Gradient Descent in Feature Space

Linear Functions in Feature Space

$$f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle$$

Empirical Risk Functional

$$R_{\text{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)$$

Regularized Risk Functional

$$R_{\text{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \frac{\lambda}{2} \|w\|^2$$

Idea

Minimize $R_{\text{reg}}[f]$ by performing gradient descent. This leads to

$$\mathbf{w} \;\rightarrow\; \mathbf{w} - \Lambda \partial_{\mathbf{w}} R_{\text{reg}}[f]$$

Gradient Descent in Feature Space

Gradient Descent Update

$$\begin{aligned}
\mathbf{w} &\rightarrow \mathbf{w} - \Lambda \partial_{\mathbf{w}} R_{\text{reg}}[f] \\
&= \mathbf{w} - \frac{\Lambda}{m} \sum_{i=1}^{m} c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i)\Phi(\mathbf{x}_i) - \Lambda\lambda\mathbf{w} \\
&= (1 - \Lambda\lambda)\mathbf{w} - \frac{\Lambda}{m} \sum_{i=1}^{m} c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i)\Phi(\mathbf{x}_i)
\end{aligned}$$

Coefficient Notation

If we represent $\mathbf{w}$ as a linear combination of $\Phi(\mathbf{x}_i)$, i.e.

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i)$$

we obtain the following update rules for α_i

$$\alpha_i \rightarrow (1 - \Lambda\lambda)\alpha_i - c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$$

Stochastic Gradient Descent

Problem

Gradient descent may be expensive, in particular if the observations are similar.

Idea

Minimize $R_{\text{reg}}[f]$ by performing stochastic gradient descent over the individual terms under the sum of

$$R_{\text{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \frac{\lambda}{2}\|w\|^2$$

Stochastic Gradient

$$\mathbf{w} \to \mathbf{w} - \Lambda c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i))\Phi(\mathbf{x}) - \Lambda\lambda\mathbf{w}$$

Coefficient Notation

We obtain the following update rule for i

$$\alpha_i \to (1 - \Lambda\lambda)\alpha_i - c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$$

for all other $j \neq i$ we have $\alpha_j \to (1 - \lambda\Lambda)\alpha_j$.

Stochastic Gradient Descent Algorithm

```
argument:   Training sample, $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathcal{X}$, $\{y_1, \ldots, y_m\} \subset \{\pm 1\}$, $\lambda, \Lambda$
returns:    Weight vector $\mathbf{w}$.
function StochasticGradientDescent($X, Y, \lambda, \Lambda$)
   initialize $\alpha_i = 0$ for all $i$
   repeat
      for all $i$ from $i = 1, \ldots, m$
         Compute $f(\mathbf{x}_i) = \left\langle \sum_{l=1}^{i} \alpha_l \Phi(x_l), \Phi(\mathbf{x}_i) \right\rangle = \sum_{l=1}^{i} \alpha_l k(\mathbf{x}_l, \mathbf{x}_i)$
         Update $\alpha_i$ according to $\alpha_i = (1 - \lambda\Lambda)\alpha_i - \Lambda c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$
         and $\alpha_j = (1 - \lambda\Lambda)\alpha_j$ (for $j \neq i$).
      endfor
   until for all $1 \leq i \leq m$ we have $|\frac{1}{m} c'(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \lambda \alpha_i| \leq \varepsilon$ end
```

So why do kernels work?

Problem

We map into a very high dimensional space. This is in conflict with the curse of dimensionality (recall the “10 observations per dimension” rule of thumb), in particular we have several orders of magnitude more dimensions than observations.

Empirical Finding

Kernels work.

Idea

Maybe, the idea of **flatness** of f, i.e. the idea of having small $\|\mathbf{w}\|$ is helping us.

We are not using all possible combinations of coefficients in $\mathbf{w}$ but rather only those for which the overall sum of coefficients is small.

The regularization constant λ specifies the trade-off between goodness of fit on the training set (the empirical error $R_{\text{emp}}[f]$) and a small value of $\|\mathbf{w}\|^2$ which is responsible for an effectively smaller class of functions.

Regularization Functionals

Goal

Quite often we would want to find a **smooth function** rather than one with small $\mathbf{w}$ in a high dimensional space which we have no real control over.

Question

What function space does $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)$ *really* correspond to?

Idea

Use a regulariztion functional $\Omega[f]$ instead, where we can specify the smoothness properties of f directly, such as

$$\Omega[f] := \frac{1}{2}\left(\|f\|^2 + \|f'\|^2\right) \text{ or } \Omega[f] := \frac{1}{2}\left(\|f\|^2 + \|f'\|^2 + \|f''\|^2\right).$$

This means that we are explicitly favouring functions with small values, flat functions, etc.

Regularization Functionals in Fourier Space

THE AUSTRALIAN NATIONAL UNIVERSITY

Idea

Maybe the regularization functional becomes easier to handle in Fourier domain.

Recall

The Fourier transform $\tilde{f}(\omega)$ of the derivative of $f(\mathbf{x})$ is given by $i\omega\tilde{f}(\omega)$. Proof:

$$\begin{aligned}
\frac{d}{d\mathbf{x}}f(\mathbf{x}) &= \frac{d}{d\mathbf{x}}(2\pi)^{-\frac{n}{2}}\int_{\mathcal{X}}\exp(i\langle\omega,\mathbf{x}\rangle)\tilde{f}(\omega)d\omega \\
&= (2\pi)^{-\frac{n}{2}}\int_{\mathcal{X}}\frac{d}{d\mathbf{x}}\exp(i\langle\omega,\mathbf{x}\rangle)\tilde{f}(\omega)d\omega \\
&= (2\pi)^{-\frac{n}{2}}\int_{\mathcal{X}}i\omega\exp(i\langle\omega,\mathbf{x}\rangle)\tilde{f}(\omega)d\omega
\end{aligned}$$

and therefore $i\omega\tilde{f} = \tilde{f}'$.

Application

$$\|f\|^2 + \|f'\|^2 = \int_{\mathcal{X}}|\tilde{f}(\omega)|^2 d\omega + \int_{\mathcal{X}}|\tilde{f}(\omega)|^2\|\omega\|^2 d\omega = \int_{\mathcal{X}}\left(\frac{1}{1+\|\omega\|^2}\right)^{-1}|\tilde{f}(\omega)|^2 d\omega$$

Regularization Functionals in Fourier Space

Application 2

$$
\begin{aligned}
& \|f\|^2 + \|f'\|^2 + c\|f''\|^2 \\
= & \int_{\mathcal{X}} |\tilde{f}(\omega)|^2 d\omega + \int_{\mathcal{X}} |\tilde{f}(\omega)|^2 \|\omega\|^2 d\omega + c \int_{\mathcal{X}} |\tilde{f}(\omega)|^2 \|\omega\|^4 d\omega \\
= & \int_{\mathcal{X}} \left(\frac{1}{1 + \|\omega\|^2 + c\|\omega\|^4} \right)^{-1} |\tilde{f}(\omega)|^2 d\omega
\end{aligned}
$$

Idea

We can extend this formalism such that it includes more general positive, symmetric functions $Q(\omega)$ with $Q(\omega) \to 0$ for $\omega \to \infty$. It leads to

$$\Omega[f] := \int_{\mathcal{X}} \frac{|\tilde{f}(\omega)|^2}{Q(\omega)} d\omega$$

In this case $Q(\omega)$ describes the desired frequency properties of f.

Regularization Functionals and Dot Products

Idea

The functional $\Omega[f]$ in Fourier space also defines a dot product between functions. It is given by

$$\langle f, g\rangle_{\mathcal{H}} := \int_{\mathcal{X}} \frac{\overline{\tilde{f}(\omega)}g(\omega)}{Q(\omega)} d\omega$$

We can check that $\langle f, f\rangle_{\mathcal{H}}$ leads to the original form of $\Omega[f]$.

Kernels and Feature Spaces

Maybe we can find a kernel function k such that for the map

$$\mathbf{x} \to k(\mathbf{x}, \cdot)$$

we have at the same time

$$\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot)\rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}').$$

In this case the map $\mathbf{x} \to k(\mathbf{x}, \cdot)$ would be the **feature map** and the Hilbert space given by $\langle, \cdot, \cdot\rangle_{\mathcal{H}}$ the corresponding **feature space**.

Regularization and Fourier Transforms

Wild Guess

Given a regularization functional via

$$\Omega[f] := \int_{\mathcal{X}} \frac{|\tilde{f}(\omega)|^2}{Q(\omega)} d\omega$$

we posit (of course it is true and we will prove it on the next slide) that the kernel $k(\mathbf{x} - \mathbf{x}')$, for which $Q(\omega) = \tilde{k}(\omega)$, satisfies the condition

$$\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot) \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}').$$

Interpretation

This means that the Fourier transform of the kernel tells us how smooth the function class is that we are approximating.

It also allows us to construct a kernel corresponding to the regularization terms, such as $\|f\|^2 + \|f'\|^2$, described before, simply by computing the Fourier representation of the regularization term first and subsequently performing a Fourier transform back.

Proof

Computing the Dot Product

Translations in the $\mathbf{x}$ domain correspond to multiplications in the Fourier domain, i.e.

$$g(\mathbf{x}') := k(\mathbf{x}-\mathbf{x}') \Longrightarrow \tilde{g}(\mathbf{x}) = \tilde{k}(\omega)\exp(-i\langle\omega,\mathbf{x}'\rangle)$$

For $\tilde{k}(\omega) = Q(\omega)$ we have

$$\begin{aligned}
\langle k(\mathbf{x},\cdot), k(\mathbf{x}',\cdot)\rangle_{\mathcal{H}} &= \int_{\mathcal{X}} \frac{\overline{\tilde{k}(\omega)\exp(-i\langle\omega,\mathbf{x}\rangle)}\tilde{k}(\omega)\exp(-i\langle\omega,\mathbf{x}'\rangle)}{Q(\omega)} d\omega \\
&= \int_{\mathcal{X}} \frac{Q^2(\omega)}{Q(\omega)} \exp(i\langle\omega,\mathbf{x}-\mathbf{x}'\rangle) d\Omega \\
&= \int_{\mathcal{X}} Q(\omega)\exp(i\langle\omega,\mathbf{x}-\mathbf{x}'\rangle) d\Omega = k(\mathbf{x}-\mathbf{x}')
\end{aligned}$$

The space given by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is often also referred to as the **Reproducing Kernel Hilbert Space**.

Application to Dot Product Kernels

Gaussian RBF Kernel

For a kernel of type

$$k(x, x') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}|x - x'|)$$

the Fourier transform is given by

$$\tilde{k}(\omega) = \sqrt{\frac{\sigma^2}{2\pi}} \exp(-\frac{\sigma^2}{2}|x - x'|)$$

The narrower we choose the kernels in input space, the wider their Fourier transform becomes. Consequently we accept more high frequency components.

Laplacian Kernel

For $k(x, x') = \exp(-|x - x'|)$ we have $\tilde{k}(\omega) = \sqrt{\frac{2}{\pi}} \frac{1}{1+\omega^2}$. This means that the Laplacian kernel is less smooth than the Gaussian one.

Application to Regularization Terms

Operator Representation

If $\Omega[f]$ is represented as a sum of norms of derivatives, i.e.

$$\Omega[f] = \sum_j c_j \left\| \left(\frac{d}{dx} \right)^j f \right\|^2$$

we can find a corresponding Fourier representation where

$$\frac{1}{Q(\omega)} = \sum_i c_j \|\omega\|^{2j}.$$

Therefore we can obtain the kernel corresponding to $Q[\omega]$ by computing the Fourier transform of Q.

Laplacian Kernel

We already showed that the Fourier transform of $e^{-|x|}$ is $\frac{1}{1+\omega^2}$. From this we immediately read off the regularization term as $c_0 = 1$ and $c_1 = 1$, i.e.

$$\Omega[f] = \|f\|^2 + \|f'\|^2.$$