Nonparametric Methods

Wei Li

Syracuse University

Spring 2021

Nonparametric classification

Nonparametric logistic regression

Additive models

Variable selection in nonparametric regression

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.

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

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

Nonparametric density estimation

- Data $X_1, \ldots, X_n$ are contained in the unit cube $\mathcal{X} = [0, 1]^p$.
- Divide $\mathcal{X}$ into bins, or sub-cubes, of length h.
- There are $M \approx (1/h)^p$ such bins and each has volume h^p.
- Denote the bins by $B_1, \ldots, B_M$.

1. Assuming the density estimate should be constant in each cube.
2. Letting that constant value be proportional to the number of observations falling in the cube

Roughly, this gives a heuristic estimator for a given point $x \in B_j$:

$$\widehat{p}_n(x) = \frac{\text{number of observations within } B_\ell}{n} \times \frac{1}{\text{volume of the bin}}$$

The **histogram density estimator** is

$$\widehat{p}_h(x) = \sum_{j=1}^{M} \frac{\widehat{\pi}_j}{h^p} I\left(x \in B_j\right)$$

where

$$\widehat{\pi}_j = \#\{i : X_i \in B_j\}/n$$

is the fraction of data points in bin B_j.

Parzen estimate

Suppose $p \geq 1$. The smooth **Parzen** estimate is

$$\hat{f}_X(x) = \frac{1}{nh^p} \sum_{i=1}^{n} K_h(x, x_i)$$

Here, $K_h(x, y) = \bar{K}(\|x - y\|/h)$ for some kernel function $\bar{K}$.

The kernel is assumed to satisfy

- $\int \bar{K}(x)dx = 1$, $\int x\bar{K}(x)dx = 0$
- $\sigma_{\bar{K}}^2 \equiv \int x^2 \bar{K}(x)dx > 0$.

Some commonly used kernels are the following:

$$
\begin{array}{ll}
\text{Boxcar:} & \bar{K}(x) = \frac{1}{2} 1\{x : |x| \leq 1\} \\
\text{Gaussian:} & \bar{K}(x) = \frac{1}{\sqrt{2}} e^{-x^2/2} \\
\text{Epanechnikov:} & \bar{K}(x) = \frac{3}{4}\left(1 - x^2\right) 1\{x : |x| \leq 1\} \\
\text{Tricube:} & \bar{K}(x) = \frac{70}{81}\left(1 - |x|^3\right)^3 1\{x : |x| \leq 1\}
\end{array}
$$

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).

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

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 smoothing spline estimate of polynomial degree 3 is defined by

$$\hat{f} = \operatorname*{argmin}_{f} \sum_{i=1}^{n} \left(-y_i f\left(x_i\right) + \log\left(1 + e^{-f\left(x_i\right)}\right)\right) + \frac{\lambda}{2} \left(f^{(2)}(x)\right)^2 dx$$

- $N_1, \ldots, N_n$ 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 $p(x_i)$,
- $\mathbf{W}$ is a diagonal matrix of weights $p(x_i)(1-p(x_i))$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{N}^T(\mathbf{y}-\mathbf{p}) - \lambda \boldsymbol{\Omega}\theta$$
$$\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} = -\mathbf{N}^T\mathbf{W}\mathbf{N} - \lambda \boldsymbol{\Omega}$$

The update equation is

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

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

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:

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

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

Suppose our univariate smoother $Smooth(z, y)$ 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 $\hat{f}_j \leftarrow \operatorname{Smooth}\left(\mathbf{x}_j, r\right)$, 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

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

Consider using smoothing splines solution:

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

Inference

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

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

We have

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

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

- its variance $var(\hat{f}_j(x_j)) = h_j^T(x_j)\hat{\Sigma}_{j,j} h_j(x_j)$.
- The pointwise confidence band (biased): $\hat{f}(x_j) \pm 2\sqrt{var(\hat{f}_j(x_j))}$.

Alleviation of the Curse of Dimensionality

If the true function is indeed 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 neural network, the curse of dimensionality can be similarly circumvented.

Variable selection in nonparametric regression

Variable selection in nonparametric regression

Additive models

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

Claim X_i as unimportant if the function $f_i = 0$

Two-way interaction model

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

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)
- Goup-LASSO Methods (Huang et al. 2010)
- Sparse Additive Models (Ravikuma et al. 2009)