SPARROW

SPARse app**RO**ximation **W**eighted 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 $\mathcal{D} := \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$
  - $y_i \in \mathbb{R}$ is the output
  - at the input $\mathbf{x}_i := [x_{i1}, \ldots, x_{iM}]^\mathsf{T} \in \mathbb{R}^M$
- our task is to estimate the regression function

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

such that

$$y_i = f(\mathbf{x}_i) + \epsilon_i$$

the ϵ_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

$$f(\mathbf{x}_0) = \sum_{j=1}^{M} \beta_j x_{0j} + \epsilon$$

- we can also add higher order terms but still have a model that is linear in the parameters β_j, γ_j

$$f(\mathbf{x}_0) = \sum_{j=1}^{M} \left(\beta_j x_{0j} + \gamma_j x_{0j}^2\right) + \epsilon$$

Global methods

Continued

- Example of a global nonparametric approach:
- ϵ-**support vector regression** (ϵ-SVR) (Smola and Schölkopf, 2004)

$$f(\mathbf{x}_0) = \sum_{i=1}^{N} \beta_j K(\mathbf{x}_0, \mathbf{x}_i) + \epsilon$$

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:

$$f(\mathbf{x}_0) = \sum_{i=1}^{N} l_i(\mathbf{x}_0) y_i + \epsilon$$

Local methods

Continued

- For e.g. in k-**nearest neighbor regression** (k-NNR)

$$f(\mathbf{x}_0) = \sum_{i=1}^{N} \frac{\alpha_i(\mathbf{x}_0)}{\sum_{p=1}^{N} \alpha_p(\mathbf{x}_0)} y_i$$

  - where $\alpha_i(\mathbf{x}_0) := \mathrm{I}_{\mathcal{N}_k(\mathbf{x}_0)}(\mathbf{x}_i)$
  - $\mathcal{N}_k(\mathbf{x}_0) \subset \mathcal{D}$ is the set of the k-nearest neighbors of $\mathbf{x}_0$

Local methods

Continued

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

$$f(\mathbf{x}_0) = \sum_{i=1}^{N} \frac{\alpha_i(\mathbf{x}_0)}{\sum_{p=1}^{N} \alpha_p(\mathbf{x}_0)} y_i$$

  - $\alpha_i(\mathbf{x}_0) := S(\mathbf{x}_0, \mathbf{x}_i)^{-1}\, \mathrm{I}_{\mathcal{N}_k(\mathbf{x}_0)}(\mathbf{x}_i)$
  - $S(\mathbf{x}_0, \mathbf{x}_i) = (\mathbf{x}_0 - \mathbf{x}_i)^{\mathsf{T}} \mathbf{V}^{-1} (\mathbf{x}_0 - \mathbf{x}_i)$ is the scaled Euclidean distance

Local methods

Continued

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

$$f(\mathbf{x}_0) = \sum_{j=1}^{M} f_j(x_{0j}) + \epsilon$$

- 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

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

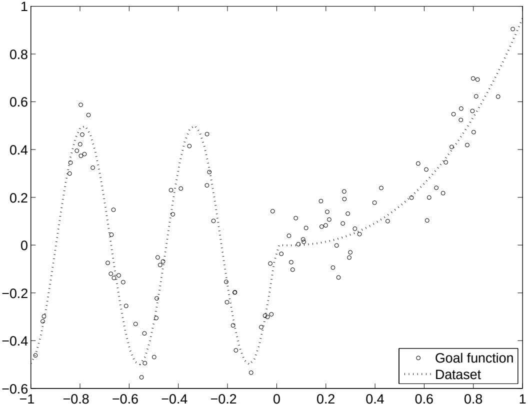

Figure: Our generated dataset. $y_i = f(x_i) + \epsilon_i$, where $f(x) = (x^3 + x^2)\,\mathrm{I}(x) + \sin(x)\,\mathrm{I}(-x)$.

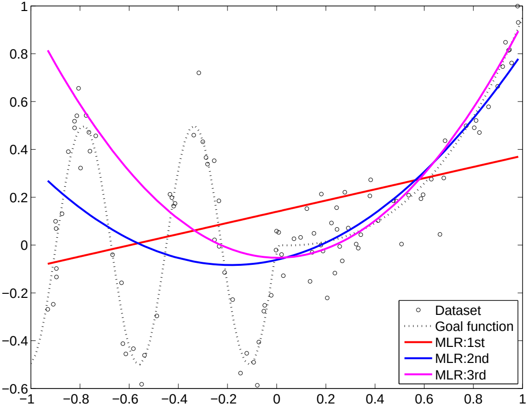

Figure: Multiple linear regression with first-, second-, and third-order terms.

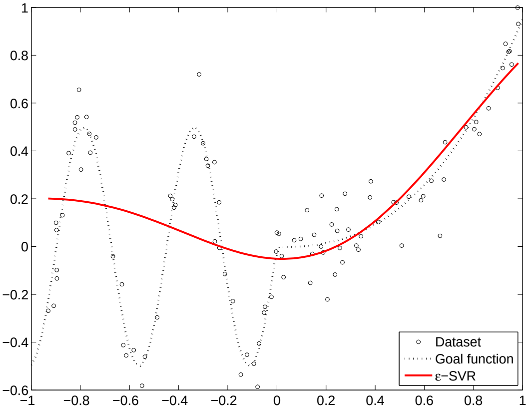

Figure: ϵ-support vector regression with an RBF kernel.

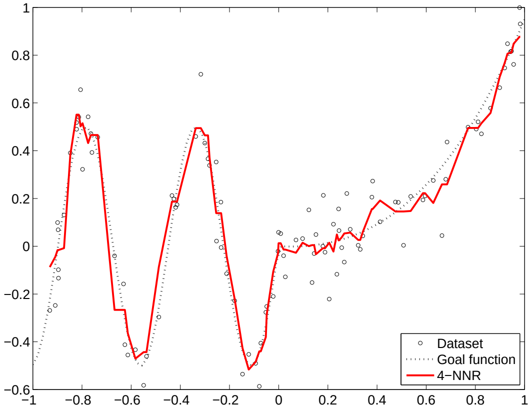

Figure: 4-nearest neighbor regression.

Effective weights in SPARROW

- In local methods:

$$f(\mathbf{x}_0) = \sum_{i=1}^{N} l_i(\mathbf{x}_0) y_i + \epsilon$$

- Now we define $l_i(\mathbf{x}_0)$

Local estimation by a Taylor polynomial

- To locally estimate the regression function near $\mathbf{x}_0$
- let us approximate $f(\mathbf{x})$ by a second-degree Taylor polynomial about $\mathbf{x}_0$

$$\mathrm{P}_2(\mathbf{x}) = \phi + (\mathbf{x} - \mathbf{x}_0)^\mathsf{T}\boldsymbol{\theta} + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\mathsf{T}\mathbf{H}(\mathbf{x} - \mathbf{x}_0) \tag{1}$$

- $\phi := f(\mathbf{x}_0)$,
  $\boldsymbol{\theta} := \nabla f(\mathbf{x}_0)$ is the gradient of $f(\mathbf{x})$,
  $\mathbf{H} := \nabla^2 f(\mathbf{x}_0)$ is its Hessian
  both evaluated at $\mathbf{x}_0$

Local estimation by a Taylor polynomial

Continued

$$\mathrm{P}_2(\mathbf{x}) = \phi + (\mathbf{x} - \mathbf{x}_0)^\mathsf{T}\boldsymbol{\theta} + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\mathsf{T}\mathbf{H}(\mathbf{x} - \mathbf{x}_0)$$

- We need to solve the locally weighted least squares problem

$$\min_{\phi,\boldsymbol{\theta},\mathbf{H}} \sum_{i \in \Omega} \alpha_i \{y_i - \mathrm{P}_2(\mathbf{x}_i)\}^2 \tag{2}$$

Local estimation by a Taylor polynomial

Continued

- Express (2) as

$$\min_{\boldsymbol{\Theta}(\mathbf{x}_0)} \left\| \mathbf{A}^{1/2} \{ \mathbf{y} - \mathbf{X}\boldsymbol{\Theta}(\mathbf{x}_0) \} \right\|^2 \tag{3}$$

- $a_{ii} = \alpha_i$, $\mathbf{y} := [y_1, y_2, \ldots, y_N]^\mathsf{T}$
- $\mathbf{X} := \begin{bmatrix} 1 & (\mathbf{x}_1 - \mathbf{x}_0)^\mathsf{T} & \mathrm{vech}^\mathsf{T}\{(\mathbf{x}_1 - \mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0)^\mathsf{T}\} \\ \vdots & \vdots & \vdots \\ 1 & (\mathbf{x}_N - \mathbf{x}_0)^\mathsf{T} & \mathrm{vech}^\mathsf{T}\{(\mathbf{x}_N - \mathbf{x}_0)(\mathbf{x}_N - \mathbf{x}_0)^\mathsf{T}\} \end{bmatrix}$
- parameter supervector: $\boldsymbol{\Theta}(\mathbf{x}_0) := \left[\phi, \boldsymbol{\theta}, \mathrm{vech}(\mathbf{H}) \right]^\mathsf{T}$

Local estimation by a Taylor polynomial

Continued

- The solution:

$$\widehat{\mathbf{\Theta}}(\mathbf{x}_0) = \left(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{y}$$

- And so the local quadratic estimate is

$$\hat{\phi} = \hat{f}(\mathbf{x}_0) = \mathbf{e}_1^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{y}$$

- Since $f(\mathbf{x}_0) = \sum_{i=1}^{N} l_i(\mathbf{x}_0)y_i$,
  the vector of effective weights for SPARROW is

$$[l_1(\mathbf{x}_0), \ldots, l_N(\mathbf{x}_0)]^\mathsf{T} = \mathbf{A}^\mathsf{T}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{X}\right)^{-1}\mathbf{e}_1$$

Local estimation by a Taylor polynomial

Continued

- The local constant regression estimate is

$$\hat{f}(\mathbf{x}_0) = (\mathbf{1}^\mathsf{T}\mathbf{A}\mathbf{1})^{-1}\mathbf{1}^\mathsf{T}\mathbf{A}\mathbf{y} = \sum_{i=1}^{N} \frac{\alpha_i(\mathbf{x}_0)}{\sum_{k=1}^{N} \alpha_k(\mathbf{x}_0)} y_i.$$

- Look familiar?

Observation weights in SPARROW

- We have to assign the weights here

$$\min_{\phi,\boldsymbol{\theta},\mathbf{H}} \sum_{i \in \Omega} \alpha_i \{y_i - f(\mathbf{x}_i)\}^2$$

- that is, the diagonal elements of $\mathbf{A}$

$$\min_{\boldsymbol{\Theta}(\mathbf{x}_0)} \left\| \mathbf{A}^{1/2} \{\mathbf{y} - \mathbf{X}\boldsymbol{\Theta}(\mathbf{x}_0)\} \right\|^2 \tag{4}$$

Observation weights in SPARROW

Continued

- To find α_i we solve the following problem (Chen et al., 1995)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \|\boldsymbol{\alpha}\|_1 \text{ subject to } \|\mathbf{x}_0 - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \sigma \tag{5}$$

  - $\sigma > 0$ limits the maximum approximation error
  - and $\mathbf{D} := \left[\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}, \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}, \ldots, \frac{\mathbf{x}_N}{\|\mathbf{x}_N\|}\right]$

Power family of penalties

ℓ_p norms raised to the pth power

$$\|\mathbf{x}\|_p^p = \left(\sum_i |x_i|^p \right) \tag{6}$$

- For $1 \leq p < \infty$, (6) is convex.
- $0 < p \leq 1$, is the range of p useful for measuring sparsity.

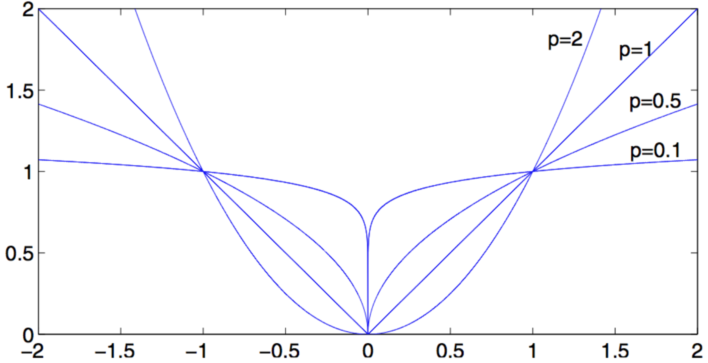

Figure: As p goes to 0, $|x|^p$ becomes the indicator function and $|\mathbf{x}|^p$ becomes a count of the nonzeros in $\mathbf{x}$ (Bruckstein et al., 2009).

Representation by sparse approximation

Continued

- 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

$$\mathbf{x}_0 = \mathbf{D}\boldsymbol{\alpha}$$

- An example is the recent work by Coates and Ng (2011).
  - where $\mathbf{x}_0$ is the input vector
  - could be a vectorized image patch, or a SIFT descriptor
  - $\boldsymbol{\alpha}$ is the **higher-dimensional sparse representation** of $\mathbf{x}_0$
  - $\mathbf{D}$ is usually learned

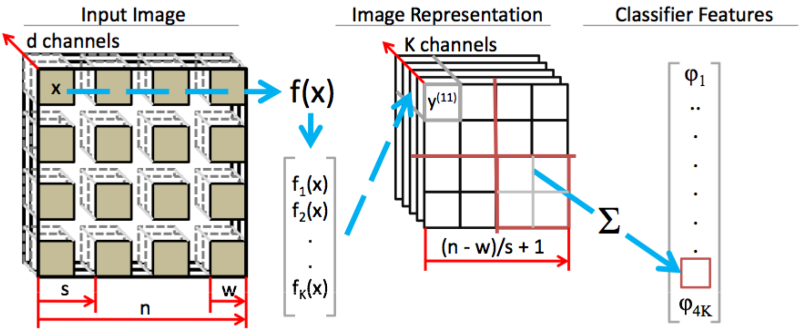

Figure: Image classification (Coates et al., 2011).

Multiclass classification

(Wright et al., 2009)

- $\mathcal{D} := \left\{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^m, y_i \in \{1, \ldots, c\}, i \in \{1, \ldots, N\}\right\}$
- Given a test sample $\mathbf{x}_0$
    1. Solve $\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \|\boldsymbol{\alpha}\|_1$ subject to $\|\mathbf{x}_0 - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \sigma$
    2. Define $\{\boldsymbol{\alpha}_y : y \in \{1, \ldots, c\}\}$ where $[\boldsymbol{\alpha}_y]_i = \boldsymbol{\alpha}_i$ if $\mathbf{x}_i$ belongs to class y, o.w. 0
    3. Construct $\mathcal{X}(\boldsymbol{\alpha}) := \left\{\hat{\mathbf{x}}_y(\boldsymbol{\alpha}) = \mathbf{D}\boldsymbol{\alpha}_y, y \in \{1, \ldots, c\}\right\}$
    4. Predict $\hat{y} := \arg\min_{y \in \{1, \ldots, c\}} \|\mathbf{x}_0 - \hat{\mathbf{x}}_y(\boldsymbol{\alpha})\|_2^2$

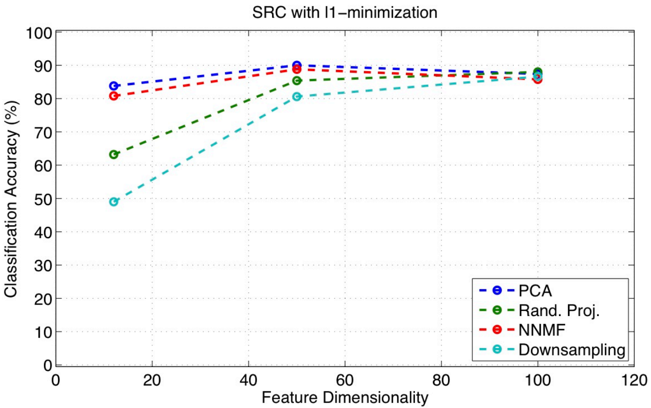

Figure: SRC on handwritten image dataset.

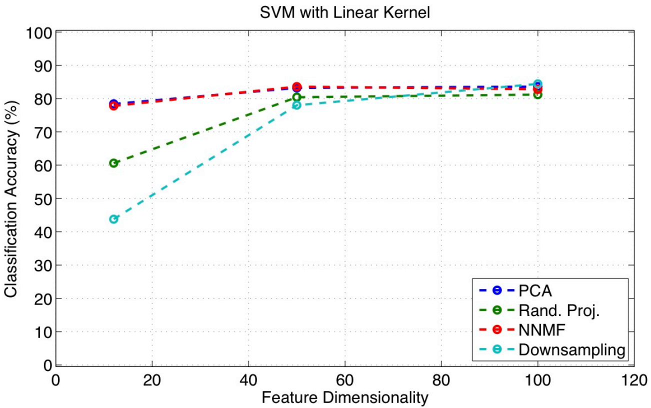

Figure: SVM with linear kernel on handwritten image dataset.

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)

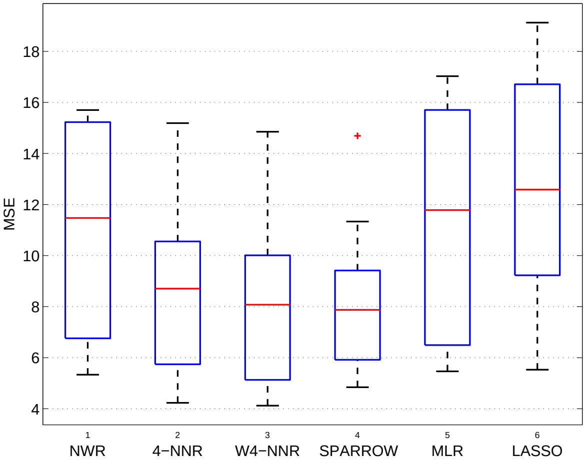

Figure: Average mean squared error values achieved by various methods over 10-fold cross-validation.

Looking ahead

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

References I

Alfred M. Bruckstein, David L. Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):435–555, 1989.

Scott S. Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, May 1995.

Adam Coates and Andrew Ng. The importance of encoding versus training with sparse coding and vector quantization. In *International Conference on Machine Learning (ICML)*, pages 921–928, 2011.

Adam Coates, Honglak Lee, and Andrew Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *International Conference on AI and Statistics*, 2011.

References II

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

Wolfgang Härdle and Oliver Linton. Applied nonparametric methods. Technical Report 1069, Yale University, 1994.

T. J. Hastie and C. Loader. Local regression: Automatic kernel carpentry. *Statistical Science*, 8(2):120–129, 1993.

D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22:1346–1370, 1994.

Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, August 2004.

John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, 2009.

Acknowledgements

- 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.