The Role of a Kernel in Statistical Learning

Dr. Jimmy Risk
Cal Poly Pomona

4/6/21

What Is A Kernel?

Statistics and Probability:

1. The kernel of a pdf (or pmf)
2. Kernel Density Estimation
3. Support Vector Machines
4. Kernel Ridge Regression
5. Kernel PCA
6. Covariance kernels in Gaussian processes

Mathematics:

1. Kernel of a linear map (aka null space)
2. Integral transform T

$$(Tf)(u) = \int_{t_1}^{t_2} f(t)K(t,u)dt,$$

   where $K(t,u)$ is a **kernel**

   - e.g. Fourier transform: $K(t,u) = e^{-2\pi i u t}$

3. Reproducing Kernel Hilbert Spaces (RKHS)

What Is A Kernel?

Statistics and Probability:

1. The kernel of a pdf (or pmf)
2. Kernel Density Estimation
3. Support Vector Machines
4. Kernel Ridge Regression
5. Kernel PCA
6. Covariance kernels in Gaussian processes

Mathematics:

1. Kernel of a linear map (aka null space)
2. Integral transform T

$$(Tf)(u) = \int_{t_1}^{t_2} f(t)K(t,u)dt,$$

   where $K(t,u)$ is a **kernel**
   - e.g. Fourier transform: $K(t,u) = e^{-2\pi iut}$
3. Reproducing Kernel Hilbert Spaces (RKHS)

Reproducing Kernel Hilbert Space

Definition (Reproducing Kernel Hilbert Space)

[a] Let $\mathcal{H}$ be a Hilbert space of real functions f defined on $\mathcal{X}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with an inner product $\langle \mathcal{X}, \mathcal{X} \rangle_{\mathcal{H}}$ (and norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:

1. for every x, $k(x, x')$ as a function of x' belongs to $\mathcal{H}$, and
2. k has the reproducing property $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

[a] From Rasmussen & Williams, Gaussian Processes for Machine Learning 2006

- $\|f\|^2_{\mathcal{H}}$ can be thought of as a generalization (to functions) of the Mahalanobis norm $\|y\|^2_{\Sigma} = y^\top \Sigma^{-1} y$.
- The second item is called the **reproducing property** (will become clear in the representer theorem)

Reproducing Kernel Hilbert Space

Definition (Reproducing Kernel Hilbert Space)

[a] Let $\mathcal{H}$ be a Hilbert space of real functions f defined on $\mathcal{X}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with an inner product $\langle \mathcal{X}, \mathcal{X} \rangle_{\mathcal{H}}$ (and norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:

1. for every x, $k(x, x')$ as a function of x' belongs to $\mathcal{H}$, and
2. k has the reproducing property $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

[a] From Rasmussen & Williams, Gaussian Processes for Machine Learning 2006

- $\|f\|^2_{\mathcal{H}}$ can be thought of as a generalization (to functions) of the Mahalanobis norm $\|\mathrm{y}\|^2_{\Sigma} = \mathrm{y}^\top \Sigma^{-1} \mathrm{y}$.
- The second item is called the **reproducing property** (will become clear in the representer theorem)

Moore-Aronszajn Theorem

Theorem (Moore-Aronszajn Theorem (Aronszajn 1950))

For every symmetric and positive definite function $k(\cdot, \cdot)$ *on* $\mathcal{X} \times \mathcal{X}$ *there exists a unique RKHS, and vice versa.*

- Ensures that defining a symmetric, positive definite function[1] (aka a kernel) yields a unique RKHS.

[1] discussed on next slide

Positive Definite Function

Definition

Suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then k is a positive definite function if for all $n \in \mathbb{N}$, and $x = [x_1, \ldots, x_n]^\top$ where each $x_i \in \mathcal{X}$ and $c = [c_1, \ldots, c_n]^\top \in \mathbb{R}^n$, we have

$$c^\top \mathrm{K} c \geq 0,$$

where K is the $n \times n$ matrix with entries $\mathrm{K}_{ij} = k(x_i, x_j)$.

- Functional generalization of a semi-positive definite[2] matrix:

$$x^\top \Sigma x \geq 0, \qquad \forall x \in \mathbb{R}^d$$

[2] for some reason, the "function" definition does not distinguish between semi-positive definite and positive definite; a positive definite matrix satisfies $x^\top \Sigma x > 0$.

Additional Results

Theorem (Corollary of Mercer's Theorem)

If k is a symmetric positive definite function, then there exists an inner product space V and a feature map ϕ such that $k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle_V$.

Theorem (Bochner's Theorem)

A stationary function $k(x, x') = \tilde{k}(|x - y|)$ is positive definite if and only if $\tilde{k}$ can be represented as

$$\tilde{k}(t) = \int_{\mathbb{R}} e^{itx} d\mu(x),$$

where μ is a probability measure.

Representer Theorem (Motivation)

Suppose

- $x_1, \ldots, x_n \in \mathcal{X}$
- $y_1, \ldots, y_n \in \mathbb{R}^d$
- $f : \mathcal{X} \to \mathbb{R}^d$

Interpretation:

- observe pairs of data $(x_1, y_1), \ldots, (x_n, y_n)$,
- want to recover an unknown function f from the data

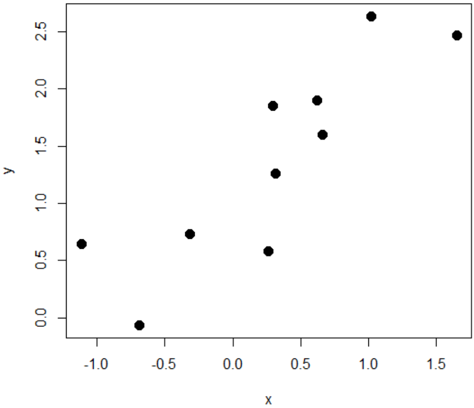
Example:

$$f(x) = y + \epsilon \qquad \text{(regression)}$$

Problem:

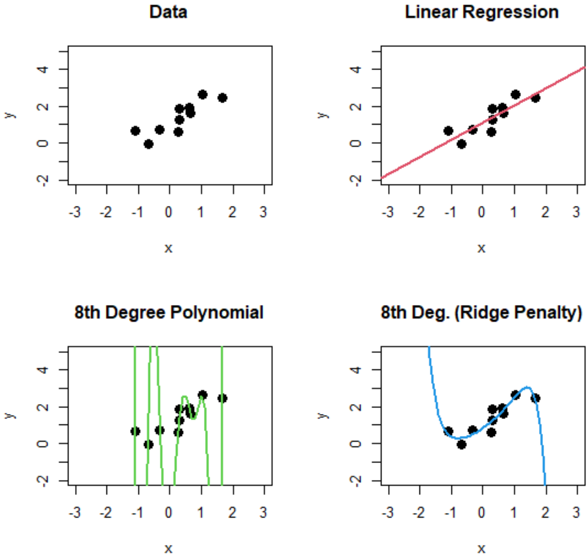
- How to choose f?

Choosing f (Issues)



What f is appropriate here?

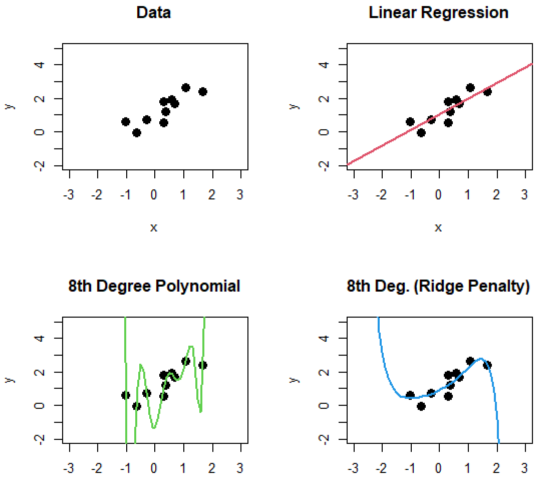
Choosing f (Issues)



Choosing f (Issues)

Perturb the data slightly...

$$x_{new} = x_{old} + 0.05\epsilon_x, \quad , y_{new} = y_{old} + 0.05\epsilon_y, \quad \epsilon_x, \epsilon_y \sim N(0,1)$$



Representer Theorem (pt 1)

Define

$$J[f] = Q(y, f) + \lambda \|f\|_{\mathcal{H}}^2$$

- $Q(y, f)$ is a data-fit term (squared error loss, negative log likelihood, etc.)
- $\lambda \|f\|_{\mathcal{H}}^2$ is the regularizer term
  - Represents smoothness assumptions on f as encoded by a suitable RKHS
  - $\lambda \in \mathbb{R}^+$ is a penalty factor

Theorem (Representer Theorem)

Let $\mathcal{H}$ be a RKHS. Each minimizer $f \in \mathcal{H}$ of $J[f]$ has the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

for some $\alpha_1, \ldots, \alpha_n$.

Representer Theorem (pt 1)

Define

$$J[f] = Q(y, f) + \lambda \|f\|_{\mathcal{H}}^2$$

- $Q(y, f)$ is a data-fit term (squared error loss, negative log likelihood, etc.)
- $\lambda \|f\|_{\mathcal{H}}^2$ is the regularizer term
  - Represents smoothness assumptions on f as encoded by a suitable RKHS
  - $\lambda \in \mathbb{R}^+$ is a penalty factor

Theorem (Representer Theorem)

Let $\mathcal{H}$ be a RKHS. Each minimizer $f \in \mathcal{H}$ of $J[f]$ has the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

for some $\alpha_1, \dots, \alpha_n$.

Representer Theorem (Specific Cases)

$$J[f] = Q(y, f) + \lambda \|f\|_{\mathcal{H}}^2$$

- Least Squares Ridge Regression ($f(\mathrm{x}_i) = \beta^\top \mathrm{x}_i$)

$$J[f] = \sum_{i=1}^{n} (y_i - \beta^\top \mathrm{x}_i)^2 + \lambda \|\beta\|_2^2 \qquad \text{(squared error loss)}$$

- Support Vector Machines

$$J[f] = \sum_{i=1}^{n} \max(0, 1 - y_i(w^\top \mathrm{x}_i - b)) + \lambda \|w\|_2^2 \qquad \text{(hinge loss)}$$

- Gaussian Process Regression

$$J[f] = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\mathrm{x}_i))^2 + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \qquad \text{(Gaussian likelihood)}$$

Representer Theorem (Specific Cases)

$$J[f] = Q(y, f) + \lambda \|f\|_{\mathcal{H}}^2$$

- Least Squares Ridge Regression $(f(\mathrm{x}_i) = \beta^\top \mathrm{x}_i)$

$$J[f] = \sum_{i=1}^{n} (y_i - \beta^\top \mathrm{x}_i)^2 + \lambda \|\beta\|_2^2 \qquad \text{(squared error loss)}$$

- Support Vector Machines

$$J[f] = \sum_{i=1}^{n} \max(0, 1 - y_i(w^\top \mathrm{x}_i - b)) + \lambda \|w\|_2^2 \qquad \text{(hinge loss)}$$

- Gaussian Process Regression

$$J[f] = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\mathrm{x}_i))^2 + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \qquad \text{(Gaussian likelihood)}$$

Representer Theorem (Specific Cases)

$$J[f] = Q(y, f) + \lambda \|f\|_{\mathcal{H}}^2$$

- Least Squares Ridge Regression $(f(\mathrm{x}_i) = \beta^\top \mathrm{x}_i)$

$$J[f] = \sum_{i=1}^{n} (y_i - \beta^\top \mathrm{x}_i)^2 + \lambda \|\beta\|_2^2 \qquad \text{(squared error loss)}$$

- Support Vector Machines

$$J[f] = \sum_{i=1}^{n} \max(0, 1 - y_i(w^\top \mathrm{x}_i - b)) + \lambda \|w\|_2^2 \qquad \text{(hinge loss)}$$

- Gaussian Process Regression

$$J[f] = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\mathrm{x}_i))^2 + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \qquad \text{(Gaussian likelihood)}$$

Using the Representer Theorem (RKHS Norm)

- Representer Theorem: The minimizer has form $f(\mathrm{x}) = \sum_{i=1}^{n} \alpha_i k(\mathrm{x}, \mathrm{x}_i)$
- Reproducing Property: $\langle k(\cdot, \mathrm{x}_i), k(\cdot, \mathrm{x}_j)\rangle_{\mathcal{H}} = k(\mathrm{x}_i, \mathrm{x}_j)$

$$
\begin{aligned}
\|f\|_{\mathcal{H}} = \|f(\cdot)\|_{\mathcal{H}} &= \left\| \sum_{i=1}^{n} \alpha_i k(\cdot, \mathrm{x}_i) \right\|_{\mathcal{H}} && \text{(representer theorem)} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, \mathrm{x}_i), \sum_{j=1}^{n} \alpha_j k(\cdot, \mathrm{x}_j) \right\rangle_{\mathcal{H}} && \text{(write as inner product)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle k(\cdot, \mathrm{x}_i), k(\cdot, \mathrm{x}_j)\rangle_{\mathcal{H}} && \text{(inner product bilinearity)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) && \text{(reproducing property)} \\
&= \alpha^\top K \alpha
\end{aligned}
$$

Using the Representer Theorem (GP Case)

In Gaussian Process Regression:

$$J[f] = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\mathrm{x}_i))^2 + \frac{1}{2}\|f\|_{\mathcal{H}}^2$$
$$= \frac{1}{2\sigma^2}(\mathrm{y} - K\alpha)^\top(\mathrm{y} - K\alpha) + \frac{1}{2}\alpha^\top K\alpha$$
$$= \frac{1}{2}\alpha^\top\left(K + \frac{1}{2\sigma^2}K^\top K\right)\alpha - \frac{1}{2\sigma^2}y^\top K\alpha + \frac{1}{2\sigma^2}y^\top y$$

Minimize J with respect to $\alpha = [\alpha_1, \ldots, \alpha_n]^\top$:

$$\Rightarrow \hat{\alpha} = (K + \sigma^2 I)^{-1}\mathrm{y}$$

$$\Rightarrow \hat{f}(\mathrm{x}_*) = \sum_{i=1}^{n}\hat{\alpha}_i k(\mathrm{x}_*, \mathrm{x}_i) = \mathrm{k}(\mathrm{x}_*)^\top(K + \sigma^2 I)^{-1}\mathrm{y}$$

where $\mathrm{k}(\mathrm{x}_*) = [k(\mathrm{x}_*, \mathrm{x}_1), \ldots, k(\mathrm{x}_*, \mathrm{x}_n)]^\top$.

Using the Representer Theorem (GP Case)

In Gaussian Process Regression:

$$
\begin{aligned}
J[f] &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\mathrm{x}_i))^2 + \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\
&= \frac{1}{2\sigma^2}(\mathrm{y} - K\alpha)^\top(\mathrm{y} - K\alpha) + \frac{1}{2}\alpha^\top K\alpha \\
&= \frac{1}{2}\alpha^\top\left(K + \frac{1}{2\sigma^2}K^\top K\right)\alpha - \frac{1}{2\sigma^2}y^\top K\alpha + \frac{1}{2\sigma^2}y^\top y
\end{aligned}
$$

Minimize J with respect to $\alpha = [\alpha_1, \ldots, \alpha_n]^\top$:

$$\Rightarrow \hat{\alpha} = (K + \sigma^2 I)^{-1}\mathrm{y}$$

$$\Rightarrow \hat{f}(\mathrm{x}_*) = \sum_{i=1}^{n}\hat{\alpha}_i k(\mathrm{x}_*, \mathrm{x}_i) = \mathrm{k}(\mathrm{x}_*)^\top(K + \sigma^2 I)^{-1}\mathrm{y}$$

where $\mathrm{k}(\mathrm{x}_*) = [k(\mathrm{x}_*, \mathrm{x}_1), \ldots, k(\mathrm{x}_*, \mathrm{x}_n)]^\top$.

Using the Representer Theorem (GP Case)

In Gaussian Process Regression:

$$
\begin{aligned}
J[f] &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\mathrm{x}_i))^2 + \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\
&= \frac{1}{2\sigma^2}(\mathrm{y} - K\alpha)^\top(\mathrm{y} - K\alpha) + \frac{1}{2}\alpha^\top K\alpha \\
&= \frac{1}{2}\alpha^\top\left(K + \frac{1}{2\sigma^2}K^\top K\right)\alpha - \frac{1}{2\sigma^2}y^\top K\alpha + \frac{1}{2\sigma^2}y^\top y
\end{aligned}
$$

Minimize J with respect to $\alpha = [\alpha_1, \ldots, \alpha_n]^\top$:

$$\Rightarrow \hat{\alpha} = (K + \sigma^2 I)^{-1}\mathrm{y}$$

$$\Rightarrow \hat{f}(\mathrm{x}_*) = \sum_{i=1}^{n}\hat{\alpha}_i k(\mathrm{x}_*, \mathrm{x}_i) = \mathrm{k}(\mathrm{x}_*)^\top(K + \sigma^2 I)^{-1}\mathrm{y}$$

where $\mathrm{k}(\mathrm{x}_*) = [k(\mathrm{x}_*, \mathrm{x}_1), \ldots, k(\mathrm{x}_*, \mathrm{x}_n)]^\top$.

Recap: Kernel Method Roadmap

Goal: recover f from data $(x_1, y_1), \ldots, (x_n, y_n)$

1. Choose a kernel (symmetric, positive definite function)
   - Imposes restrictions on f
2. Representer theorem ensures a minimizer to the penalized minimization problem

$$J[f] = Q(y, f) + \lambda \|f\|_{\mathcal{H}}^2$$

Two Common Kernel Methods

- **Gaussian Process Regression**: $k(\mathrm{x}, \mathrm{x}') = \mathrm{cov}(f(\mathrm{x}), f(\mathrm{x}'))$

- **Support Vector Machines**: maps input space into feature space:

$$k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle_V$$

  where $\phi : \mathcal{X} \to V$ is a map that transforms the input data to be more appropriate to the task at hand

  - Can be done with Mercer's theorem by choosing an appropriate kernel

Focus on Gaussian Processes

- In this work we focus on Gaussian process regression
- Kernels have similar interpretation in other methods (e.g. support vector machines, kernel ridge regression, kernel PCA)

Gaussian Process Interlude

Definition (Gaussian Process)

Let $f : \mathcal{X} \to \mathbb{R}$. Then f is a **Gaussian process** if for all $n \in \mathbb{N}$, the vector $[f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n)]^\top$ is multivariate normal.

- Specified by
  - mean function μ: $\mathbb{E}[f(x)] = \mu(x)$
  - covariance kernel k: $\mathrm{cov}(f(x), f(x')) = k(x, x')$
- Generalization of a multivariate normal distribution to infinite dimensional indices

The covariance kernel k is crucial – it determines underlying properties of f considering it as a function of x, e.g.

- continuity,
- differentiability,
- overall shape (linear? polynomial? periodic?)

Gaussian Process Interlude

Definition (Gaussian Process)

Let $f : \mathcal{X} \to \mathbb{R}$. Then f is a **Gaussian process** if for all $n \in \mathbb{N}$, the vector $[f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n)]^\top$ is multivariate normal.

- Specified by
  - mean function μ: $\mathbb{E}[f(x)] = \mu(x)$
  - covariance kernel k: $\mathrm{cov}(f(x), f(x')) = k(x, x')$
- Generalization of a multivariate normal distribution to infinite dimensional indices

The covariance kernel k is crucial – it determines underlying properties of f considering it as a function of x, e.g.

- continuity,
- differentiability,
- overall shape (linear? polynomial? periodic?)

Gaussian Process Regression

Given data $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$, assume

1. $y_i = f(\mathrm{x}_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
2. f is a Gaussian process with mean function μ and covariance kernel k
    - Without loss of generality, assume $\mu = 0$

Then if x_* is a test point, $[y_1, \ldots, y_n, f(\mathrm{x}_*)]^\top$ is multivariate normal and thus

$$f(\mathrm{x}_*) | y_1, \ldots, y_n \sim N(m(\mathrm{x}_*), s(\mathrm{x}_*, \mathrm{x}_*))$$

where

$$\begin{aligned} m(\mathrm{x}_*) &= \mathrm{k}(\mathrm{x}_*)^\top \left[K + \sigma^2 I\right]^{-1} y, \\ s(\mathrm{x}_*, \mathrm{x}_*) &= k(\mathrm{x}_*, \mathrm{x}_*) - \mathrm{k}(\mathrm{x}_*)^\top \left[K + \sigma^2 I\right]^{-1} \mathrm{k}(\mathrm{x}_*) \end{aligned}$$

Consistency With Representer Theorem

- Using $Q(y, f) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i))^2$ (Gaussian likelihood)
- posterior mean function m is the minimizer of $J[f]$, i.e.

$$m = \operatorname{argmin}_{f \in \mathcal{H}} J[f] = \operatorname{argmin}_{f \in \mathcal{H}} \left\{ Q(y, f) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right\}$$

Hence:

- Conditions on the covariance kernel k determines behavior of the posterior mean function
- The posterior Gaussian process f itself (i.e. with mean function m and covariance kernel $s(\cdot, \cdot)$) has slightly different, but related properties

In other settings, e.g. support vector machines, replace m with the (kevin's thesis?)

Consistency With Representer Theorem

- Using $Q(y, f) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i))^2$ (Gaussian likelihood)
- posterior mean function m is the minimizer of $J[f]$, i.e.

$$m = \text{argmin}_{f \in \mathcal{H}} J[f] = \text{argmin}_{f \in \mathcal{H}} \left\{ Q(y, f) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right\}$$

Hence:

- Conditions on the covariance kernel k determines behavior of the posterior mean function
- The posterior Gaussian process f itself (i.e. with mean function m and covariance kernel $s(\cdot, \cdot)$) has slightly different, but related properties

In other settings, e.g. support vector machines, replace m with the (kevin's thesis?)

The Punchline

- The kernel determines several properties of the statistical problem at hand
- The following slides provide examples of commonly used kernels, along with some real world examples

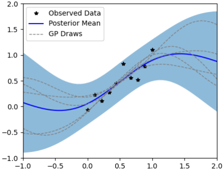
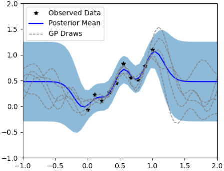
Common Kernels (Squared Exponential Kernel)

Let $x, x' \in \mathbb{R}$ for simplicity

- **Squared Exponential Kernel**[3]

$$k_{\text{SE}}(x, x') = \eta^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

  - ℓ is a lengthscale that determines the length of information borrowing in the function.
  - η^2 determines the average distance the function is away from its mean.
  - Gaussian processes with this kernel are infinitely differentiable.

$\ell = 0.857$	$\ell = 0.163$
	

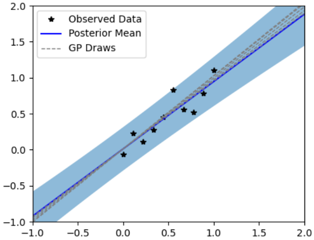
[3] also called the radial basis function kernel, or Gaussian kernel

Common Kernels (Linear Kernel)

- **Linear Kernel**

$$k_{\text{Lin}}(x, x') = \sigma_b^2 + \sigma_v^2 (x - c)(x' - c)$$

  - The offset c determines the x-coordinate of the point that all lines in the posterior go through
  - The constant variance σ_b^2 determines how far from 0 the height of the function will be at $x = 0$.
  - Gaussian processes with this kernel corresponds **exactly** with Bayesian linear regression



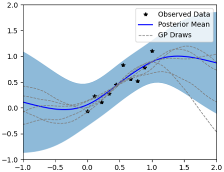
- **Matérn**

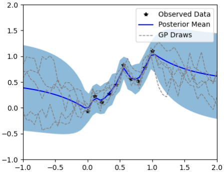
$$k_{\text{Mat}}(x, x'; \nu) = \frac{\eta^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{\ell} |x - x'| \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}}{\ell} |x - x'| \right)$$

Where $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is a modified Bessel function

- ℓ is a lengthscale
- ν controls the smoothness of f
    - The resulting Gaussian process is $\nu-$times differentiable
    - e.g. $\nu = 2.5 \Rightarrow f$ is 2 times differentiable, $\nu = 0.5 \Rightarrow f$ is not differentiable

$\nu = 2.5$	$\nu = 0.5$



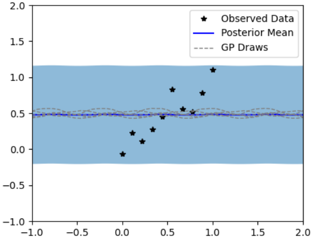


- **Periodic Kernel**

$$k_{\mathrm{Per}}(x, x') = \eta^2 \exp\left(-\frac{2\sin^2(\pi|x - x'|/p)}{\ell^2}\right)$$

Where $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is a modified Bessel function

  - ℓ is a lengthscale
  - p determines the period (distance between repeating patterns of the function)



More Complex Kernels

- **Changepoint Kernels**
  - Expresses change from one kernel to another
- **Heteroskedastic Kernel**
  - Automatically accounts for varying noise amplitude
- **Translation and Rotation Invariant Kernels**
  - Useful with image data

See Automatic Model Construction with Gaussian Processes by *Duvenaud* for more examples and thorough discussion: https://www.cs.toronto.edu/ duvenaud/thesis.pdf

Constructing New Kernels

Two common ways to construct new kernels:

- Adding two kernels yields a kernel[4]

$$k_{a+b}(x, x') = k_a(x, x') + k_b(x, x')$$

- Multiplying two kernels yields a kernel

$$k_{a \cdot b}(x, x') = k_a(x, x') \cdot k_b(x, x')$$

[4] recall by kernel, we mean a symmetric and positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

Additive Kernel (Interpretation)

In a Gaussian process, if

$$f_1 \sim \mathrm{GP}(\mu_1, k_1)$$
$$f_2 \sim \mathrm{GP}(\mu_2, k_2)$$

Then

$$f_1 + f_2 \sim \mathrm{GP}(\mu_1 + \mu_2, k_1 + k_2).$$

Kernel Multiplication and Dimensionality

If $\mathrm{x} = [x^{(1)}, \cdots, x^{(d)}]^\top \in \mathbb{R}^d$, it may make sense to define

$$k(\mathrm{x}, \mathrm{x}') = \prod_{j=1}^{d} k_j(x^{(j)}, x'^{(j)})$$

Example. Suppose $[x^{(1)}, x^{(2)}]^\top \in \mathbb{R}^2$ where

- $x^{(1)}$ represents an individuals age, and
- $x^{(2)}$ represents the current calendar year.

$$k(\mathrm{x}, \mathrm{x}') = k_1(x^{(1)}, x'^{(1)}) \cdot k_2(x^{(2)}, x'^{(2)})$$

For example

$$k(\mathrm{x}, \mathrm{x}') = \eta^2 \exp\left(\frac{-|x^{(1)} - x'^{(1)}|^2}{2\ell_{\text{age}}}\right) \cdot \exp\left(\frac{-|x^{(2)} - x'^{(2)}|^2}{2\ell_{\text{year}}}\right)$$

Anecdotal Example Of Kernel Abuse

4.2. Covariance Functions

The covariance function plays the central role in GPR as it encodes our assumptions about the underlying process by defining the similarity between functions. We model the image as a locally stationary Gaussian Process and choose the squared exponential covariance function:

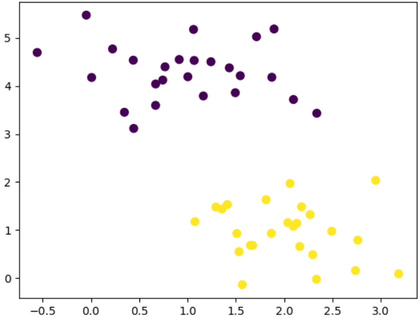
$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{\ell^2}\right), \quad (10)$$

SVM Classification

SVM is generally a classification method.

Task:

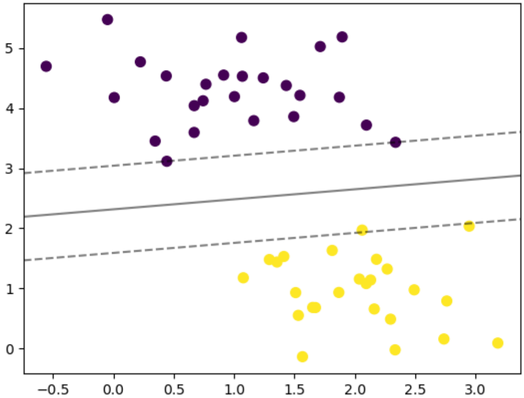
- Decide a rule that labels a point to be purple or yellow.
- The mechanics of the rule with SVM are dependent on the kernel chosen.



Linear Decision Boundary

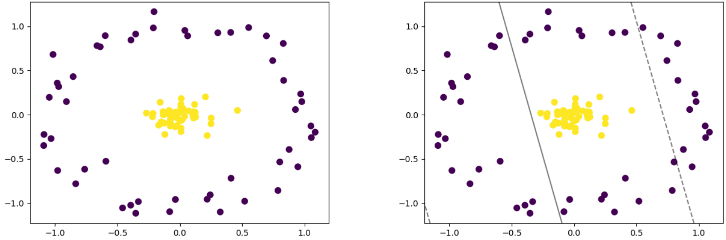
Choosing the linear kernel yields a linear decision boundary:

$$k(x, x') = x^\top x'$$



“Circular” Decision Boundary

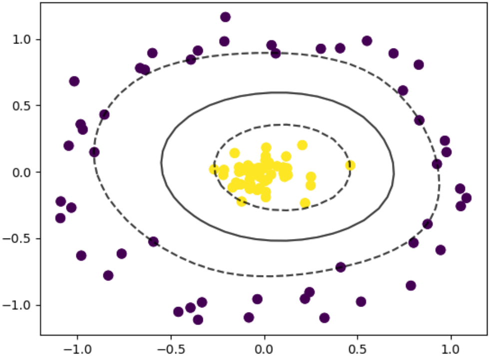
A linear decision boundary is **not** appropriate here...



“Circular” Decision Boundary

The radial basis function kernel gives a decision boundary based on “closeness” of points

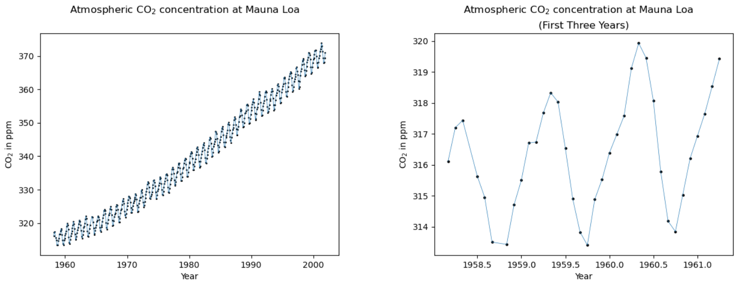
$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\theta^2}\right)$$



Example: Mauna Loa Data Set

- y: **monthly** average atmospheric CO_2 concentrations (in ppm by volume) derived from air samples at the Mauna Loa Observatory, Hawaii, between 1958 and 2003, with some missing values
- x: month

Goal: model $f(x)$



Example: Mauna Loa Data Set (Kernel Choice)

Model the apparent features[5]:

- Long term rising trend

$$k_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right)$$

where θ_1 is the amplitude, and θ_2 is the characteristic length-scale

- Yearly decaying periodicity

$$k_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2}\right) \exp\left(-\frac{2\sin^2(\pi(x - x'))}{2\theta_5^2}\right)$$

where θ_3 is the magnitude, θ_4 is the decay-time, and θ_5 is the smoothness of the periodic component.

[5] This particular construction is taken from Gaussian Processes for Machine Learning by Rasmussen and Williams

Example: Mauna Loa Data Set (Kernel Choice)

Model the apparent features[5]:

- Long term rising trend

$$k_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right)$$

where θ_1 is the amplitude, and θ_2 is the characteristic length-scale

- Yearly decaying periodicity

$$k_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2}\right) \exp\left(-\frac{2\sin^2(\pi(x - x'))}{2\theta_5^2}\right)$$

where θ_3 is the magnitude, θ_4 is the decay-time, and θ_5 is the smoothness of the periodic component.

[5] This particular construction is taken from Gaussian Processes for Machine Learning by Rasmussen and Williams

Example: Mauna Loa Data Set (Kernel Choice, Continued)

- (Small) medium term irregularities

$$k_3(x, x') = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8 \theta_7^2}\right)^{-\theta_8}$$

where θ_6 is the magnitude, θ_7 is the typical length-scale, and θ_8 is the shape parameter

- Noise term

$$k_4(x, x') = \theta_9^2 \exp\left(-\frac{(x - x')^2}{2\theta_{10}^2}\right) + \theta_{11}^2 \delta_{x=x'},$$

where θ_9 is the magnitude of the correlated noise component, θ_{10} is its length-scale, and θ_{11} is the magnitude of the independent noise component.

Final covariance function:

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x')$$

Example: Mauna Loa Data Set (Kernel Choice, Continued)

- (Small) medium term irregularities

$$k_3(x, x') = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$$

where θ_6 is the magnitude, θ_7 is the typical length-scale, and θ_8 is the shape parameter

- Noise term

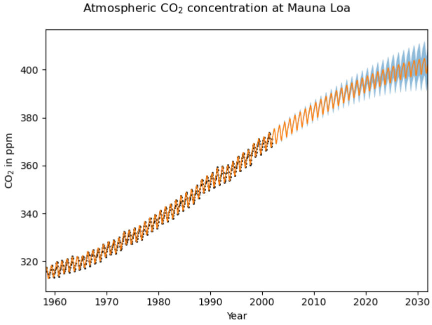
$$k_4(x, x') = \theta_9^2 \exp\left(-\frac{(x - x')^2}{2\theta_{10}^2}\right) + \theta_{11}^2 \delta_{x=x'},$$

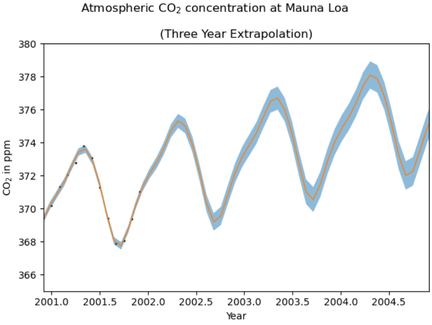
where θ_9 is the magnitude of the correlated noise component, θ_{10} is its length-scale, and θ_{11} is the magnitude of the independent noise component.

Final covariance function:

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x')$$

Example: Mauna Loa Data Set (Posterior Prediction)






```
Learned kernel:
2.63**2 * RBF(length_scale=51.6) +
0.155**2 * RBF(length_scale=91.5) * ExpSineSquared(length_scale=1.48,
                                                   periodicity=1) +
0.0314**2 * RationalQuadratic(alpha=2.89, length_scale=0.968) +
0.011**2 * RBF(length_scale=0.122) + WhiteKernel(noise_level=0.000126)
```

Duvenaud's Thesis (Part 1)

Automatic Model Construction with Gaussian Processes by *Duvenaud* gives an algorithm that searches over kernel combinations and expresses the structure discovered

Example[6]

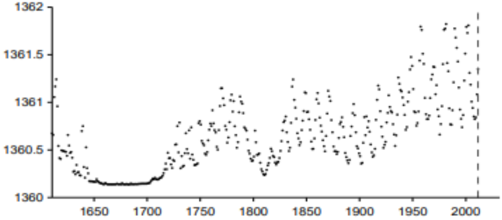


Figure 4.1: Solar irradiance data (Lean et al., 1995).

[6] Figures are taken from *Duvenaud*

Duvenaud's Thesis (Part 2)

This component is constant. This component applies from 1643 until 1716.

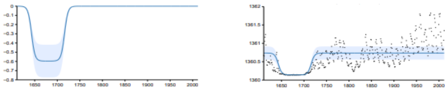


Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

This component is a smooth function with a typical lengthscale of 23.1 years. This component applies until 1643 and from 1716 onwards.

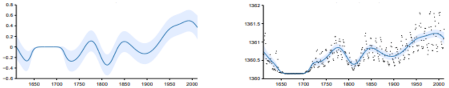


Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.

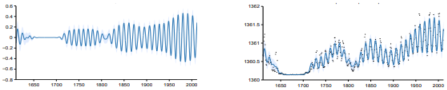


Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

Ongoing and Completed Projects (Part 1)

Gaussian Process Models for Computer Vision (*Student Thesis (Hakeem Frank)*)

- Comparing classification metrics in changing kernels (using a GP classifier), in three settings: handwritten digit classification, object detection (airplane or not), brain scans (tumor detection)
- Found that results varied heavily among using polynomial, linear, and squared exponential kernels

Sample table (handwritten digit classification)

Model	Accuracy	time (m)
Polynomial-2	**99.40 %**	34.88
Polynomial-3	99.30 %	28.75
Squared Exponential	98.20 %	28.58
Rational Quadratic	98.20 %	54.22
Matern 5/2	98.00 %	33.61
Matern 3/2	97.90 %	38.77
Dot Product	96.10 %	16.54
Logistic Regression	95.90 %	**0.10**
SVM_RBF	98.05 %	0.29

Ongoing and Completed Projects (Part 2)

Kernel Selection in Gaussian Process Superresolution (*Student Thesis (Charles Amelin)*)

- Comparing kernels in image "superresolution" techniques
- Current literature almost exclusively uses squared exponential kernel
- Preliminary results show that images with sharp details (e.g. corners of stairs) are upscaled with better details in more relaxed kernels (e.g. Matérn kernel)

Ongoing and Completed Projects (Part 3)

Kernel Selection in Multipopulation Mortality Modelling

- Idea: use a special kernel that allows for vector-valued functions
- Model multi-population mortality through latent GP's

Example.

$$\begin{aligned} f_{\mathrm{USA,M}}(x) &= a_{1,1}u_1(x) + a_{1,2}u_2(x) + a_{1,3}u_3(x) \\ f_{\mathrm{USA,F}}(x) &= a_{2,1}u_1(x) + a_{2,2}u_2(x) + a_{2,3}u_3(x) \\ f_{\mathrm{JPN,M}}(x) &= a_{3,1}u_1(x) + a_{3,2}u_2(x) + a_{3,3}u_3(x) \\ f_{\mathrm{JPN,F}}(x) &= a_{4,1}u_1(x) + a_{4,2}u_2(x) + a_{4,3}u_3(x) \end{aligned}$$

- Model latent GPs $[u_1(x), u_2(x), u_3(x)]^\top$ as a vector-valued GP
- $a_{i,j}$ coefficients are hyperparameters
- Latent GPs can express unique fundamental mortality structures through different kernels
- There exists a tensor covariance structure which significantly reduces fitting time

Conclusion

- Kernel methods are gaining in popularity
- Kernel choice is a nontrivial topic
  - If there is domain knowledge, the modeler can use this in choosing a kernel
  - If there is no domain knowledge, the modeler can try different kernels similarly to model selection

References

- Williams, Christopher KI, and Carl Edward Rasmussen. "Gaussian processes for regression." (1996).
- Duvenaud, David. "Automatic model construction with Gaussian processes." *Diss. University of Cambridge*, (2014).
- Aronszajn, Nachman. "Theory of reproducing kernels." *Transactions of the American mathematical society* 68.3 (1950): 337-404.
- Huynh, Nhan, and Mike Ludkovski. "Multi-Output Gaussian Processes for Multi-Population Longevity Modeling." *arXiv preprint arXiv:2003.02443* (2020).
- Frank, Hakeem. "Gaussian Process Models for Computer Vision." *Diss. California State Polytechnic University, Pomona*, (2020).