UNIVERSITY OF CAMBRIDGE

Neural Nets, $\mathcal{GP}$s, and where the kernel lives

Paul Rubenstein[12] Matthias Bauer[12]

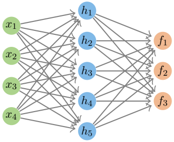
[1]University of Cambridge

[2]Max-Planck Institute for Intelligent Systems, Tübingen

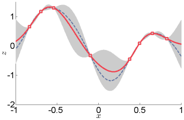
12th November 2015

Agenda

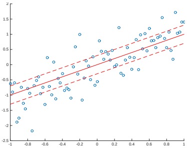
Priors over infinite NN = $\mathcal{GP}$



Relationship between Kernel Ridge Regression and GPs



Support Vector Regression and GPs



Acknowledgements/References

This talk is based mostly on the following:

- Arthur Gretton's course on RKHS theory: `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html`
- Bishop's Pattern Recognition and Machine Learning
- Stulp and Sigaud, Many regression algorithms, one unified model: A review

Ordinary Least Squares (OLS) Linear Regression

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^p$ and $y_i \in \mathcal{Y} = \mathbb{R}$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ that explains* $\mathcal{D}$.

Ordinary Least Squares (OLS) Linear Regression

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^p$ and $y_i \in \mathcal{Y} = \mathbb{R}$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ that explains* $\mathcal{D}$.

An approach:

- Assume f is linear: $f(\mathbf{x}) = \mathbf{x}^\intercal \beta$ for some β
- Choose β to minimise the sum of squared errors.

Ordinary Least Squares (OLS) Linear Regression

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^p$ and $y_i \in \mathcal{Y} = \mathbb{R}$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ that explains* $\mathcal{D}$.

An approach:

- Assume f is linear: $f(\mathbf{x}) = \mathbf{x}^\intercal \beta$ for some β
- Choose β to minimise the sum of squared errors.

Writing $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^\intercal$ and $Y = (y_1, \ldots, y_N)^\intercal$, we wish to minimise

$$L(\beta) = (Y - X\beta)^\intercal (Y - X\beta)$$

OLS (cont)

$$\begin{aligned} L(\beta) &= (Y - X\beta)^\intercal (Y - X\beta) \\ &= Y^\intercal Y - 2\beta^\intercal X^\intercal Y + \beta^\intercal X^\intercal X\beta \\ \implies \frac{dL}{d\beta} &= -2X^\intercal Y + 2X^\intercal X\beta \end{aligned}$$

So $\frac{dL}{d\beta} = 0 \implies \beta = (X^\intercal X)^{-1} X^\intercal Y$ if $(X^\intercal X)^{-1}$ exists.

OLS (cont)

Two problems.

1. What if $X^\intercal X$ is not invertible?
2. What if y is not well approximated by a linear function of $\mathbf{x}$?

OLS (cont)

Two problems.

1. What if $X^\intercal X$ is not invertible?
2. What if y is not well approximated by a linear function of $\mathbf{x}$?

Solutions:

1. Eigenvalues of $X^\intercal X$ are always ≥ 0
   $\implies X^\intercal X + \lambda I$ invertible for $\lambda > 0$... why?
2. Can replace $\mathbf{x}$ with $\phi(\mathbf{x})$, where $\phi : \mathcal{X} \longrightarrow \mathbb{R}^p$.
   Write $\Phi := \phi(X)$

Ridge Regression in Feature Space

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y} = \mathbb{R}$ and feature map $\phi : \mathcal{X} \longrightarrow \mathbb{R}^p$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ so that $f \circ \phi$ explains* $\mathcal{D}$.

Ridge Regression in Feature Space

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y} = \mathbb{R}$ and feature map $\phi : \mathcal{X} \longrightarrow \mathbb{R}^p$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ so that $f \circ \phi$ explains* $\mathcal{D}$.

An approach:

- Assume f is linear: $f(\phi(\mathbf{x})) = \phi(\mathbf{x})^\intercal \beta$ for some β
- If p is large compared to N then we may overfit
- So choose β to minimise the sum of squared errors plus complexity penalty.

$$
\begin{aligned}
L(\beta) &= \sum_i (f(\phi(\mathbf{x}_i)) - y_i)^2 + \lambda \|\beta\|^2 \\
&= (Y - \Phi\beta)^\intercal (Y - \Phi\beta) + \lambda \beta^\intercal \beta
\end{aligned}
$$

Ridge Regression in Feature Space

$$L(\beta) = (Y - \Phi\beta)^{\intercal}(Y - \Phi\beta) + \lambda\beta^{\intercal}\beta$$

$$\begin{aligned}\frac{dL}{d\beta} = 0 &\implies 0 = -2\Phi^{\intercal}Y + 2\Phi^{\intercal}\Phi\beta + 2\lambda\beta \\ &\implies \beta = (\Phi^{\intercal}\Phi + \lambda_p I)^{-1}\Phi^{\intercal}Y\end{aligned}$$

Ridge Regression in Feature Space

$$\beta = (\Phi^\intercal \Phi + \lambda_p I)^{-1} \Phi^\intercal Y$$

Ridge Regression in Feature Space

$$\beta = (\Phi^\intercal \Phi + \lambda_p I)^{-1} \Phi^\intercal Y$$

Two observations:

- $\Phi^\intercal \Phi$ is NOT the Gram matrix

Ridge Regression in Feature Space

$$\beta = (\Phi^\intercal \Phi + \lambda_p I)^{-1} \Phi^\intercal Y$$

Two observations:

- $\Phi^\intercal \Phi$ is NOT the Gram matrix
- $(\Phi^\intercal \Phi + \lambda I_p)\Phi^\intercal = \Phi^\intercal \Phi \Phi^\intercal + \lambda \Phi^\intercal = \Phi^\intercal (\Phi \Phi^\intercal + \lambda I_N)$

Ridge Regression in Feature Space

$$\beta = (\Phi^\intercal \Phi + \lambda_p I)^{-1} \Phi^\intercal Y$$

Two observations:

- $\Phi^\intercal \Phi$ is NOT the Gram matrix
- $(\Phi^\intercal \Phi + \lambda I_p)\Phi^\intercal = \Phi^\intercal \Phi \Phi^\intercal + \lambda \Phi^\intercal = \Phi^\intercal(\Phi\Phi^\intercal + \lambda I_N)$

All eigenvalues of $\Phi^\intercal \Phi$ and $\Phi\Phi^\intercal$ are ≥ 0 and so both bracketed expressions are invertible. Thus

$$\Phi^\intercal(\Phi\Phi^\intercal + \lambda I_N)^{-1} = (\Phi^\intercal \Phi + \lambda I_p)^{-1}\Phi^\intercal$$

Regularised feature-mapped regression

So instead we can write

$$\beta = \Phi^\intercal(\Phi\Phi^\intercal + \lambda I_N)^{-1}Y$$
$$\implies f(\mathbf{x}_*) = \phi(\mathbf{x}_*)^\intercal\Phi^\intercal(\Phi\Phi^\intercal + \lambda I_N)^{-1}Y$$

Regularised feature-mapped regression

So instead we can write

$$\beta = \Phi^\intercal(\Phi\Phi^\intercal + \lambda I_N)^{-1}Y$$
$$\implies f(\mathbf{x}_*) = \phi(\mathbf{x}_*)^\intercal\Phi^\intercal(\Phi\Phi^\intercal + \lambda I_N)^{-1}Y$$

Some reasons this might be good:

- If $p > N$ then the matrix inversion takes $O(N^3)$ operations compared to $O(p^3)$
- $\phi(\mathbf{x})$ only ever appears as an inner product - so might not need to explicitly represent ϕ

Definition (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *kernel* if it is symmetric and if, for any $x_1, \ldots, x_n \in \mathcal{X}$, the matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

Definition (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *kernel* if it is symmetric and if, for any $x_1, \ldots, x_n \in \mathcal{X}$, the matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

K positive semi-definite $\iff a^\intercal K a \geq 0$ for any $a \in \mathbb{R}^n$.

Definition (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *kernel* if it is symmetric and if, for any $x_1, \ldots, x_n \in \mathcal{X}$, the matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

K positive semi-definite $\iff$ $a^\intercal K a \geq 0$ for any $a \in \mathbb{R}^n$.

Example: Let $\phi : \mathcal{X} \longrightarrow \mathcal{H}$ be any map into a Hilbert space, then $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ is a kernel.

- **Symmetry**: inherited from $\langle ., . \rangle$
- **+ve semidefinite**:
  $a^\intercal K a = \sum_{ij} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} = \| \sum_i a_i \phi(x_i) \|^2_{\mathcal{H}} \geq 0$

Brief Introduction to RKHS Theory

Definition (Reproducing Kernel Hilbert Space)

Let $\mathcal{H}$ be a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$. We say that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) if the evaluation operators $\delta_x : \mathcal{H} \longrightarrow \mathbb{R}, f \mapsto f(x)$ are continuous for all $x \in \mathcal{X}$

Brief Introduction to RKHS Theory

Definition (Reproducing Kernel Hilbert Space)

Let $\mathcal{H}$ be a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$. We say that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) if the evaluation operators $\delta_x : \mathcal{H} \longrightarrow \mathbb{R}, f \mapsto f(x)$ are continuous for all $x \in \mathcal{X}$

δ_x continuous means...

- convergence in norm of a sequence of functions implies pointwise convergence at every point so functions are 'smooth'

Brief Introduction to RKHS Theory

Definition (Reproducing Kernel Hilbert Space)

Let $\mathcal{H}$ be a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$. We say that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) if the evaluation operators $\delta_x : \mathcal{H} \longrightarrow \mathbb{R}, f \mapsto f(x)$ are continuous for all $x \in \mathcal{X}$

δ_x continuous means...

- convergence in norm of a sequence of functions implies pointwise convergence at every point so functions are ‘smooth’
- by Riesz, there exists a unique $\phi_x \in \mathcal{H}$ such that $f(x) = \langle f, \phi_x \rangle$ for all $f \in \mathcal{H}$.

Definition (Reproducing Kernel Hilbert Space)

Let $\mathcal{H}$ be a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$. We say that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) if the evaluation operators $\delta_x : \mathcal{H} \longrightarrow \mathbb{R}, f \mapsto f(x)$ are continuous for all $x \in \mathcal{X}$

δ_x continuous means...

- convergence in norm of a sequence of functions implies pointwise convergence at every point so functions are 'smooth'
- by Riesz, there exists a unique $\phi_x \in \mathcal{H}$ such that $f(x) = \langle f, \phi_x \rangle$ for all $f \in \mathcal{H}$.

We call $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}, (x, x') \mapsto \langle \phi_x, \phi_{x'} \rangle$ the (unique) Reproducing Kernel of $\mathcal{H}$

Brief Introduction to RKHS Theory

Summary:

- An RKHS on a base set $\mathcal{X}$ is just[1] a set of functions $\mathcal{X} \longrightarrow \mathbb{R}$
- Given an RKHS, we can construct a kernel on $\mathcal{X}$

[1] with some previously mentioned caveats

Summary:

- An RKHS on a base set $\mathcal{X}$ is just[1] a set of functions $\mathcal{X} \longrightarrow \mathbb{R}$
- Given an RKHS, we can construct a kernel on $\mathcal{X}$

Remarkably, the converse holds.

Theorem (Moore-Aronszajn)

Suppose that $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a kernel. Then there exists an RKHS $\mathcal{H}$ and feature map $\phi : \mathcal{X} \longrightarrow \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

[1] with some previously mentioned caveats

Summary:

- An RKHS on a base set $\mathcal{X}$ is just[1] a set of functions $\mathcal{X} \longrightarrow \mathbb{R}$
- Given an RKHS, we can construct a kernel on $\mathcal{X}$

Remarkably, the converse holds.

Theorem (Moore-Aronszajn)

Suppose that $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a kernel. Then there exists an RKHS $\mathcal{H}$ and feature map $\phi : \mathcal{X} \longrightarrow \mathcal{H}$ such that
$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

- $\mathcal{H}$ is the smallest Hilbert space containing each $k(\cdot, x)$
- properties of functions determined through properties of $k(\cdot, x)$

[1] with some previously mentioned caveats

Kernel Ridge Regression

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y} = \mathbb{R}$
- Want to infer a function $f : \mathcal{X} \longrightarrow \mathbb{R}$ so that f explains* $\mathcal{D}$.

Kernel Ridge Regression

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y} = \mathbb{R}$
- Want to infer a function $f : \mathcal{X} \longrightarrow \mathbb{R}$ so that f explains* $\mathcal{D}$.

An approach:

- Pick a kernel k such that the functions $k(\cdot, x)$ are 'good'
- Consider the RKHS $\mathcal{H}$ corresponding to k
- Find the $f \in \mathcal{H}$ that minimises empirical squared error (with penalty for complexity)

$$\arg\min_{f \in \mathcal{H}} \sum_i (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Kernel Ridge Regression

$$\arg\min_{f \in \mathcal{H}} \sum_i (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}}$$

How do we find the argmin?

Kernel Ridge Regression

$$\arg\min_{f \in \mathcal{H}} \sum_i (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}}$$

How do we find the argmin? Answer:

Theorem (Representer theorem)

The solution f_* to the above problem lies in the subspace of $\mathcal{H}$ spanned by the set $\{k(\cdot, x_i) | i = 1, \ldots, N\}$. ie
$f_* = \sum_i \alpha_i k(\cdot, x_i)$ for some coefficients α_i

Kernel Ridge Regression

$$\arg\min_{f \in \mathcal{H}} \sum_i (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

How do we find the argmin? Answer:

Theorem (Representer theorem)

The solution f_* to the above problem lies in the subspace of $\mathcal{H}$ spanned by the set $\{k(\cdot, x_i) | i = 1, \ldots, N\}$. ie
$f_* = \sum_i \alpha_i k(\cdot, x_i)$ for some coefficients α_i

$$\arg\min_{\alpha \in \mathbb{R}^N} \sum_i (f_\alpha(\mathbf{x}_i) - y_i)^2 + \lambda \|f_\alpha\|_{\mathcal{H}}^2$$

Proof:

- Let $f \in \mathcal{H}$
- Let f_s be the projection of f onto $\mathrm{span}\{k(\cdot, x_i)\}$
- Let $f_\perp = f - f_s \perp \mathrm{span}\{k(\cdot, x_i)\}$

Kernel Ridge Regression

Proof:

- Let $f \in \mathcal{H}$
- Let f_s be the projection of f onto $\mathrm{span}\{k(\cdot, x_i)\}$
- Let $f_\perp = f - f_s \perp \mathrm{span}\{k(\cdot, x_i)\}$

We show that f_s is better than f in the sense that:

- The loss function is the same: $(f_s(x) - y)^2 = (f(x) - y)^2$
- The complexity penalty is smaller: $\|f_s\|_{\mathcal{H}}^2 \leq \|f\|_{\mathcal{H}}^2$

For each term in the loss function we have:

$$
\begin{aligned}
(f(x_i) - y_i)^2 &= (f_s(x_i) + f_\perp(x_i) - y_i)^2 \\
&= (\langle f_s, k(\cdot, x_i)\rangle + \langle f_\perp, k(\cdot, x_i)\rangle - y_i)^2 \\
&= (\langle f_s, k(\cdot, x_i)\rangle - y_i)^2 \\
&= (f_s(x_i) - y_i)^2
\end{aligned}
$$

Kernel Ridge Regression

For each term in the loss function we have:

$$\begin{aligned}(f(x_i) - y_i)^2 &= (f_s(x_i) + f_\perp(x_i) - y_i)^2 \\ &= (\langle f_s, k(\cdot, x_i)\rangle + \langle f_\perp, k(\cdot, x_i)\rangle - y_i)^2 \\ &= (\langle f_s, k(\cdot, x_i)\rangle - y_i)^2 \\ &= (f_s(x_i) - y_i)^2\end{aligned}$$

Considering the complexity penalty:

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 &= \|f_s + f_\perp\|_{\mathcal{H}}^2 \\ &= \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2\end{aligned}$$

Kernel Ridge Regression

For each term in the loss function we have:

$$\begin{aligned}(f(x_i) - y_i)^2 &= (f_s(x_i) + f_\perp(x_i) - y_i)^2 \\ &= (\langle f_s, k(\cdot, x_i)\rangle + \langle f_\perp, k(\cdot, x_i)\rangle - y_i)^2 \\ &= (\langle f_s, k(\cdot, x_i)\rangle - y_i)^2 \\ &= (f_s(x_i) - y_i)^2\end{aligned}$$

Considering the complexity penalty:

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 &= \|f_s + f_\perp\|_{\mathcal{H}}^2 \\ &= \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2\end{aligned}$$

So f_s is better than f! Thus optimal f_* must lie in $\mathrm{span}\{k(\cdot, x_i)\}$

Kernel Ridge Regression

Writing $f = \sum_j \alpha_j k(\cdot, x_j)$, we wish to minimise the following quantity over α:

$$
\begin{aligned}
L(\alpha) &= \sum_i (f(x_i) - y_i)^2 + \lambda \|f\|^2 \\
&= \sum_i (\sum_j \langle \alpha_j k(\cdot, x_j), k(\cdot, x_i) \rangle - y_i)^2 + \lambda \langle f, f \rangle \\
&= \sum_i ((K\alpha)_i - y_i)^2 + \lambda \sum_{ij} \langle \alpha_i k(\cdot, x_i), \alpha_j k(\cdot, x_j) \rangle \\
&= (K\alpha - Y)^\intercal (K\alpha - Y) + \lambda \alpha^\intercal K \alpha
\end{aligned}
$$

Kernel Ridge Regression

Differentiating with respect to α yields

$$
\begin{aligned}
\frac{dL}{d\alpha} &= 2KK\alpha - 2KY + 2\lambda K\alpha \\
&= 2K(K\alpha - Y + \lambda\alpha) \\
&= 2K((K + \lambda I_N)\alpha - Y) = 0 \\
\implies \alpha &= (K + \lambda I_N)^{-1} Y
\end{aligned}
$$

Kernel Ridge Regression

Differentiating with respect to α yields

$$
\begin{aligned}
\frac{dL}{d\alpha} &= 2KK\alpha - 2KY + 2\lambda K\alpha \\
&= 2K(K\alpha - Y + \lambda\alpha) \\
&= 2K((K + \lambda I_N)\alpha - Y) = 0 \\
\implies \alpha &= (K + \lambda I_N)^{-1} Y
\end{aligned}
$$

For a new point x_*, writing $\mathbf{k}$ to be the vector with $\mathbf{k}_i = k(x_*, x_i)$ we see that
$f(x_*) = \sum_i \alpha_i k(x_*, x_i) = \mathbf{k}^\intercal (K + \lambda I_N)^{-1} Y$

Solution is the same as before

Old solution:

$$f(\mathbf{x}_*) = \phi(\mathbf{x}_*)^\intercal \Phi^\intercal (\Phi\Phi^\intercal + \lambda I_N)^{-1} Y$$

New solution:

$$f(x_*) = \mathbf{k}^\intercal (K + \lambda I_N)^{-1} Y$$

- If we look 'inside' the $\mathbf{k}$ and K, we see that these are the same.

Summary so far

- Starting with linear regression, we have derived Kernel Ridge Regression.

Summary so far

- Starting with linear regression, we have derived Kernel Ridge Regression.
- Crucial idea 1: Regulariser λ let us write all computations in terms of inner products between feature mapped observations.

Summary so far

- Starting with linear regression, we have derived Kernel Ridge Regression.
- Crucial idea 1: Regulariser λ let us write all computations in terms of inner products between feature mapped observations.
- Crucial idea 2: Representer theorem $\implies$ can project infinite dimensional optimisation problem to finite dimensional space

Summary so far

- Starting with linear regression, we have derived Kernel Ridge Regression.
- Crucial idea 1: Regulariser λ let us write all computations in terms of inner products between feature mapped observations.
- Crucial idea 2: Representer theorem $\implies$ can project infinite dimensional optimisation problem to finite dimensional space

Diferent approach?

- Motivation to use regulariser was to prevent overfitting
- Could instead adopt a Bayesian approach

GP Regression

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y} = \mathbb{R}$ and feature map $\phi : \mathcal{X} \longrightarrow \mathbb{R}^p$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ so that $f \circ \phi$ explains $\mathcal{D}$.

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y} = \mathbb{R}$ and feature map $\phi : \mathcal{X} \longrightarrow \mathbb{R}^p$
- Want to infer a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ so that $f \circ \phi$ explains $\mathcal{D}$.

An approach:

- Assume f is linear: $f(\phi(\mathbf{x})) = \phi(\mathbf{x})^\intercal \alpha$ for some α
- Place prior over α, add noise and perform Bayesian inference

$$y(\mathbf{x}) = \phi(\mathbf{x})^\intercal \mathbf{w} + \epsilon \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{w} | 0, \sigma_w^2 I), \quad \epsilon \sim \mathcal{N}(\epsilon | 0, \sigma_\epsilon^2)$$

GP Regression

$\mathbf{w}, \epsilon$ Gaussian $\implies$ y is Gaussian with

$$\mathbb{E}(y(\mathbf{x})) = \phi(\mathbf{x})^\intercal \mathbb{E}(\mathbf{w}) + \mathbb{E}(\epsilon) = 0$$
$$\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}')) = \sigma_w^2 \phi(\mathbf{x})^\intercal \phi(\mathbf{x}') + \delta_{\mathbf{x}=\mathbf{x}'} \sigma_\epsilon^2$$

GP Regression

$\mathbf{w}, \epsilon$ Gaussian $\implies$ y is Gaussian with

$$\mathbb{E}(y(\mathbf{x})) = \phi(\mathbf{x})^\intercal \mathbb{E}(\mathbf{w}) + \mathbb{E}(\epsilon) = 0$$
$$\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}')) = \sigma_w^2 \phi(\mathbf{x})^\intercal \phi(\mathbf{x}') + \delta_{\mathbf{x}=\mathbf{x}'} \sigma_\epsilon^2$$

- Write K for the matrix with $K_{ij} = \phi(\mathbf{x}_i)^\intercal \phi(\mathbf{x}_j)$
- $\mathbf{k}$ for the vector with $\mathbf{k}_i = \phi(\mathbf{x}_*)^\intercal \phi(\mathbf{x}_i)$
- $c = \phi(\mathbf{x}_*)^\intercal \phi(\mathbf{x}_*)$
- $\mathbf{y} = (y_1, \ldots, y_N, y_*)^\intercal$

GP Regression

$\mathbf{w}, \epsilon$ Gaussian $\implies$ y is Gaussian with

$$\mathbb{E}(y(\mathbf{x})) = \phi(\mathbf{x})^\intercal \mathbb{E}(\mathbf{w}) + \mathbb{E}(\epsilon) = 0$$
$$\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}')) = \sigma_w^2 \phi(\mathbf{x})^\intercal \phi(\mathbf{x}') + \delta_{\mathbf{x}=\mathbf{x}'} \sigma_\epsilon^2$$

- Write K for the matrix with $K_{ij} = \phi(\mathbf{x}_i)^\intercal \phi(\mathbf{x}_j)$
- $\mathbf{k}$ for the vector with $\mathbf{k}_i = \phi(\mathbf{x}_*)^\intercal \phi(\mathbf{x}_i)$
- $c = \phi(\mathbf{x}_*)^\intercal \phi(\mathbf{x}_*)$
- $\mathbf{y} = (y_1, \ldots, y_N, y_*)^\intercal$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y} | 0, \sigma_w^2 \begin{pmatrix} K + \frac{\sigma_\epsilon^2}{\sigma_w^2} I_N & \mathbf{k} \\ \mathbf{k}^\intercal & c + \frac{\sigma_\epsilon^2}{\sigma_w^2} \end{pmatrix})$$

GP Regression

Manipulating Gaussians shows that

$$y_*|(y_1, \ldots, y_N) \sim \mathcal{N}(y_*|\mu, \mathbf{\Sigma})$$

where

$$\mu = \mathbf{k}^\mathsf{T}(K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{y}_o$$

$$\Sigma = \sigma_w^2 c + \sigma_\epsilon^2 - \sigma_w^2 \mathbf{k}^\mathsf{T}(K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{k}$$

GP Regression

$\mu = \mathbf{k}^\intercal (K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{y}_o$

GP Regression

$\mu = \mathbf{k}^\intercal (K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{y}_o$

Some observations:

- Posterior mean depends on the ratio $\frac{\sigma_\epsilon^2}{\sigma_w^2}$

GP Regression

$\mu = \mathbf{k}^\intercal (K + \frac{\sigma_\epsilon^2}{\sigma_w^2} \mathbf{I}_N)^{-1} \mathbf{y}_o$

Some observations:

- Posterior mean depends on the ratio $\frac{\sigma_\epsilon^2}{\sigma_w^2}$
- Setting $\sigma_w^2 = 1$ and $\sigma_\epsilon^2 = \lambda$, we have KRR solution

GP Regression

$\mu = \mathbf{k}^{\mathsf{T}}(K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{y}_o$

Some observations:

- Posterior mean depends on the ratio $\frac{\sigma_\epsilon^2}{\sigma_w^2}$
- Setting $\sigma_w^2 = 1$ and $\sigma_\epsilon^2 = \lambda$, we have KRR solution
- $\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}'))$ was in terms of inner products - can replace with any kernel function.

GP Regression

$\mu = \mathbf{k}^{\intercal}(K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{y}_o$

Some observations:

- Posterior mean depends on the ratio $\frac{\sigma_\epsilon^2}{\sigma_w^2}$
- Setting $\sigma_w^2 = 1$ and $\sigma_\epsilon^2 = \lambda$, we have KRR solution
- $\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}'))$ was in terms of inner products - can replace with any kernel function.

Conclusion:

- KRR with kernel k and regularisation $\lambda \subset$ GP regression with kernel $k' = k + \lambda\delta_{x=x'}$

GP Regression

$\mu = \mathbf{k}^\intercal (K + \frac{\sigma_\epsilon^2}{\sigma_w^2}\mathbf{I}_N)^{-1}\mathbf{y}_o$

Some observations:

- Posterior mean depends on the ratio $\frac{\sigma_\epsilon^2}{\sigma_w^2}$
- Setting $\sigma_w^2 = 1$ and $\sigma_\epsilon^2 = \lambda$, we have KRR solution
- $\mathrm{Cov}(y(\mathbf{x}), y(\mathbf{x}'))$ was in terms of inner products - can replace with any kernel function.

Conclusion:

- KRR with kernel k and regularisation $\lambda \subset$ GP regression with kernel $k' = k + \lambda\delta_{x=x'}$

In fact, if we use the kernel k' for KRR without regularisation and just work through, we get the same answer[2].

[2] This is cheating really, because there is no unique optimum in this case

Frequentist Regression as MAP

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
- Want to infer a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ so that f explains $\mathcal{D}$.

Frequentist Regression as MAP

Problem:

- Given observations $\mathcal{D} = \{\mathbf{x}_i, y_i | i = 1, \ldots, N\}$ with $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
- Want to infer a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ so that f explains $\mathcal{D}$.

An approach:

- Choose some set of candidate functions $\mathcal{F}$
- Choose some *loss function* $L(f, \mathcal{D})$ to penalise misfitting the data
- Choose some *complexity penalty* $\Omega(f)$ to prevent overfitting
- Find best $f \in \mathcal{F}$ to minimise sum:

$$\arg\min_{f \in \mathcal{F}} L(f, \mathcal{D}) + \Omega(f)$$

Frequentist Regression as MAP

If $L(f, \mathcal{D}) = \sum_i L(f(x_i), y_i)$ then the problem is equivalent to

$$\arg\max_{f \in \mathcal{F}} \prod_i e^{-L(f(x_i), y_i)} e^{-\Omega(f)} \qquad (*)$$

Frequentist Regression as MAP

If $L(f, \mathcal{D}) = \sum_i L(f(x_i), y_i)$ then the problem is equivalent to

$$\arg\max_{f \in \mathcal{F}} \prod_i e^{-L(f(x_i), y_i)} e^{-\Omega(f)} \qquad (*)$$

If we can interpret

- $e^{-\Omega(f)}$ as a prior over $\mathcal{F}$
- $e^{-L(f(x_i), y_i)}$ as a likelihood

Then solving (*) is the same as performing MAP inference over $\mathcal{F}$.

Kernel Ridge Regression as MAP

In Kernel Ridge Regression, the Representer theorem allowed us to restrict ourselves from $\mathcal{F} = \mathcal{H}$ to $\mathrm{span}\{k(\cdot, x_i)\}$. We parameterise f by α, and have $\Omega(f) = \lambda \alpha^\intercal K \alpha$. So we seek

$$\begin{aligned}
&\arg\max_\alpha \prod_i e^{-(f_\alpha(x_i)-y_i)^2} e^{-\lambda\alpha^\intercal K \alpha} \\
=&\arg\max_\alpha \prod_i e^{-\frac{1}{2\lambda}(f_\alpha(x_i)-y_i)^2} e^{-\frac{1}{2}\alpha^\intercal K \alpha}
\end{aligned}$$

Kernel Ridge Regression as MAP

In Kernel Ridge Regression, the Representer theorem allowed us to restrict ourselves from $\mathcal{F} = \mathcal{H}$ to $\mathrm{span}\{k(\cdot, x_i)\}$. We parameterise f by α, and have $\Omega(f) = \lambda \alpha^\intercal K \alpha$. So we seek

$$\begin{aligned} &\arg\max_{\alpha} \prod_i e^{-(f_\alpha(x_i) - y_i)^2} e^{-\lambda \alpha^\intercal K \alpha} \\ = &\arg\max_{\alpha} \prod_i e^{-\frac{1}{2\lambda}(f_\alpha(x_i) - y_i)^2} e^{-\frac{1}{2}\alpha^\intercal K \alpha} \end{aligned}$$

This is like finding the MAP solution in the model:

$$y | \alpha, x \sim \mathcal{N}(f_\alpha(x), \lambda) \qquad\qquad \alpha \sim \mathcal{N}(0, K^{-1})$$

Kernel Ridge Regression as MAP

$$y|\alpha, x \sim \mathcal{N}(f_\alpha(x), \lambda) \qquad\qquad \alpha \sim \mathcal{N}(0, K^{-1})$$

- Prior over α is Gaussian, $f_\alpha(x) = \mathbf{k}^\intercal \alpha \implies y$ is Gaussain
- $p(\alpha|\mathcal{D})$ also Gaussian due to self-conjugacy of Gaussian
- posterior over y is Gaussian

Kernel Ridge Regression as MAP

$$y|\alpha, x \sim \mathcal{N}(f_\alpha(x), \lambda) \qquad\qquad \alpha \sim \mathcal{N}(0, K^{-1})$$

- Prior over α is Gaussian, $f_\alpha(x) = \mathbf{k}^\intercal\alpha \implies y$ is Gaussain
- $p(\alpha|\mathcal{D})$ also Gaussian due to self-conjugacy of Gaussian
- posterior over y is Gaussian

So this model is a GP, and KRR gives its MAP solution

Kernel Ridge Regression as MAP

$$y|\alpha, x \sim \mathcal{N}(f_\alpha(x), \lambda) \qquad \alpha \sim \mathcal{N}(0, K^{-1})$$

- Prior over α is Gaussian, $f_\alpha(x) = \mathbf{k}^\intercal \alpha \implies y$ is Gaussain
- $p(\alpha|\mathcal{D})$ also Gaussian due to self-conjugacy of Gaussian
- posterior over y is Gaussian

So this model is a GP, and KRR gives its MAP solution
Question: are there regression methods that are not strictly worse than GPs?

Support Vector Regression

We can do the same as Kernel Ridge Regression but with a different loss function:

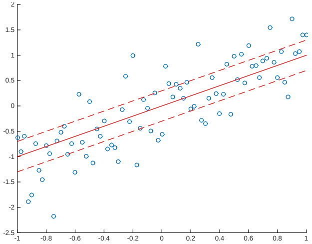
$$L(f(x), y) = \begin{cases} 0 & \text{if } |f(x) - y| < \epsilon \\ |f(x) - y| - \epsilon & \text{if } |f(x) - y| \geq \epsilon \end{cases}$$

We can do the same as Kernel Ridge Regression but with a different loss function:

$$L(f(x), y) = \begin{cases} 0 & \text{if } |f(x) - y| < \epsilon \\ |f(x) - y| - \epsilon & \text{if } |f(x) - y| \geq \epsilon \end{cases}$$

Why this might be a sensible L:

1. Robust to outliers - linear rather than quadratic loss
2. Sparse solutions - any points inside ϵ-tube around function are ignored



Support Vector Regression

Want to solve:

$$\arg\min_{f \in \mathcal{H}} \sum_i L(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

Support Vector Regression

Want to solve:

$$\arg\min_{f \in \mathcal{H}} \sum_i L(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

- Representer theorem $\implies$ solution lies in $\mathrm{span}\{k(\cdot, x_i)\}$
- Parameterise this subspace by α writing $f_\alpha(x) = \sum_i \alpha_i k(x, x_i)$

Support Vector Regression

Want to solve:

$$\arg\min_{f \in \mathcal{H}} \sum_i L(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

- Representer theorem $\implies$ solution lies in $\mathrm{span}\{k(\cdot, x_i)\}$
- Parameterise this subspace by α writing $f_\alpha(x) = \sum_i \alpha_i k(x, x_i)$
- As before, problem reduces to:

$$\arg\min_{\alpha} \sum_i L(f_\alpha(x_i), y_i) + \lambda \alpha^\intercal K \alpha$$

Support Vector Regression

This corresponds to the problem

$$\arg\max_{\alpha} \prod_i e^{-\frac{1}{2\lambda} L(f_\alpha(x_i), y_i)} e^{-\frac{1}{2}\alpha^\intercal K \alpha}$$

Equivalently, finding the MAP solution in the model:

$$p(y|\alpha, x) \propto e^{-\frac{1}{2\lambda} L(f_\alpha(x), y)} \qquad \alpha \sim \mathcal{N}(0, K^{-1})$$

Support Vector Regression

This corresponds to the problem

$$\arg\max_{\alpha} \prod_i e^{-\frac{1}{2\lambda} L(f_\alpha(x_i), y_i)} e^{-\frac{1}{2}\alpha^\intercal K \alpha}$$

Equivalently, finding the MAP solution in the model:

$$p(y|\alpha, x) \propto e^{-\frac{1}{2\lambda} L(f_\alpha(x), y)} \qquad \alpha \sim \mathcal{N}(0, K^{-1})$$

- Prior on α is Gaussian, likelihood not Gaussian
- $\implies$ y not Gaussian, posterior $p(\alpha|\mathcal{D})$ not Gaussian
- $\implies$ latent function values $f_\alpha(x) = \mathbf{k}^\intercal \alpha$ will not be Gaussian

So Support Vector Regression is distinct from Gaussian Process Regression.

Conclusion

1. Derived Kernel Ridge Regression

Conclusion

1. Derived Kernel Ridge Regression
2. KRR solution same as GP posterior mean (so KRR $\subset$ GP)

Conclusion

1. Derived Kernel Ridge Regression
2. KRR solution same as GP posterior mean (so KRR $\subset$ GP)
3. KRR is like MAP inference in GP model

Conclusion

1. Derived Kernel Ridge Regression
2. KRR solution same as GP posterior mean (so KRR $\subset$ GP)
3. KRR is like MAP inference in GP model
4. Support Vector Regression is not comparable to GP regression