Lecture II – Nearest Neighbor and Kernel predictors

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527
Spring 2023

The Nearest-Neighbor predictor

- **1-Nearest Neighbor** The label of a point x is assigned as follows:
  1. find the example x^i that is nearest to x in $\mathcal{D}$ (in Euclidean distance)
  2. assign x the label y^i, i.e.

$$\hat{y}(x) = y^i$$

The Nearest-Neighbor predictor

- **1-Nearest Neighbor** The label of a point x is assigned as follows:
  1. find the example x^i that is nearest to x in $\mathcal{D}$ (in Euclidean distance)
  2. assign x the label y^i, i.e.

$$\hat{y}(x) = y^i$$

- **K-Nearest Neighbor** (with $K = 3, 5$ or larger)
  1. find the K nearest neighbors of x in $\mathcal{D}$: $x^{i_1, \ldots i_K}$
  2.
     - for classification $f(x) =$ the most frequent label among the K neighbors (well suited for multiclass)
     - for regression $f(x) = \frac{1}{K} \sum_{i \text{ neighbor of } x} y^i$ = mean of neighbors' labels

The Nearest-Neighbor predictor

- **1-Nearest Neighbor** The label of a point x is assigned as follows:
  1. find the example x^i that is nearest to x in $\mathcal{D}$ (in Euclidean distance)
  2. assign x the label y^i, i.e.
  $$\hat{y}(x) = y^i$$
- **K-Nearest Neighbor** (with $K = 3, 5$ or larger)
  1. find the K nearest neighbors of x in $\mathcal{D}$: $x^{i_1, \cdots i_K}$
  2. - for classification $f(x) =$ the most frequent label among the K neighbors (well suited for multiclass)
     - for regression $f(x) = \frac{1}{K} \sum_{i \text{ neighbor of } x} y^i =$ mean of neighbors' labels

- No parameters to estimate!
- No training!
- But all data must be stored (also called memory-based learning)

Kernel regression and classification

- Like the K-nearest neighbor but with "smoothed" neighborhoods
- The predictor

$$f(x) = \sum_{i=1}^{n} \beta_i b(x, x^i) y^i \tag{1}$$

where β_i are coefficients

Kernel regression and classification

- Like the K-nearest neighbor but with "smoothed" neighborhoods
- The predictor

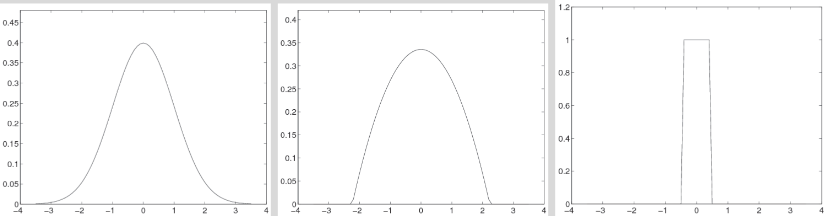
$$f(x) = \sum_{i=1}^{n} \beta_i b(x, x^i) y^i \tag{1}$$

  where β_i are coefficients
- Intuition: center a "bell-shaped" *kernel* function b on each data point, and obtain the prediction $f(x)$ as a weighted sum of the values y^i, where the weights are $\beta_i b(x, x^i)$
- Requirements for a kernel function $b(x, x')$
  1. non-negativity
  2. symmetry in the arguments x, x'
  3. optional: radial symmetry, bounded support, smoothness
- A typical kernel function is the **Gaussian kernel** (or **Radial Basis Function (RBF)**)

$$b(z) \propto e^{-z^2/2} \tag{2}$$

$$b_h(x, x') \propto e^{-\frac{||x-x'||^2}{2h^2}} \quad \text{with} \quad h = \textbf{the kernel width} \tag{3}$$

Kernels


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4 −3 −2 −1 0 1 2 3 4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4 −3 −2 −1 0 1 2 3 4
1.2
1
0.8
0.6
0.4
0.2
0
−4 −3 −2 −1 0 1 2 3 4

Regression example

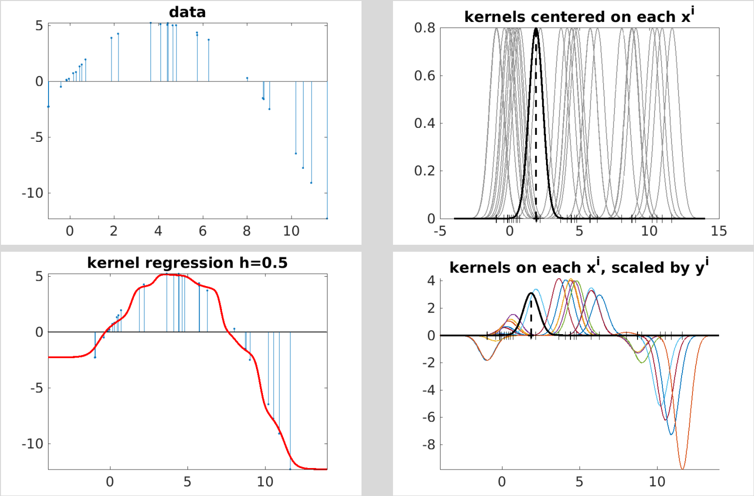
A special case in wide use is the Nadaraya-Watson regressor

$$f(x) \;=\; \frac{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right)}. \tag{4}$$

In this regressor, $f(x)$ is always a convex combination of the y^i's, and the weigths are proportional to $b_h(x, x^i)$.
The Nadaraya-Watson regressor is biased if the density of P_X varies around x.

STAT/BIOST 527 Spring 2023

An example: noisy data from a parabola



Local Linear Regression

To correct for the bias (to first order) one can estimate a regression line around x.

1. Given **query point** x
2. Compute kernel $b_h(x, x^i) = w_i$ for all $i = 1, \ldots N$
3. Solve weighted regression $\min_{\beta,\beta_0} \sum_{i=1}^{d} w_i \left(y^i - \beta^T x^i - \beta_0\right)^2$ to obtain β, β_0
   (β, β_0 depend on x through w_i)
4. Calculate $f(x) = \beta^T x + \beta_0$

Exercise Show that Nadaraya-Watson solves a local linear regression with fixed $\beta = 0$

Kernel binary classifiers

- Obtained from Nadaraya-Watson by setting y^i to ± 1.
- Note that the classifier can be written as the difference of two non-negative functions

$$f(x) \propto \sum_{i:y^i=1} b\left(\frac{||x-x^i||}{h}\right) - \sum_{i:y^i=-1} b\left(\frac{||x-x^i||}{h}\right). \tag{5}$$

Kernel regression by Nadaraya-Watson

$$\hat{y}(x) \;=\; \frac{\sum_{i=1}^n b\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^n b\left(\frac{||x-x^i||}{h}\right)} \tag{6}$$

Let $w_i = \frac{b\left(\frac{||x-x^i||}{h}\right)}{\sum_{i'=1}^n b\left(\frac{||x-x^{i'}||}{h}\right)}$.

Assumptions

A0 For simplicity, in this analysis we assume $x \in \mathbb{R}$.

A1 There is a true smooth[1] function $f(x)$ so that

$$y \;=\; f(x) + \varepsilon, \tag{7}$$

where ε is sampled independently for each x from a distribution P_ε, with $E_{P_\varepsilon}[\varepsilon] = 0$, $Var_{P_\varepsilon}(\varepsilon) = \sigma^2$.

A2 The kernel $b(z)$ is smooth, $\int_{\mathbb{R}} b(z)dz = 1$, $\int_{\mathbb{R}} zb(z) = 0$, and we denote $\sigma_b^2 = \int_{\mathbb{R}} z^2 b(z)dz$, $\gamma_b^2 = \int_{\mathbb{R}} b^2(z)dz$.

In this first analysis, we consider that the values x, $x^{1:N}$ are fixed; hence, the randomness is only in $\varepsilon^{1:N}$.

[1] with continuous derivatives up to order 2

Expectation of $\hat{y}(x)$ – a simple analysis

Expanding f in Taylor series around x we obtain

$$f(x^i) \;=\; f(x) + f'(x)(x^i - x) + \frac{f''(x)}{2}(x^i - x)^2 + o((x^i - x)^2) \tag{8}$$

We also have

$$y^i = f(x^i) + \varepsilon^i. \tag{9}$$

We now write the expectation of $\hat{y}(x)$ from (6), replacing in it y^i and $f(x^i)$ as above. What we would like to happen is that this expectation equals $f(x)$. Let us see if this is the case.

$$E_{P_\varepsilon^n}[\hat{y}(x)] \;=\; E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i y^i\right] \;=\; E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i\left(f(x^i) + \varepsilon^i\right)\right] \tag{10}$$

$$= \sum_{i=1}^n w_i f(x) + \sum_{i=1}^n w_i f'(x)(x^i - x) + \sum_{i=1}^n w_i \frac{f''(x)}{2}(x^i - x)^2 + \underbrace{E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \varepsilon^i\right]}_{=0]} \tag{11}$$

$$= \underbrace{f(x) + f'(x)\sum_{i=1}^n w_i(x^i - x) + \frac{f''(x)}{2}\sum_{i=1}^n w_i(x^i - x)^2}_{\text{bias}} \tag{12}$$

In the above, the expressions in red depend of f and x, those in blue depend on $x^{1:n}$.

Qualitative analysis of the bias terms

- The first order term $f'(x)\sum_{i=1}^{n} w_i(x^i - x)$ is responsible for **border effects**.
- The second order term **smooths out** sharp peaks and valleys.

Bias, Variance and h for $x \in \mathbb{R}$

[2]

The **bias** of $\hat{y}$ at x is defined as $E_{P_X^n} E_{P_\varepsilon^n}[\hat{y}(x) - f(x)]$.

$$E_{P_X^n} E_{P_\varepsilon^n}[\hat{y}(x) - f(x)] = h^2 \sigma_b^2 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right) + o(h^2) \tag{13}$$

The **variance** $\hat{y}$ at x is defined as $Var_{P_X^n P_\varepsilon^n}(\hat{y}(x))$.

$$Var_{P_X^n} P_\varepsilon^n(\hat{y}(x)) = \frac{\gamma^2}{nh} \sigma^2 + o\left(\frac{1}{nh}\right). \tag{14}$$

The **MSE (Mean Squared Error)** is defined as $E_{P_X^n} E_{P_\varepsilon^n}\left[(\hat{y}(x) - f(x))^2\right]$, which equals

$$MSE(x) = \textbf{bias}^2 + \textbf{variance} = h^4 \sigma_b^4 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right)^2 + \frac{\gamma_b^2}{nh} \sigma^2 + \ldots \tag{15}$$

[2] After []

Optimal selection of h

If the MSE is integrated over $\mathbb{R}$ we obtain the **MISE**$= \int_{\mathbb{R}} MSE(x) p_X(x) dx$.

$$MISE(h) \;=\; h^4 \left(\blacksquare \right) + \frac{(\blacksquare)}{nh} \;=\; 0. \tag{16}$$

The kernel width h can be chosen to minimize the MISE, for fixed f, p_X and b.
We set to 0 the partial derivative

$$\frac{\partial MISE}{\partial h} \;=\; h^3 \left(\blacksquare \right) - \frac{(\blacksquare)}{nh^2} \;=\; 0. \tag{17}$$

It follows that $h^5 \propto \frac{1}{n}$, or

$$h \;\propto\; \frac{1}{n^{1/5}}. \tag{18}$$

In d dimensions, the optimal h depends on the sample size n as

$$h \;\propto\; \frac{1}{n^{1/(d+4)}}. \tag{19}$$

The MISE with optimal h decreases as $\sim \frac{1}{n} n^{1/(d+4)} \;=\; \frac{1}{n^{1-1/(d+4)}}$
Compare this with the MSE of the mean of a distribution, which decreases $\sim \frac{1}{n}$