Academy of PhD Training in Statistics

Statistical Machine Learning

Louis J.M. Aslett (`louis.aslett@durham.ac.uk`)

Local Methods

Durham University

This Section

- Direct empirical estimation
- k-nearest neighbour
- Smoothing kernels
  - Kernel densities
  - Nadaraya-Watson estimator
  - Kernel density classification
  - Naïve Bayes

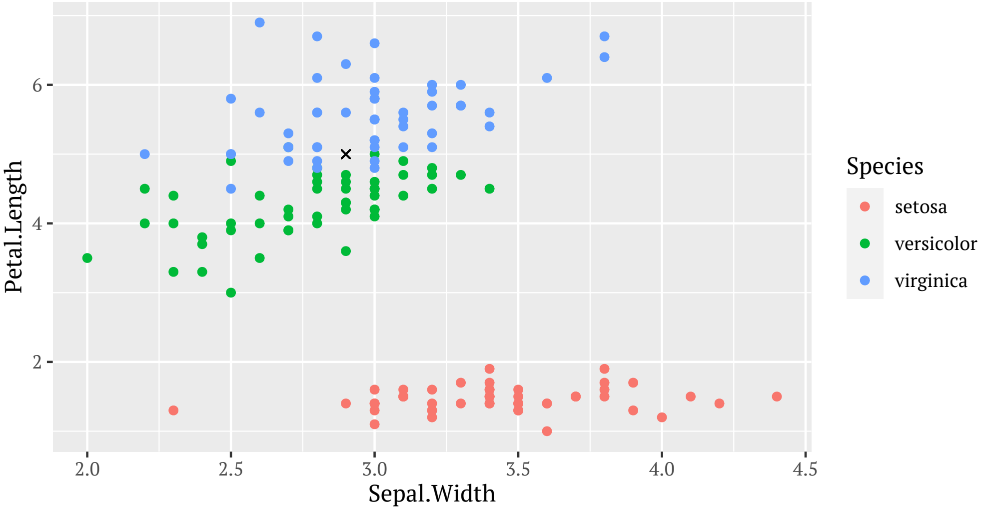
Direct empirical estimation

Recall, Bayes predictor:

$$\begin{aligned} g^\star(\mathbf{x}) &:= \arg\min_{z \in \mathcal{Y}} \mathbb{E}_{Y \mid X} \left[\mathcal{L}(Y, z) \mid X = \mathbf{x} \right] \\ &= \arg\min_{z \in \mathcal{Y}} \int_{\mathcal{Y}} \mathcal{L}(y, z) \, d\pi_{Y \mid X = \mathbf{x}} \end{aligned}$$

Could construct empirical estimate of the measure $\pi_{Y \mid X=\mathbf{x}}$, by looking at data "near" $\mathbf{x}$, so in a sense $\hat{\pi}_{Y \mid X \approx \mathbf{x}}$.

k-nearest neighbour



Formalising k-nearest neighbour

Given $\mathcal{D}_n = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n))$, reorder wrt new prediction value $\mathbf{x}$,

$$\Big((\mathbf{x}_{(1,\mathbf{x})}, y_{(1,\mathbf{x})}), \dots, (\mathbf{x}_{(n,\mathbf{x})}, y_{(n,\mathbf{x})})\Big)$$

where

$$d(\mathbf{x}_{(i,\mathbf{x})}, \mathbf{x}) \leq d(\mathbf{x}_{(j,\mathbf{x})}, \mathbf{x}) \quad \forall\, i < j$$

Formalising k-nearest neighbour

Given $\mathcal{D}_n = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n))$, reorder wrt new prediction value $\mathbf{x}$,

$$\left((\mathbf{x}_{(1,\mathbf{x})}, y_{(1,\mathbf{x})}), \dots, (\mathbf{x}_{(n,\mathbf{x})}, y_{(n,\mathbf{x})}) \right)$$

where

$$d(\mathbf{x}_{(i,\mathbf{x})}, \mathbf{x}) \le d(\mathbf{x}_{(j,\mathbf{x})}, \mathbf{x}) \quad \forall\, i < j$$

Then, for a general loss,

$$\begin{aligned} g^\star(\mathbf{x}) &= \arg\min_{z \in \mathcal{Y}} \int_{\mathcal{Y}} \mathcal{L}(y, z)\, d\pi_{Y \mid X = \mathbf{x}} \\ &\approx \arg\min_{z \in \mathcal{Y}} \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}\left(y_{(i,\mathbf{x})}, z \right) \end{aligned}$$

Formalising k-nearest neighbour: particular losses

$$g^\star(\mathbf{x}) \approx \begin{cases} \frac{1}{k}\sum_{i=1}^{k} y_{(i,\mathbf{x})} & \text{for squared loss} \\ \text{median}\{y_{(i,\mathbf{x})} : i \in \{1,\ldots,k\}\} & \text{for absolute loss (odd } k\text{)} \end{cases}$$

Formalising k-nearest neighbour: particular losses

$$g^\star(\mathbf{x}) \approx \begin{cases} \frac{1}{k}\sum_{i=1}^k y_{(i,\mathbf{x})} & \text{for squared loss} \\ \text{median}\{y_{(i,\mathbf{x})} : i \in \{1,\dots,k\}\} & \text{for absolute loss (odd } k\text{)} \end{cases}$$

0-1 loss,

$$g^\star(\mathbf{x}) \approx \begin{cases} 1 & \text{if } \sum_{i=1}^k y_{(i,\mathbf{x})} > \frac{k}{2} \\ 0 & \text{otherwise} \end{cases}$$

Formalising k-nearest neighbour: particular losses

$$g^\star(\mathbf{x}) \approx \begin{cases} \frac{1}{k}\sum_{i=1}^k y_{(i,\mathbf{x})} & \text{for squared loss} \\ \text{median}\{y_{(i,\mathbf{x})} : i \in \{1,\dots,k\}\} & \text{for absolute loss (odd } k\text{)} \end{cases}$$

0-1 loss,

$$g^\star(\mathbf{x}) \approx \begin{cases} 1 & \text{if } \sum_{i=1}^k y_{(i,\mathbf{x})} > \frac{k}{2} \\ 0 & \text{otherwise} \end{cases}$$

Or empirical probabilistic estimate,

$$\mathbb{P}(Y = j \mid X = \mathbf{x}) = \frac{1}{k}\sum_{i=1}^k \mathbb{1}\{y_{(i,\mathbf{x})} = j\}$$

Example

Example in notes

Bayesian k-nearest neighbour (Holmes and Adams, 2002) I

Can formulate a posterior on k and β (parameter for strength of effect between neighbours):

$$p(k, \beta \,|\, \mathcal{D} = (\mathbf{X}, \mathbf{y})) = \mathbb{P}(Y = \mathbf{y} \,|\, \mathbf{X}, k, \beta) p(k, \beta)$$

where

$$\mathbb{P}(Y = \mathbf{y} \,|\, \mathbf{X}, k, \beta) = \prod_{i=1}^{n} \frac{\exp\left(\frac{\beta}{k} \sum_{j=1}^{k} \mathbb{1}\{y_i = y_{(j,\mathbf{x}_i)}\}\right)}{\sum_{\ell=1}^{g} \exp\left(\frac{\beta}{k} \sum_{j=1}^{k} \mathbb{1}\{\ell = y_{(j,\mathbf{x}_i)}\}\right)}$$

Priors are independent, $p(k, \beta) = p(k)p(\beta)$, e.g. uniform k on $\{1, \ldots, n\}$ and improper uniform on $\beta \in \mathbb{R}^+$.

Bayesian k-nearest neighbour (Holmes and Adams, 2002) II

Posterior predictive for new observation $\mathbf{x}_{n+1}$,

$$\mathbb{P}(Y = y_{n+1} \,|\, \mathbf{x}_{n+1}, \mathbf{X}, \mathbf{y}) = \sum_{k=1}^{n} \int \mathbb{P}(Y = y_{n+1} \,|\, \mathbf{x}_{n+1}, \mathbf{X}, \mathbf{y}, k, \beta) p(k, \beta \,|\, \mathbf{X}, \mathbf{y}) \, d\beta$$

where

$$\mathbb{P}(Y = y_{n+1} \,|\, \mathbf{x}_{n+1}, \mathbf{X}, \mathbf{y}, k, \beta) = \frac{\exp\left(\frac{\beta}{k} \sum_{j=1}^{k} \mathbb{1}\{y_{n+1} = y_{(j,\mathbf{x}_{n+1})}\}\right)}{\sum_{\ell=1}^{g} \exp\left(\frac{\beta}{k} \sum_{j=1}^{k} \mathbb{1}\{\ell = y_{(j,\mathbf{x}_{n+1})}\}\right)}$$

Theoretical behaviour

- knn is asymptotically consistent (Stone, 1977)
    - as long as k is st $k/n \to 0$ as n increases

Theoretical behaviour

- knn is asymptotically consistent (Stone, 1977)
  - as long as k is st $k/n \to 0$ as n increases

- Convergence of expected error to Bayes error bounded above (regression) by $O\left(n^{-2\alpha/(d+2\alpha)}\right)$ term (Kohler et al., 2006)
  - if f bounded, α-Hölder continuous plus finite moment conditions on X

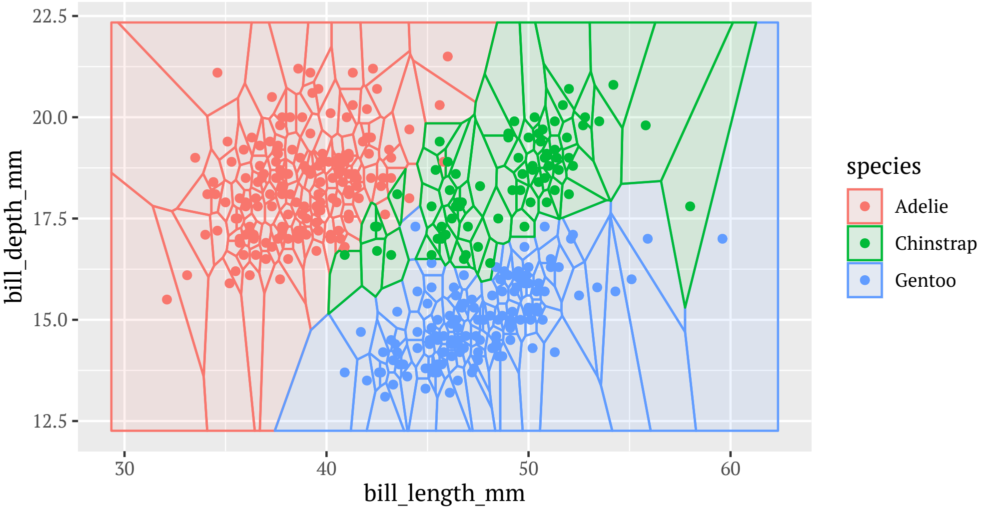
Theoretical behaviour

- knn is asymptotically consistent (Stone, 1977)
  - as long as k is st $k/n \to 0$ as n increases

- Convergence of expected error to Bayes error bounded above (regression) by $O\left(n^{-2\alpha/(d+2\alpha)}\right)$ term (Kohler et al., 2006)
  - if f bounded, α-Hölder continuous plus finite moment conditions on X

- For classification, no $k > 1$ has lower error against all possible distributions than $k = 1$ (Cover and Hart, 1967)
  - practically not a helpful result as usually more worried about striking a bias-variance tradeoff
  - $k = 1$ low bias, high variance;
  - $k \gg 1$ higher bias, lower variance.

$k = 1 \implies$ Voronoi diagrams



Computational considerations

Naïvely, computational cost $\propto nkd$ and memory cost $\propto nd$

Computational considerations

Naïvely, computational cost $\propto nkd$ and memory cost $\propto nd$

- Memory (García et al., 2012)
  - Condensing (Hart, 1968 etc): keep points near decision boundary, eliminate redundancy. e.g. interior points on Voronoi diagram
  - Editing (Wilson, 1972 etc): remove boundary points in noisy region to keep core representatives

Computational considerations

Naïvely, computational cost $\propto nkd$ and memory cost $\propto nd$

- Memory (García et al., 2012)
  - Condensing (Hart, 1968 etc): keep points near decision boundary, eliminate redundancy. e.g. interior points on Voronoi diagram
  - Editing (Wilson, 1972 etc): remove boundary points in noisy region to keep core representatives

- Compute
  - k-d tree algorithms (Bentley, 1975): partition space into tree structure to guide search for nearest neighbours (then $\propto \log n$)
  - Approximate nearest neighbour, eg via random projections (Andoni et al., 2018)

Curse of dimensionality (I)

e.g. in convergence rate, $O\left(n^{-2\alpha/(d+2\alpha)}\right)$

Curse of dimensionality (I)

e.g. in convergence rate, $O\left(n^{-2\alpha/(d+2\alpha)}\right)$

knn assumes there is other data 'nearby', but as dimension grows, neighbourhoods grow very fast.

Let π_X be uniform on $\mathcal{X} = [0,1]^d$ and consider $k = 5$. We measure 'local' by size of hypercube that contains an observation and it's $k = 5$ nearest neighbours.
Example in notes

Curse of dimensionality (I)

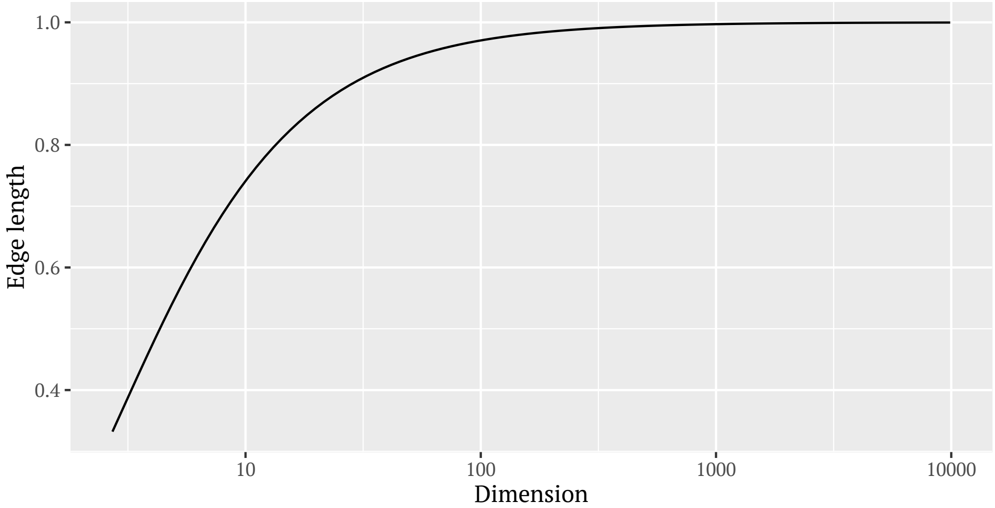
e.g. in convergence rate, $O\left(n^{-2\alpha/(d+2\alpha)}\right)$

knn assumes there is other data 'nearby', but as dimension grows, neighbourhoods grow very fast.

Let π_X be uniform on $\mathcal{X} = [0,1]^d$ and consider $k = 5$. We measure 'local' by size of hypercube that contains an observation and it's $k = 5$ nearest neighbours.
Example in notes

Uniform $\implies$ Volume $\ell^d \approx \frac{k}{n} \implies \ell \approx \left(\frac{k}{n}\right)^{\frac{1}{d}}$ … not so local!

Curse of dimensionality (II)



Distances

- ℓ_1-norm (Manhattan), ℓ_2-norm (Euclidean) or Mahalanobis distance common for numeric predictors.

Distances

- ℓ_1-norm (Manhattan), ℓ_2-norm (Euclidean) or Mahalanobis distance common for numeric predictors.
- Gower's distance (Gower, 1971). $d(\mathbf{x}_i, \mathbf{x}_j)$: take each variable, $\ell \in \{1, \dots, d\}$
  - Numeric: $\delta_\ell = \frac{|x_{i\ell} - x_{j\ell}|}{R_\ell}$ where $R_\ell = \max_i\{x_{i\ell}\} - \min_i\{x_{i\ell}\}$
  - Categorical: $\delta_\ell = \begin{cases} 1 & \text{if } x_{i\ell} = x_{j\ell} \\ 0 & \text{otherwise} \end{cases}$
  - Total distance: $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{d}\sum_{\ell=1}^{d} \delta_\ell$

Distances

- ℓ_1-norm (Manhattan), ℓ_2-norm (Euclidean) or Mahalanobis distance common for numeric predictors.
- Gower's distance (Gower, 1971). $d(\mathbf{x}_i, \mathbf{x}_j)$: take each variable, $\ell \in \{1, \dots, d\}$
  - Numeric: $\delta_\ell = \frac{|x_{i\ell} - x_{j\ell}|}{R_\ell}$ where $R_\ell = \max_i\{x_{i\ell}\} - \min_i\{x_{i\ell}\}$
  - Categorical: $\delta_\ell = \begin{cases} 1 & \text{if } x_{i\ell} = x_{j\ell} \\ 0 & \text{otherwise} \end{cases}$
  - Total distance: $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{d} \sum_{\ell=1}^{d} \delta_\ell$
- Many other special distances customised for particular scenarios, eg tangent data for images etc. (Hastie et al., 2009, eg §13.3.3)

Scaling

Important final word: k-nearest neighbours is not invariant to the scale of individual variables!

$\therefore$ **do** centre and scale your data

Easy mistake: *be careful!* to ensure you apply the mean and standard deviation from the training data to scale new observations (ie don't scale your test observations independently)

Example

Hays and Efros (2008) application example

Smoothing kernels

knn defined 'local' by looking for k near observations, irrespective of their distance.

Smoothing kernels

knn defined ‘local’ by looking for k near observations, irrespective of their distance.

Dual idea? Fix the distance we look, irrespective of the number of observations within that radius.

Smoothing kernels

knn defined 'local' by looking for k near observations, irrespective of their distance.

Dual idea? Fix the distance we look, irrespective of the number of observations within that radius.

$$\hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i \mathbb{1}\{d(\mathbf{x}_i, \mathbf{x}) < h\}}{\sum_{i=1}^n \mathbb{1}\{d(\mathbf{x}_i, \mathbf{x}) < h\}}$$

Now, rather than choose $k \in \mathbb{N}$, we choose $h \in \mathbb{R}^+$, called the *bandwidth*.

Kernel density estimation: justification (I)

Tackle third classical ML problem first, and use this to solve the other two: ie construct density estimator for π_X.

Kernel density estimation: justification (I)

Tackle third classical ML problem first, and use this to solve the other two: ie construct density estimator for π_X.

Simplest density estimator is a histogram. Could we instead place bins at the observations and accumulate? Yes!

$$\begin{aligned} f_X(x) &= \frac{\partial F_X}{\partial x}(x) \\ &= \lim_{h \to 0} \frac{F_X(x+h) - F_X(x)}{h} \\ &\equiv \lim_{h \to 0} \frac{F_X(x) - F_X(x-h)}{h} \end{aligned}$$

Kernel density estimation: justification (II)

$$f_X(x) = \lim_{h\to 0} \frac{F_X(x+h) - F_X(x)}{h} \equiv \lim_{h\to 0} \frac{F_X(x) - F_X(x-h)}{h}$$

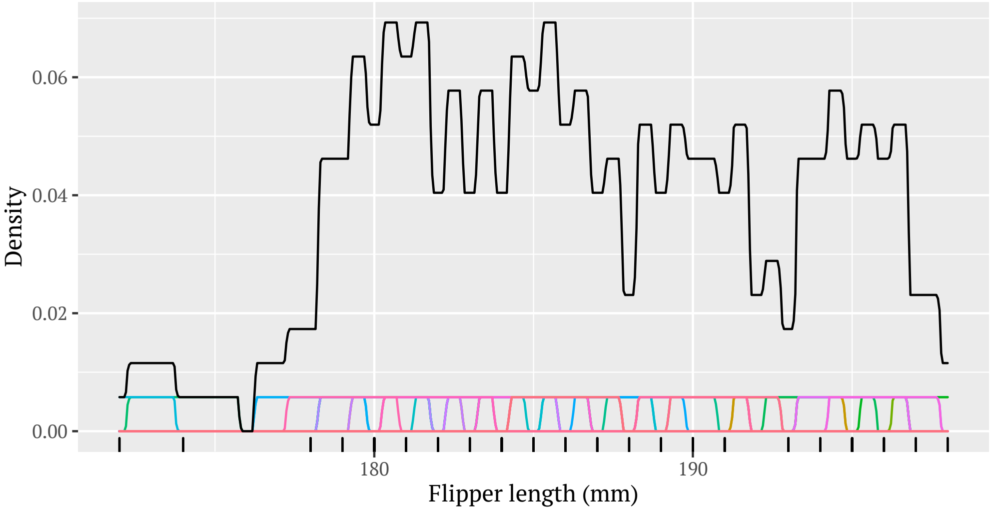
$$\implies \frac{f_X(x) + f_X(x)}{2} = \lim_{h\to 0} \frac{F_X(x+h) - \cancel{F_X(x)} + \cancel{F_X(x)} - F_X(x-h)}{2h}$$

$$f_X(x) = \lim_{h\to 0} \frac{\mathbb{P}(x - h < X < x + h)}{2h}$$

$$\approx \frac{1}{2h}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x - h < x_i < x + h\}\right) \text{ for small } h > 0$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{1}{2h}\mathbb{1}\left\{\left|\frac{x - x_i}{h}\right| < 1\right\}}_{\star \text{ valid uniform pdf}}$$

Kernel density estimation: justification (III)



Kernel function

Replace uniform density by smoother alternative, a *kernel function*.

A *kernel function* is any function K such that,

1. $K : \mathcal{X} \to [0, \infty)$
2. $K(\cdot)$ is a valid probability density (integrating to 1),

$$\int_{\mathcal{X}} K(d\pi_X) = 1$$

3. $K(\cdot)$ is symmetric, $K(\mathbf{x}) = K(-\mathbf{x}) \, \forall \, \mathbf{x} \in \mathcal{X}$.

Some authors only require (i) or (i) and (ii), whilst others add explicit moment conditions on $K(\cdot)$ (ie second moment =1, all moments finite).

The kernel density estimator (KDE)

The *kernel density estimator* of the density $f_X(\cdot)$ based on n iid observations $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ drawn from π_X is given by:

$$\hat{f}_X(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\mathbf{z} - \mathbf{x}_i}{h}\right)$$

Where $K(\cdot)$ is a valid kernel function. NB, $\hat{f}_X(\mathbf{z})$ a valid pdf.

The kernel density estimator (KDE)

The *kernel density estimator* of the density $f_X(\cdot)$ based on n iid observations $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ drawn from π_X is given by:

$$\hat{f}_X(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\mathbf{z} - \mathbf{x}_i}{h}\right)$$

Where $K(\cdot)$ is a valid kernel function. NB, $\hat{f}_X(\mathbf{z})$ a valid pdf.

Practically: product of univariate kernel functions

$$\hat{f}_X(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{\ell=1}^{d} \frac{1}{h_\ell} K\left(\frac{z_\ell - x_{i\ell}}{h_\ell}\right)$$

where $K : \mathbb{R} \to [0, \infty)$, possibly different bandwidths per dimension, h_ℓ.

Common univariate kernel functions

- Епанечников (1969) (translation Epanechnikov (1969))

$$K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}\{|x| \le 1\}$$

original scaled by factor $\sqrt{5}$ for unit second moment,

$$K(x) = \left(\frac{3}{4\sqrt{5}} - \frac{3x^2}{20\sqrt{5}} \right) \mathbb{1}\{|x| \le \sqrt{5}\}$$

- Gaussian

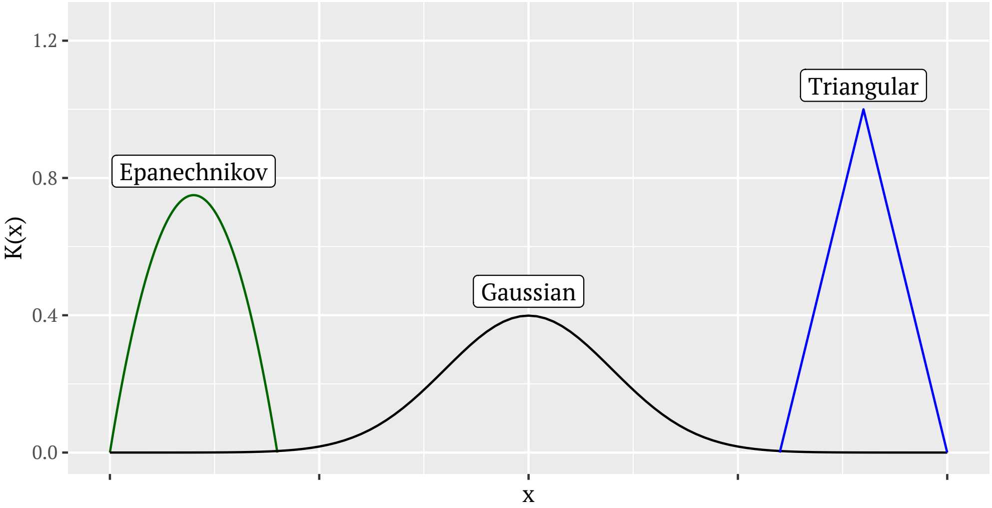
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{x^2}{2} \right)$$

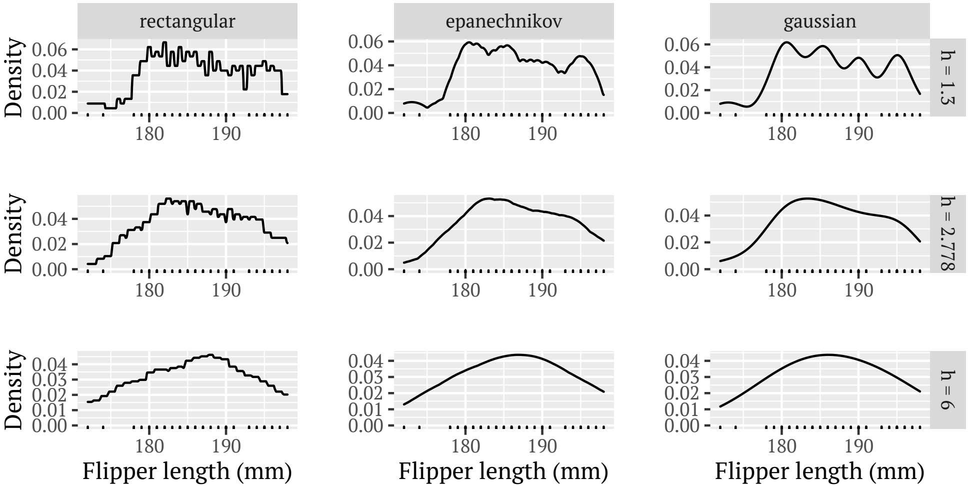
- Triangular

$$K(x) = (1 - |x|)\mathbb{1}\{|x| \le 1\}$$

Practically speaking, any choices does not have huge impact on result.

Visualising univariate kernel functions




rectangular
epanechnikov
gaussian
h = 1.3
h = 2.778
h = 6
Density
0.00
0.01
0.02
0.03
0.04
0.06
180
190
Flipper length (mm)

Errors in KDE

KDE is about density estimation in the general setting, so uses the following error metrics rather than the ML metrics presented before:

$$\mathbb{E}_{\pi_X^n}\left[\left(f_X(\mathbf{x}) - \hat{f}_X(\mathbf{x})\right)^2\right] \qquad \text{mean square error (MSE)}$$

$$\int_{\mathcal{X}} \mathbb{E}_{\pi_X^n}\left[\left(f_X(\mathbf{x}) - \hat{f}_X(\mathbf{x})\right)^2\right] d\mathbf{x} \qquad \text{mean integrated square error (MISE)}$$

Note

- MSE is at fixed $\mathbf{x}$;
- MISE is *not* integrated wrt π_X

KDE bias/variance

$$\text{Bias}[\hat{f}_X(x)] \approx \frac{h^2}{2} f''_X(x) \int z^2 K(z)\, dz + o(h)^2 = O\left(h^2\right)$$

$$\text{Var}[\hat{f}_X(x)] \approx \frac{f_X(x)}{nh} \int K^2(z)\, dz = O\left((nh)^{-1}\right)$$

- $h \gg \implies$ higher bias, lower variance
- $h \ll \implies$ lower bias, higher variance

KDE bias/variance

$$\mathrm{Bias}[\hat{f}_X(x)] \approx \frac{h^2}{2} f_X''(x) \int z^2 K(z)\, dz + o(h)^2 = O\left(h^2\right)$$

$$\mathrm{Var}[\hat{f}_X(x)] \approx \frac{f_X(x)}{nh} \int K^2(z)\, dz = O\left((nh)^{-1}\right)$$

- $h \gg \implies$ higher bias, lower variance
- $h \ll \implies$ lower bias, higher variance

Leads to asymptotic mean integrated square error,

$$\text{AMISE} = \frac{1}{nh} \int K^2(z)\, dz + h^4 \left(\int f_X''(z)^2\, dz \right) \left(\int z^2 K(z)\, dz \right)^2$$

Optimal / practical bandwidth selection

$$h = \left(\frac{\int K^2(z)\,dz}{n \left(\int f_X''(z)^2\,dz\right) \left(\int z^2 K(z)\,dz\right)^2} \right)^{\frac{1}{5}}$$

Optimal / practical bandwidth selection

$$h = \left(\frac{\int K^2(z)\,dz}{n \left(\int f''_X(z)^2\,dz\right) \left(\int z^2 K(z)\,dz\right)^2} \right)^{\frac{1}{5}}$$

- Rules of thumb: estimate $\int f''_X(z)^2\,dz$ by substituting Normal density for f.
  - default method used by `density()` function in R.

Optimal / practical bandwidth selection

$$h = \left(\frac{\int K^2(z)\, dz}{n \left(\int f_X''(z)^2\, dz\right) \left(\int z^2 K(z)\, dz\right)^2} \right)^{\frac{1}{5}}$$

- Rules of thumb: estimate $\int f_X''(z)^2\, dz$ by substituting Normal density for f.
    - default method used by `density()` function in R.
- Cross validation (see next lecture): either unbiased (estimate square error directly) or biased (estimate the AMISE).
    - `density(..., method = "ucv")` or `density(..., method = "bcv")`

Optimal / practical bandwidth selection

$$h = \left(\frac{\int K^2(z)\,dz}{n\left(\int f_X''(z)^2\,dz\right)\left(\int z^2 K(z)\,dz\right)^2} \right)^{\frac{1}{5}}$$

- Rules of thumb: estimate $\int f_X''(z)^2\,dz$ by substituting Normal density for f.
  - default method used by `density()` function in R.
- Cross validation (see next lecture): either unbiased (estimate square error directly) or biased (estimate the AMISE).
  - `density(..., method = "ucv")` or `density(..., method = "bcv")`
- Plug-in: make pilot estimate of derivative and seek fixed-point solution of above.
  - `density(..., method = "SJ")`

Jones et al. (1996) recommend plug-in approach (`method = "SJ"`)

Pivoting to regression

$$\mathbb{E}\left[Y \mid X = \mathbf{x}\right]$$

Pivoting to regression

$$\mathbb{E}\left[Y \mid X = \mathbf{x}\right] = \int_{\mathcal{Y}} y \, f_{Y \mid X}(y \mid \mathbf{x}) \, dy$$

Pivoting to regression

$$\begin{aligned}\mathbb{E}\left[Y \mid X = \mathbf{x}\right] &= \int_{\mathcal{Y}} y f_{Y \mid X}(y \mid \mathbf{x}) \, dy \\ &= \int_{\mathcal{Y}} y \frac{f_{XY}(\mathbf{x}, y)}{f_X(\mathbf{x})} \, dy\end{aligned}$$

Pivoting to regression

$$
\begin{aligned}
\mathbb{E}\left[Y \mid X = \mathbf{x}\right] &= \int_{\mathcal{Y}} y f_{Y \mid X}(y \mid \mathbf{x}) \, dy \\
&= \int_{\mathcal{Y}} y \frac{f_{XY}(\mathbf{x}, y)}{f_X(\mathbf{x})} \, dy \\
&\approx \int_{\mathcal{Y}} y \frac{\cancel{\frac{1}{n}} \sum_{i=1}^n \frac{1}{h^{d+1}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) K\left(\frac{y-\mathbf{y}_i}{h}\right)}{\cancel{\frac{1}{n}} \sum_{i=1}^n \frac{1}{h^d} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} \, dy
\end{aligned}
$$

Pivoting to regression

$$\begin{aligned}
\mathbb{E}\left[Y \mid X = \mathbf{x}\right] &= \int_{\mathcal{Y}} y f_{Y \mid X}(y \mid \mathbf{x}) \, dy \\
&= \int_{\mathcal{Y}} y \frac{f_{XY}(\mathbf{x}, y)}{f_X(\mathbf{x})} \, dy \\
&\approx \int_{\mathcal{Y}} y \frac{\cancel{\frac{1}{n}} \sum_{i=1}^n \frac{1}{h^{d+1}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) K\left(\frac{y-\mathbf{y}_i}{h}\right)}{\cancel{\frac{1}{n}} \sum_{i=1}^n \frac{1}{h^d} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} \, dy \\
&= \frac{\sum_{i=1}^n \cancel{\frac{1}{h^d}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \int_{\mathcal{Y}} \frac{y}{h} K\left(\frac{y-\mathbf{y}_i}{h}\right) \, dy}{\sum_{i=1}^n \cancel{\frac{1}{h^d}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}
\end{aligned}$$

Pivoting to regression

$$
\begin{aligned}
\mathbb{E}\left[Y \mid X = \mathbf{x}\right] &= \int_{\mathcal{Y}} y f_{Y \mid X}(y \mid \mathbf{x}) \, dy \\
&= \int_{\mathcal{Y}} y \frac{f_{XY}(\mathbf{x}, y)}{f_X(\mathbf{x})} \, dy \\
&\approx \int_{\mathcal{Y}} y \frac{\cancel{\frac{1}{n}} \sum_{i=1}^n \frac{1}{h^{d+1}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) K\left(\frac{y-\mathbf{y}_i}{h}\right)}{\cancel{\frac{1}{n}} \sum_{i=1}^n \frac{1}{h^d} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} \, dy \\
&= \frac{\sum_{i=1}^n \cancel{\frac{1}{h^d}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \int_{\mathcal{Y}} \frac{y}{h} K\left(\frac{y-\mathbf{y}_i}{h}\right) \, dy}{\sum_{i=1}^n \cancel{\frac{1}{h^d}} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} \\
&= \frac{\sum_{i=1}^n y_i K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}
\end{aligned}
$$

Nadaraya-Watson estimator

The kernel based *Nadaraya-Watson estimator* of a regression function based on training data $\mathcal{D}_n = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n))$ is:

$$\hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}$$

More common to use cross-validation to select h here (see next lecture).

Naïve Bayes classifier

$$\mathbb{P}(Y = i \,|\, X = \mathbf{x}) = \frac{f_{X\,|\,Y}(\mathbf{x} \,|\, Y = i)\mathbb{P}(Y = i)}{\sum_{j=1}^{g} f_{X\,|\,Y}(\mathbf{x} \,|\, Y = j)\mathbb{P}(Y = j)}$$

Use KDE for $f_{X\,|\,Y}(\mathbf{x} \,|\, Y = i)$ and empirical estimate for $\mathbb{P}(Y = i)$? Dimensionality ...

Naïve Bayes classifier

$$\mathbb{P}(Y = i \,|\, X = \mathbf{x}) = \frac{f_{X\,|\,Y}(\mathbf{x} \,|\, Y = i)\mathbb{P}(Y = i)}{\sum_{j=1}^{g} f_{X\,|\,Y}(\mathbf{x} \,|\, Y = j)\mathbb{P}(Y = j)}$$

Use KDE for $f_{X\,|\,Y}(\mathbf{x} \,|\, Y = i)$ and empirical estimate for $\mathbb{P}(Y = i)$? Dimensionality …

Naïve Bayes classifier assumes conditional independence of all dimensions in $\mathcal{X}$. That is,

$$\mathbb{P}(Y = i \,|\, X = \mathbf{x}) = \frac{\mathbb{P}(Y = i) \prod_{k=1}^{d} f_{X_k\,|\,Y}(x_k \,|\, Y = i)}{\sum_{j=1}^{g} \mathbb{P}(Y = i) \prod_{k=1}^{d} f_{X_k\,|\,Y}(x_k \,|\, Y = i)}$$

and construct KDE of univariate marginal densities $f_{X_k\,|\,Y}(x_k \,|\, Y = i) \,\forall\, k$.

`naivebayes::naive_bayes(..., kernel = TRUE)` (Majka, 2019) to fit in R.

References I

Andoni, A., Indyk, P., Razenshteyn, I. (2018). Approximate nearest neighbor search in high dimensions, in: Proceedings of the International Congress of Mathematicians (ICM 2018). pp. 3287–3318. DOI: 10.1142/9789813272880_0182

Bentley, J.L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM* **18**(9), 509–517. DOI: 10.1145/361002.361007

Cover, T.M., Hart, P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**(1), 21–27. DOI: 10.1109/TIT.1967.1053964

Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* **14**(1), 153–158. DOI: 10.1137/1114019

References II

García, S., Derrac, J., Cano, J.R., Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(3), 417–435. DOI: 10.1109/TPAMI.2011.142

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**(4), 857–871. DOI: 10.2307/2528823

Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory* **14**(3), 515–516. DOI: 10.1109/TIT.1968.1054155

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning, 2nd ed, Springer Series in Statistics. Springer. URL `https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf`, ISBN: 978-0387848570

References III

Hays, J., Efros, A.A. (2008). IM2GPS: estimating geographic information from a single image, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. DOI: 10.1109/CVPR.2008.4587784

Holmes, C.C., Adams, N.M. (2002). A probabilistic nearest neighbour method for statistical pattern recognition. *Journal of the Royal Statistical Society: Series B* **64**, 295–306. DOI: 10.1111/1467-9868.00338

Jones, M.C., Marron, J.S., Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**(433), 401–407. DOI: 10.2307/2291420

Kohler, M., Krzyżak, A., Walk, H. (2006). Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis* **97**(2), 311–323. DOI: 10.1016/j.jmva.2005.03.006

References IV

Majka, M. (2019). naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R. *R package version 0.9.7*. URL `https://CRAN.R-project.org/package=naivebayes`

Stone, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics* **5**(4), 595–645. DOI: 10.1214/aos/1176343886

Wilson, D.L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-2**(3), 408–421. DOI: 10.1109/TSMC.1972.4309137

Епанечников, В.А. (1969). Непараметрическая оценка многомерной плотности вероятности. *Теория вероятн. и ее примен.* **14**(1), 156–161. URL `http://mi.mathnet.ru/tvp1130`