\sum_{p=1}^{N} \alpha_p(x_0)} y_i
\]

- where \( \alpha_i(x_0) := I_{\mathcal{N}_k(x_0)}(x_i) \)

- \( \mathcal{N}_k(x_0) \subset \mathcal{D} \) is the set of the \textit{k}-nearest neighbors of \( x_0 \)
Local methods

Continued

In weighted $k$-NNR ($Wk$-NNR),

$$f(x_0) = \sum_{i=1}^{N} \frac{\alpha_i(x_0)}{\sum_{p=1}^{N} \alpha_p(x_0)} y_i$$

- $\alpha_i(x_0) := S(x_0, x_i)^{-1} I_{N_k(x_0)}(x_i)$
- $S(x_0, x_i) = (x_0 - x_i)^T V^{-1} (x_0 - x_i)$
  is the scaled Euclidean distance

$S$
Local methods

Continued

- Just so you know, here’s another example of a local method:
- **additive model (AM)**
  (Buja et al., 1989)

\[
f(x_0) = \sum_{j=1}^{M} f_j(x_{0j}) + \epsilon
\]

- Estimate univariate functions of predictors locally
Local methods
Continued

- In local methods: estimate the regression function *locally* by a *simple parametric model*
- In *local polynomial regression*: estimate the regression function locally, by a *Taylor polynomial*
- This is what happens in SPARROW, as we will explain
SPARROW is a Local Method

Sparrow
I meant this sparrow
SPARROW is a local method

- Before we get into the details,
- see a few examples showing benefits of local methods
- then we’ll talk about SPARROW
Figure: Our generated dataset. $y_i = f(x_i) + \epsilon_i$, where $f(x) = (x^3 + x^2) I(x) + \sin(x) I(-x)$. 
Figure: Multiple linear regression with first-, second-, and third-order terms.
Figure: $\epsilon$-support vector regression with an RBF kernel.
Figure: 4-nearest neighbor regression.
Effective weights in SPARROW

- In local methods:

\[ f(x_0) = \sum_{i=1}^{N} l_i(x_0) y_i + \epsilon \]

- Now we define \( l_i(x_0) \)
Local estimation by a Taylor polynomial

- To locally estimate the regression function near $x_0$
- let us approximate $f(x)$ by a second-degree Taylor polynomial about $x_0$

$$P_2(x) = \phi + (x - x_0)^T \theta + \frac{1}{2}(x - x_0)^T H(x - x_0) \quad (1)$$

- $\phi := f(x_0)$,
- $\theta := \nabla f(x_0)$ is the gradient of $f(x)$,
- $H := \nabla^2 f(x_0)$ is its Hessian
- both evaluated at $x_0$
Defining the Effective Weights

Local estimation by a Taylor polynomial
Continued

\[ P_2(x) = \phi + (x - x_0)^T \theta + \frac{1}{2} (x - x_0)^T H (x - x_0) \]

We need to solve the locally weighted least squares problem

\[
\min_{\phi, \theta, H} \sum_{i \in \Omega} \alpha_i \left( y_i - P_2(x_i) \right)^2
\]  \hspace{1cm} (2)
Local estimation by a Taylor polynomial

Continued

- Express (2) as

\[
\min_{\Theta(x_0)} \left\| A^{1/2} \{ y - X\Theta(x_0) \} \right\|^2
\]

- \( a_{ii} = \alpha_i \), \( y := [y_1, y_2, \ldots, y_N]^T \)

\[
\begin{bmatrix}
1 & (x_1 - x_0)^T & \text{vech}^T \{(x_1 - x_0)(x_1 - x_0)^T\} \\
\vdots & \vdots & \vdots \\
1 & (x_N - x_0)^T & \text{vech}^T \{(x_N - x_0)(x_N - x_0)^T\}
\end{bmatrix}
\]

- parameter supervector: \( \Theta(x_0) := [\phi, \theta, \text{vech}(H)]^T \)
Local estimation by a Taylor polynomial

Continued

- The solution:
  \[
  \hat{\Theta}(x_0) = (X^TAX)^{-1}X^T Ay
  \]

- And so the local quadratic estimate is
  \[
  \hat{\phi} = \hat{f}(x_0) = e_1^T(X^TAX)^{-1}X^T Ay
  \]

- Since \( f(x_0) = \sum_{i=1}^{N} l_i(x_0)y_i \),
  the vector of effective weights for SPARROW is
  \[
  [l_1(x_0), \ldots, l_N(x_0)]^T = A^TX(X^TAX)^{-1}e_1
  \]
Local estimation by a Taylor polynomial

Continued

- The local constant regression estimate is

\[
\hat{f}(x_0) = (1^T A 1)^{-1} 1^T A y = \sum_{i=1}^{N} \frac{\alpha_i(x_0)}{\sum_{k=1}^{N} \alpha_k(x_0)} y_i.
\]

- Look familiar?
We have to assign the weights here

\[ \min_{\phi, \theta, H} \sum_{i \in \Omega} \alpha_i \{ y_i - f(x_i) \}^2 \]

that is, the diagonal elements of \( A \)

\[ \min_{\Theta(x_0)} \left\| A^{1/2} \{ y - X\Theta(x_0) \} \right\|^2 \] (4)
Observation weights in SPARROW
Continued

To find $\alpha_i$ we solve the following problem (Chen et al., 1995)

$$\min_{\alpha \in \mathbb{R}^N} \|\alpha\|_1 \quad \text{subject to} \quad \|x_0 - D\alpha\|_2^2 \leq \sigma$$

- $\sigma > 0$ limits the maximum approximation error
- and $D := \begin{bmatrix} \frac{x_1}{\|x_1\|}, & \frac{x_2}{\|x_2\|}, & \cdots, & \frac{x_N}{\|x_N\|} \end{bmatrix}$
Defining the Observation Weights

Power family of penalties

\( \ell_p \) norms raised to the \( p \)th power

\[
\|x\|_p^p = \left( \sum_i |x_i|^p \right)
\]  

(6)

- For \( 1 \leq p < \infty \), (6) is convex.
- \( 0 < p \leq 1 \), is the range of \( p \) useful for measuring sparsity.
Figure: As $p$ goes to 0, $|x|^p$ becomes the indicator function and $|x|^p$ becomes a count of the nonzeros in $x$ (Bruckstein et al., 2009).
To motivate this idea let’s look at
- feature learning with **sparse coding**, and
- **sparse representation classification** (SRC)
  - an example of **exemplar-based sparse approximation**
Unsupervised feature learning
Application to image classification

\[ x_0 = D\alpha \]

- An example is the recent work by Coates and Ng (2011).
  - where \( x_0 \) is the input vector
  - could be a vectorized image patch, or a SIFT descriptor
  - \( \alpha \) is the **higher-dimensional sparse representation** of \( x_0 \)
  - \( D \) is usually learned
Figure: Image classification (Coates et al., 2011).
Defining the Observation Weights

**Multiclass classification**
(Wright et al., 2009)

- \( \mathcal{D} := \left\{ (x_i, y_i) : x_i \in \mathbb{R}^m, y_i \in \{1, \ldots, c\}, i \in \{1, \ldots, N\} \right\} \)

- Given a test sample \( x_0 \)
  1. Solve \( \min_{\alpha \in \mathbb{R}^N} \|\alpha\|_1 \) subject to \( \|x_0 - D\alpha\|_2^2 \leq \sigma \)
  2. Define \( \{\alpha_y : y \in \{1, \ldots, c\}\} \) where \( [\alpha_y]_i = \alpha_i \) if \( x_i \) belongs to class \( y \), o.w. 0
  3. Construct \( \mathcal{X}(\alpha) := \left\{ \hat{x}_y(\alpha) = D\alpha_y, y \in \{1, \ldots, c\} \right\} \)
  4. Predict \( \hat{y} := \arg \min_{y \in \{1, \ldots, c\}} \|x_0 - \hat{x}_y(\alpha)\|_2^2 \)
Figure: SRC on handwritten image dataset.
Figure: SVM with linear kernel on handwritten image dataset.
Back to SPARROW with evaluation on the MPG dataset

- Auto MPG Data Set
- from the UCI Machine Learning Repository (Frank and Asuncion, 2010)
- “The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 4 continuous attributes.”
- number of instances: 392
- number of attributes: 7 (cylinders, displacement, horsepower, weight, acceleration, model year, origin)
Figure: Average mean squared error values achieved by various methods over 10-fold cross-validation.
Looking ahead

- What is causing the success of SPARROW and SRC?
- How important is the bandwidth? What about in SRC?


This is ongoing work carried out under the supervision of Prof. Bob L. Sturm of Aalborg University Copenhagen.

Thanks to Isaac Nickaein and Sheida Bijani for helping out with the slides.