Nonparametric Methods III

Wei Li

Syracuse University

Spring 2024

OVERVIEW

Radial Basis Functions (RBF) network

Nonparametric classification

Nonparametric additive models

Variable selection in nonparametric regression

Radial Basis Functions (RBF) network

Radial Basis Functions (RBF) network

For basis expansion, functions are represented as expansions in basis functions, $x \in \mathbb{R}^p$:

$$f(x) = \sum_{j=1}^{M} \beta_j h(x; \gamma_j)$$

In single-hidden-layer neural networks

- $h(x; \gamma) = \sigma\left(\gamma_0 + \gamma_1^\top x\right)$, where $\sigma(t) = 1/\left(1 + e^{-t}\right)$ (with $M = 1$) is the sigmoid function (logistic function)
- γ parameterizes a linear combination of the predictors.

Radial basis expansion generalize these ideas, by treating the kernel functions $K_\lambda(\xi, x)$ as basis functions. This leads to the model

$$
\begin{aligned}
f(x) &= \sum_{j=1}^{M} K_{\lambda_j}(\xi_j, x)\, \beta_j \\
&= \sum_{j=1}^{M} \bar{K}\left(\frac{\|x - \xi_j\|}{\lambda_j}\right) \beta_j
\end{aligned}
$$

where each basis element is indexed by

- location or prototype parameter ξ_j
- a scale parameter λ_j.

RBF network

To estimate $\{\lambda_j, \xi_j, \beta_j\}$, $j = 1, \ldots, M$, optimize the sum-of-squares with respect to all the parameters:

$$\min_{\beta_0, (\lambda_j, \xi_j, \beta_j)_{j=1}^M} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{M} \beta_j \exp\Big(- \frac{(x_i - \xi_j)^\top (x_i - \xi_j)}{\lambda_j^2} \Big) \right)^2$$

Often, $\bar{K}$ is replaced by the *renormalized radial basis functions*:

$$h_j(x) = \frac{\bar{K}\left(\left\|x - \xi_j\right\| / \lambda\right)}{\sum_{k=1}^{M} \bar{K}\left(\left\|x - \xi_k\right\| / \lambda\right)}$$

The Nadaraya-Watson kernel regression estimator in $\mathbb{R}^p$ can be viewed as an expansion in renormalized radial basis functions,

$$\begin{aligned}
\hat{f}\left(x_0\right) &= \sum_{i=1}^{n} y_i \frac{K_\lambda\left(x_0, x_i\right)}{\sum_{j=1}^{n} K_\lambda\left(x_0, x_j\right)} \\
&= \sum_{i=1}^{n} y_i h_i\left(x_0\right)
\end{aligned}$$

- a basis function h_i located at every observation and coefficients y_i
  - $M = n$
  - $\xi_i = x_i, \hat{\beta}_i = y_i, i = 1, \ldots, n.$

Nonparametric classification

knn classifier

For any given $X = x_0$, we find the K closest neighbors to $X = x_0$ in the training data, and examine their corresponding Y.

$$P(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in N_K(x_0)} 1(y_i = j)$$

Estimate the conditional probability for group j by the proportion out of the k neighbors that are in group j.

Kernel density classification

Suppose for a J class problem, we fit nonparametric density estimates $\hat{f}_j(X), j = 1, \ldots, J$ separately in each of the classes, and we also have estimates of the class priors $\hat{\pi}_j$ (usually the sample proportions).

$$\hat{\Pr}(Y = j \mid X = x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\sum_{k=1}^{J} \hat{\pi}_k \hat{f}_k(x_0)}$$

Nonparametric logistic regression

Let $Y \in \{0, 1\}$.

$$f(x) = \log\left(\frac{Pr(Y = 1 \mid X = x)}{Pr(Y = 0 \mid X = x)}\right)$$

Therefore, $p(x) = Pr(Y = 1|x) = \frac{e^{f(x)}}{1+e^{f(x)}}$.

Logistic (cubic) smoothing spline estimate is defined by

$$\min_f -\ell(f) = \min_f \sum_{i=1}^{n} \left(-y_i f\left(x_i\right) + \log\left(1 + e^{-f(x_i)}\right)\right) + \frac{\lambda}{2} \int \left(f^{(2)}(x)\right)^2 dx$$

- $N_1, \ldots, N_n$: the natural cubic spline basis
- the basis matrix: $\mathbf{N} \in \mathbb{R}^{n \times n}$
- penalty matrix: $\Omega \in \mathbb{R}^{n \times n}$

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

$\mathbf{p}$ is the n-vector with elements $p(x_i;\theta)$, $\mathbf{W}$ is a diagonal matrix of weights $p(x_i;\theta)(1-p(x_i;\theta))$

$$
\begin{aligned}
\frac{\partial(-\ell(\theta))}{\partial\theta} &= -\mathbf{N}^\top(\mathbf{y}-\mathbf{p}) + \lambda\boldsymbol{\Omega}\theta \\
\frac{\partial^2(-\ell(\theta))}{\partial\theta\partial\theta^\top} &= \mathbf{N}^\top\mathbf{W}\mathbf{N} + \lambda\boldsymbol{\Omega}
\end{aligned}
$$

The gradient descent update and the Newton's update are respecitively

$$
\theta^{(k+1)} = \theta^{(k)} + \alpha \times \left(\mathbf{N}^\top(\mathbf{y}-\mathbf{p}^{(k)}) - \lambda\boldsymbol{\Omega}\theta^{(k)}\right)
$$

$$
\begin{aligned}
\theta^{(k+1)} &= \theta^{(k)} + \left(\mathbf{N}^\top\mathbf{W}^{(k)}\mathbf{N} + \lambda\boldsymbol{\Omega}\right)^{-1}\left(\mathbf{N}^\top(\mathbf{y}-\mathbf{p}^{(k)}) - \lambda\boldsymbol{\Omega}\theta^{(k)}\right) \\
&= \left(\mathbf{N}^\top\mathbf{W}^{(k)}\mathbf{N} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{N}^\top\mathbf{W}^{(k)}\left(\mathbf{N}\theta^{(k)} + {\mathbf{W}^{(k)}}^{-1}(\mathbf{y}-\mathbf{p}^{(k)})\right) \\
&= \left(\mathbf{N}^\top\mathbf{W}^{(k)}\mathbf{N} + \lambda\boldsymbol{\Omega}\right)^{-1}\mathbf{N}^\top\mathbf{W}^{(k)}\mathbf{z}^{(k)}
\end{aligned}
$$

iteratively reweighted (penalized) LS

Newton's update $\qquad \theta^{(k+1)} = \left(\mathbf{N}^\top \mathbf{W}^{(k)} \mathbf{N} + \lambda \boldsymbol{\Omega}\right)^{-1} \mathbf{N}^\top \mathbf{W}^{(k)} \mathbf{z}^{(k)}$

$$
\begin{aligned}
\mathbf{f}^{(k+1)} &= \mathbf{N} \left(\mathbf{N}^\top \mathbf{W}^{(k)} \mathbf{N} + \lambda \boldsymbol{\Omega}\right)^{-1} \mathbf{N}^\top \mathbf{W}^{(k)} \left(\mathbf{f}^{(k)} + \mathbf{W}^{(k)^{-1}} (\mathbf{y} - \mathbf{p}^{(k)})\right) \\
&= \mathbf{S}_{\lambda,\mathbf{w}}^{(k)} \mathbf{z}^{(k)}
\end{aligned}
$$

The Newton's update fits a weighted smoothing spline to the adjusted response z:

$$
\min_f \mathrm{RSS}(f, \lambda) = \sum_{i=1}^{n} w_i \big(z_i - f(x_i)\big)^2 + \lambda \int \left(f^{(2)}(x)\right)^2 dx
$$

Nonparametric additive models

In the regression setting, a generalized additive model has the form

$$\mathrm{E}\left(Y \mid X_1, X_2, \ldots, X_p\right)=\alpha+f_1\left(X_1\right)+f_2\left(X_2\right)+\cdots+f_p\left(X_p\right)$$

Let $\mu(X)=E(Y|X)$. The generalized additive models:

$$g(\mu(X))=\alpha+\sum_{j=1}^{p} f_j\left(X_j\right)$$

- $g(\mu)=\mu$: additive model for Gaussian response data.
- $g(\mu)=\operatorname{logit}(\mu)$ or $g(\mu)=\operatorname{probit}(\mu)$: logistic / probit additive models for binary response data.
- $g(\mu)=\log(\mu)$: log-additive model for Poisson count data.

Fitting additive models

$$Y = \alpha + \sum_{j=1}^{p} f_j\left(X_j\right) + \varepsilon$$

Penalized sum of squares:

$$\sum_{i=1}^{n} \left\{ y_i - \alpha - \sum_{j=1}^{p} f_j\left(x_{ij}\right) \right\}^2 + \sum_{j=1}^{p} \lambda_j \int \left(f_j^{(2)}\left(t_j\right) \right)^2 dt_j$$

where $\lambda_j \geq 0$ are tuning parameters.

The minimizer is an additive cubic spline model; each of the functions f_j is a cubic spline.

- α is not identified.
  - assume $\sum_{i=1}^{n} f_j\left(x_{ij}\right) = 0$ for any j (thus $\hat{\alpha} = \bar{y}$).

Back-fitting algorithm

For any j, $\mathrm{E}(Y - \alpha - \sum_{k \neq j} f_k(X_k) | X_j) = f_j(X_j)$.

Suppose our univariate smoothing algorithm smooth(z, y) has been chosen (smooth$(z, y) = \hat{E}(Y = y | Z = z)$).

We initialize $\hat{f}_1, \ldots, \hat{f}_p$ (say, to all to zero), let $\hat{\alpha} = \bar{y}$:

cycle over the following steps for $j = 1, \ldots, p, 1, \ldots, p, \ldots$

- define the response $r_i = y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k\left(x_{ik}\right), i = 1, \ldots, n$
- smooth $\hat{f}_j \leftarrow$ fitted smooth$\left(\mathbf{x}_j, r\right)$, where $\mathbf{x}_j = \left(x_{11}, \ldots, x_{nj}\right), r = \left(r_1, \ldots, r_n\right)$.
- center $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_j\left(x_{ij}\right)$

Generalized additive logistic regression

$$\log \frac{\Pr(Y=1 \mid X)}{\Pr(Y=0 \mid X)} = \eta(x) = \alpha + f_1(X_1) + \cdots + f_p(X_p)$$

smoothing splines solution:

$$\hat{f} = \operatorname*{argmin}_{f_1,\ldots,f_p} \sum_{i=1}^{n} \left(-y_i \eta(x_i) + \log\left(1 + e^{-\eta(x_i)}\right)\right) + \frac{\lambda}{2} \sum_{j=1}^{p} \int \left(f_j^{(2)}(t_j)\right)^2 dt_j$$

Algorithm: IRLS (iteratively reweighted least squares) + weighted backfitting

- update adjusted response $\{z_i\}$ and weights $\{w_i\}$ (IRLS loop)
  - update components $\{\hat{f}_j\}$ (backfitting loop)

Inference

Let $\mathrm{E}(Y = 1|X) = \theta_0 + \sum_{j=1}^{p} f_j(X_j)$.

- $\{\theta_{jk} : k = 1, \ldots, M_j\}$
- $h_j = \{h_{jk} : k = 1, \ldots, M_j\}$
- $\theta = (\theta_0, \theta_1^\top, \ldots, \theta_p^\top)^\top$
- $\mathbf{H}$ be the $n \times (1 + M)$ basis matrix ($M = \sum_{j=1}^{M} M_j$).

For $\hat{f}_j(x_j) = h_j^\top(x_j)\hat{\theta}_j$,

- variance $var(\hat{f}_j(x_j)) = h_j^\top(x_j)\hat{\Sigma}_{j,j}h_j(x_j)$.
  - $\hat{\Sigma}_{j,j}$ is the corresponding (θ_j) sub-matrix of $\hat{\Sigma}$
  - $cov(\hat{\theta}) := \hat{\Sigma} = (\mathbf{H}^\top\mathbf{H})^{-1}$
- pointwise confidence interval (biased): $\hat{f}_j(x_j) \pm z_{\alpha/2}\sqrt{var(\hat{f}_j(x_j))}$.

Inference (logistic regression)

Let $\text{logit}(Pr(Y=1|X)) = \theta_0 + \sum_{j=1}^{p} f_j(X_j)$,
$f_j(x_j) = \sum_{k=1}^{M_j} \theta_{jk} h_{jk}(x_j)$

- $\{\theta_{jk} : k = 1, \ldots, M_j\}$
- $h_j = \{h_{jk} : k = 1, \ldots, M_j\}$
- $\theta = (\theta_0, \theta_1^\top, \ldots, \theta_p^\top)^\top$
- $\mathbf{H}$ be the $n \times (1+M)$ basis matrix ($M = \sum_{j=1}^{M} M_j$).

$$cov(\hat{\theta}) = \hat{\Sigma} = (\mathbf{H}^\top \mathbf{W} \mathbf{H})^{-1}$$

For $\hat{f}_j(x_j) = h_j^\top(x_j)\hat{\theta}_j$,

- variance $var(\hat{f}_j(x_j)) = h_j^\top(x_j)\hat{\Sigma}_{j,j} h_j(x_j)$.
- pointwise confidence interval (biased): $\hat{f}_j(x_j) \pm z_{\alpha/2}\sqrt{var(\hat{f}_j(x_j))}$.

Alleviation of the Curse of Dimensionality

If the true function is *additive*, and each component function is s-times differentiable, then the optimal MSE rate achievable becomes $pn^{-2s/(2s+1)}$.

- p does not appear in the exponent in the rate
- p times univariate optimal rate!

See later on *deep* neural network, the curse of dimensionality can be circumvented if f has a composition and sparse structure.

Variable selection in nonparametric regression

Variable selection in nonparametric regression

$$f(x) = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$$

Claim X_j as unimportant if the function $f_j = 0$

Two-way interaction model

$$f(x) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k)$$

The interaction effect between X_j and X_k is unimportant if $f_{jk} = 0$.

- Multivariate Adaptive Regression Splines (MARS) (Friedman 1991)
- Classification and Regression Tree (CART, Brieman 1985) (not quite do the job)
- Group-LASSO Methods (Huang et al. 2010)
- Sparse Additive Models (Ravikuma et al. 2009)
  - Sparse logistic additive models