Kernel Regression and KDE

Steve Avsec

Illinois Institute of Technology

February 5, 2024

Overview

Kernels

A *kernel* is a function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that

$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$$

$$\sum_{j=1}^{n} \sum_{k=1}^{n} K(\mathbf{x}_j, \mathbf{x}_k) c_j c_k \geq 0 \text{ for all } (\mathbf{x})_{j=1}^{n} \text{ and } (c_j)_{j=1}^{n} \, .$$

Examples

Some examples

1. $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$.
2. $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2}}$
3. $K(\mathbf{x}, \mathbf{y}) = D(\|\mathbf{x} - \mathbf{y}\|)$

where

$$D(t) = \begin{cases} \left(1 - |t|^3\right)^3 & \text{if } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(This is called the Epanechnikov kernel.)

A Theorem

Theorem

For all kernels $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$*, there exists a Hilbert space* $\mathcal{H}$ *and a function* $\psi : \mathbb{R}^n \to \mathcal{H}$ *such that*

$$K(\mathbf{x}, \mathbf{y}) = \langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle$$

A Theorem

Theorem

For all kernels $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, there exists a Hilbert space $\mathcal{H}$ and a function $\psi : \mathbb{R}^n \to \mathcal{H}$ such that

$$K(\mathbf{x}, \mathbf{y}) = \langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle$$

A *Hilbert space* is a generalization of Euclidean space. It's a vector space together with an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$.

A Theorem

Theorem

For all kernels $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, there exists a Hilbert space $\mathcal{H}$ and a function $\psi : \mathbb{R}^n \to \mathcal{H}$ such that

$$K(\mathbf{x}, \mathbf{y}) = \langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle$$

A *Hilbert space* is a generalization of Euclidean space. It's a vector space together with an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$.

Example: ℓ_2, the vector space of square-summable sequences $(a_j)_{j=0}^{\infty}$ such that $\lim_N \sum_{j=0}^{N} |a_j|^2$ exists with inner product

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{j=0}^{\infty} \overline{a_j} b_j$$

Non-trivial Example

Let $\mathcal{H}$ be $L^2(d\gamma)$, the square integrable functions against the normal weight:

$$\langle f, g \rangle = \int_{-\infty}^{\infty} \overline{f(t)} g(t) e^{-\frac{t^2}{2}} dt$$

Non-trivial Example

Let $\mathcal{H}$ be $L^2(d\gamma)$, the square integrable functions against the normal weight:

$$\langle f, g \rangle = \int_{-\infty}^{\infty} \overline{f(t)} g(t) e^{-\frac{t^2}{2}} dt$$

Define $\psi : \mathbb{R} \to \mathcal{H}$ by

$$\psi(t) = e^{it}$$

Non-trivial Example

Let $\mathcal{H}$ be $L^2(d\gamma)$, the square integrable functions against the normal weight:

$$\langle f, g\rangle = \int_{-\infty}^{\infty} \overline{f(t)} g(t) e^{-\frac{t^2}{2}} dt$$

Define $\psi : \mathbb{R} \to \mathcal{H}$ by

$$\psi(t) = e^{it}$$

Through some miraculous calculations

$$\langle \psi(s), \psi(t)\rangle = \int_{-\infty}^{\infty} e^{-is} e^{it} e^{-\frac{u}{2}} du = e^{-\frac{|t-s|^2}{2}}$$

Bases

We can look at a minimization problem like:

$$\min_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w}) = \|\mathbf{y} - \langle \mathbf{w}, \sum_{j=1}^{N} \psi(\mathbf{x}_j) \rangle\|_2$$

Bases

We can look at a minimization problem like:

$$\min_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w}) = \|\mathbf{y} - \langle \mathbf{w}, \sum_{j=1}^{N} \psi(\mathbf{x}_j)\rangle\|_2$$

Lemma

The minimizer $\hat{\mathbf{w}}$ *is necessarily in span*$\{\psi(\mathbf{x}_j)\}$.

Kernel Regression

So we can assume that $\hat{\mathbf{w}} = \sum_{j=1}^{N} c_j \psi(\mathbf{x}_j)$.

Kernel Regression

So we can assume that $\hat{\mathbf{w}} = \sum_{j=1}^{N} c_j \psi(\mathbf{x}_j)$.

And our minimizer takes the form:

$$L(\mathbf{w}) = \left\| \mathbf{y} - \sum_{j,k=1}^{n} c_k \langle \psi(\mathbf{x}_k), \psi(\mathbf{x}_j) \rangle \right\|_2$$

Conclusion

This reduces to a problem we know:

$$L(\mathbf{w}) = \|\mathbf{y} - K\mathbf{c}\|_2$$

where $K = [K(\mathbf{x}_k, \mathbf{x}_j)]$ and $\mathbf{c} = (c_1, \ldots, c_N)^t$.
Our previous example then becomes:

$$L(\mathbf{c}) = \left\| \mathbf{y} - \sum_{k=1}^{N} \sum_{j=0}^{M} c_0 + c_j e^{\frac{\|\mathbf{x}_k - \xi_j\|^2}{\lambda_j}} \right\|_2$$

(This goes by the name Radial Basis Network. We could include a positive semidefinite matrix and this gets us close to an algorithm called basis pursuit that we will discuss later.)

Big Idea

Suppose we have some data $\{\mathbf{x}_j\}_{j=1}^{N} \in \mathbb{R}^d$. Assume this data is pulled from some (unknown) probability density function p.

Big Idea

Suppose we have some data $\{\mathbf{x}_j\}_{j=1}^N \in \mathbb{R}^d$. Assume this data is pulled from some (unknown) probability density function p.

Goal: Find a reasonable estimate of p.

Big Idea

Suppose we have some data $\{\mathbf{x}_j\}_{j=1}^N \in \mathbb{R}^d$. Assume this data is pulled from some (unknown) probability density function p.

Goal: Find a reasonable estimate of p.

(This is our first unsupervised technique!)

A First Guess

We could just put little circles around each data point and sum those up:

$$\hat{p}(\mathbf{x}) = \frac{C}{N\lambda^d} \sum_{j=1}^{N} \chi_{B_\lambda(\mathbf{x}_j)}(\mathbf{x})$$

A First Guess

We could just put little circles around each data point and sum those up:

$$\hat{p}(\mathbf{x}) = \frac{C}{N\lambda^d} \sum_{j=1}^{N} \chi_{B_\lambda(\mathbf{x}_j)}(\mathbf{x})$$

This works, but has some drawbacks (like that new points more than λ away from old points have 0 probability).

A Better Idea

Choose a suitable ”base” density function $k : \mathbb{R}^d \to \mathbb{R}$.

A Better Idea

Choose a suitable "base" density function $k : \mathbb{R}^d \to \mathbb{R}$.

Define

$$k_h(\mathbf{x}) = \frac{1}{h} k\left(\frac{\mathbf{x}}{h}\right).$$

A Better Idea

Choose a suitable ”base” density function $k : \mathbb{R}^d \rightarrow \mathbb{R}$.

Define

$$k_h(\mathbf{x}) = \frac{1}{h} k(\frac{\mathbf{x}}{h}).$$

Then define

$$\hat{p}(\mathbf{x}) = \frac{1}{hN} \sum_{j=1}^{N} k_h(\mathbf{x} - \mathbf{x}_j)$$

This is called the Parzen-Rosenblatt window method.

Bayes' Rule

Bayes' Rule states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Rule

Bayes' Rule states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example: Suppose the occurrence of a rare condition is 10^{-5}. Suppose that there is a test for this condition which has false positive and false negative rates of 0.01 each.

Bayes' Rule

Bayes' Rule states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example: Suppose the occurrence of a rare condition is 10^{-5}. Suppose that there is a test for this condition which has false positive and false negative rates of 0.01 each.

Let A be the event that the patient has the condition and B be the occurrence of a positive test.

Example Continued

$$\begin{aligned} P(A) &= 10^{-5} \\ P(B|A) &= 0.99 \\ P(B) &= P(B|A)P(A) + P(B|A^c)P(A^c) \\ &= 0.99 * 10^{-5} + 0.01 * (1 - 10^{-5}) \end{aligned}$$

Example Continued

$$
\begin{aligned}
P(A) &= 10^{-5} \\
P(B|A) &= 0.99 \\
P(B) &= P(B|A)P(A) + P(B|A^c)P(A^c) \\
&= 0.99 * 10^{-5} + 0.01 * (1 - 10^{-5})
\end{aligned}
$$

$$
P(A|B) = \frac{0.99 * 10^{-5}}{0.99 * 10^{-5} + 0.01 * (1 - 10^{-5})}
$$

Key Assumption

Data $\{\mathbf{x}_j\} \in \mathbb{R}^d$ and

$$p(\mathbf{x}) = \prod_{k=1}^{d} p_k(x_k)$$