Recall – Support Vector Regression

- We want to fit a function

$$g(X,\ W) = W^T \Phi(X)\ +\ b$$

- We do empirical risk minimization with ϵ-insensitive loss:

$$\begin{array}{rcll} L_\epsilon(y_i,\ g(X_i,\ W)) & = & 0 & \text{If } \ |y_i - g(X_i,\ W)| < \epsilon \\ & = & |y_i - g(X_i,\ W)| - \epsilon & \text{otherwise} \end{array}$$

- We use $W^T W$ as a regularization term

Recall – SVR Optimization Problem

- Find W, b and ξ_i, ξ_i' to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}W^TW + C\left(\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \xi_i'\right) \\ \text{subject to} & y_i - W^T\Phi(X_i) - b \leq \epsilon + \xi_i, \;\; i = 1, \ldots, n \\ & W^T\Phi(X_i) + b - y_i \leq \epsilon + \xi_i', \;\; i = 1, \ldots, n \\ & \xi_i \geq 0, \; \xi_i' \geq 0 \;\; i = 1, \ldots, n \end{array}$$

- Has similar structure as the SVM.

The dual

- The dual of this problem is

$$
\begin{aligned}
\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i') - \epsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i') \\
& -\frac{1}{2} \sum_{i,j} (\alpha_i - \alpha_i')(\alpha_j - \alpha_j')\Phi(X_i)^T \Phi(X_j) \\
\text{subject to} \quad & \sum_{i=1}^{n} (\alpha_i - \alpha_i') = 0 \\
& 0 \leq \alpha_i,\ \alpha_i' \leq C, \quad i = 1, \ldots, n
\end{aligned}
$$

- Here α_i and α_i' are the Lagrange multipliers corresponding to the first two inequalities in the primal.

The solution

- The solution is

$$
\begin{aligned}
W^* &= \sum_{i=1}^{n} (\alpha_i^* - \alpha_i^{*'})\Phi(X_i) \\
b^* &= y_j - \Phi(X_j)^T W^* + \epsilon, \;\; j \; s.t. \; 0 < \alpha_j^* < C/n
\end{aligned}
$$

- Let $K(X, X') = \Phi(X)^T \Phi(X')$.
- The optimal model learnt is

$$
\begin{aligned}
g(X,\; W^*) &= X^T W^* + b^* \\
&= \sum_{i=1}^{n} (\alpha_i^* - \alpha_i^{*'})\phi(X_i)^T \phi(X) + b^* \\
&= \sum_{i=1}^{n} (\alpha_i^* - \alpha_i^{*'}) K(X_i,\; X) + b^*
\end{aligned}
$$

(Note that b^* can also be written in terms of the Kernel function.)

Support vector regression

- Once again, the kernel trick allows us to learn non-linear models using a linear method.
- The parameters: C, ϵ and parameters of kernel function.
- The basic idea of SVR can be used in many related problems.

SV regression

- With the ϵ-insensitive loss function, points whose targets are within ϵ of the prediction do not contribute any ‘loss’.
- Gives rise to some interesting robustness of the method. It can be proved that local movements of target values of points outside the ϵ-tube do not influence the regression.
- Robustness essentially comes through the support vector representation of the regression.

- In our formulation of the regression problem we added W^TW term in the objective function.
- We are essentially minimizing

$$\frac{1}{2}W^TW \ + \ C \sum_{i=1}^{n} \max\left(|y_i - \Phi(X_i)^TW - b| - \epsilon, \ 0\right)$$

- This is 'regularized risk minimization'.
- Then W^TW is the model complexity term which is intended to favour learning of 'smoother' models.
- There are several ways to understand why W^TW is a good term to caracterize smoothness in case of linear models.

- Let $f : \Re^m \rightarrow \Re$ be a continuous function.
- Continuity means we can make $|f(X) - f(X')|$ as small as we want by taking $||X - X'||$ sufficiently small.
- There are ways to characterize the 'degree of continuity' of a function.
- We consider one such measure now.

ϵ-Margin of a function

- The ϵ-margin of a function, $f : \Re^n \rightarrow \Re$ is

$$m_\epsilon(f) = \inf\{||X - X'|| \; : \; |f(X) - f(X')| \geq 2\epsilon\}$$

- The intuitive idea is:
  How small can $||X - X'||$ be, still keeping $|f(X) - f(X')|$ 'large'
- The larger $m_\epsilon(f)$, the smoother is the function.

$$m_\epsilon(f) = \inf\{||X - X'|| \, : \, |f(X) - f(X')| \geq 2\epsilon\}$$

- Obviously, $m_\epsilon(f) = 0$ if f is discontinuous.
- $m_\epsilon(f)$ can be zero even for continuous functions,

  e.g., $f(x) = 1/x$.

- $m_\epsilon(f) > 0$ for all $\epsilon > 0$ iff f is uniformly continuous.
- Higher margin would mean the function is 'slowly varying' and hence is a 'smoother' model.

Linear Models and margin

- Consider regression with linear models. Then,

$$|f(X) - f(X')| = |W^T(X - X')|.$$

- For all X, X' with $|W^T(X - X')| \geq 2\epsilon$, we want the smallest $||X - X'||$
- It would be smallest if $|W^T(X - X')| = 2\epsilon$ and $(X - X')$ is parallel to W.

  That is, $X - X' = \pm \frac{2\epsilon W}{W^T W}$.
- Thus, $m_\epsilon(f) = || \pm \frac{2\epsilon W}{W^T W}|| = \frac{2\epsilon}{||W||}$.
- Thus in our optimization problem adding the term $W^T W$ promotes learning of smoother models.
- As we have seen linear regression models use this as the regularization term.

- The basic idea of kernel functions, as we saw in SVM, has been extended in many ways.
- There have been many extensions of the basic SVM method also.
- Some of them are essentially formulations of approximate solutions to make the algorithm more efficient.
- Some of them are reformulations to add additional features to the SVM method.
- We consider a couple of simple examples of such extensions.

- Suppose the optimization problem is changed to

$$\begin{aligned} \min_{W,b,\boldsymbol{\xi}} \quad & \frac{1}{2}W^T W \; + \; b^2 \; + \; C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(W^T X_i + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, n \\ & \xi_i \geq 0, \;\; i = 1, \ldots, n \end{aligned}$$

- We have added the b^2 term to the objective function. The main reason is that it simplifies the dual.

- The dual turns out to be

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \sum_{i=1}^{n} \mu_i - \frac{1}{2}\sum_{i,j=1}^{n} \mu_i\mu_j y_i y_j K(X_i, X_j) \\ & \qquad - \frac{1}{2}\sum_{i,j=1}^{n} \mu_i\mu_j y_i y_j \\ \text{subject to} \quad & 0 \le \mu_i \le C, \;\; i = 1, \dots, n, \end{aligned}$$

- The equality constraint is absent.
  Only bound constraints on variables.
- Allows for efficient optimization.
  (Successive overrelaxation).

- Next, we consider a reformulation of SVM optimization problem, known as ν-SVM.
- Recall that the primal problem for SVM with slack variables is

$$\begin{array}{ll} \min\limits_{W,b,\boldsymbol{\xi}} & \frac{1}{2}W^TW \; + \; C\sum\limits_{i=1}^{n} \xi_i \\ \text{subject to} & y_i(W^T\phi(X_i)+b) \geq 1-\xi_i, \;\; i=1,\ldots,n \\ & \xi_i \geq 0, \;\; i=1,\ldots,n \end{array}$$

- We will call this as C-SVM.
- In the C-SVM, one has no intuition for choosing the value of C.

ν-SVM

- Consider a changed optimization problem

$$\begin{aligned} \min_{W,b,\boldsymbol{\xi},\rho} \quad & \frac{1}{2}W^TW - \nu\rho + \frac{1}{n}\sum \xi_i \\ \text{subject to} \quad & y_i[W^T\phi(X_i)+b] \geq \rho - \xi_i \\ & \xi_i \geq 0. \end{aligned}$$

  where ν is a user-chosen constant.
- Note that $W, b, \rho, \xi_i = 0$ is a feasible solution.
- We do not need $\rho \geq 0$ constraint.

- The Lagrangian for this problem is

$$
\begin{aligned}
L(W, b, \boldsymbol{\xi}, \rho, \boldsymbol{\eta}, \boldsymbol{\mu}) &= \frac{1}{2}W^T W - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\
&\quad - \sum_{i=1}^{n} \eta_i\xi_i + \sum_{i=1}^{n} \mu_i \left(\rho - \xi_i - y_i[W^T\phi(X_i) + b]\right)
\end{aligned}
$$

- The μ_i are the Lagrange multipliers for the separability constraints and η_i are the Lagrange multipliers for the constraints $\xi_i \geq 0$.

$$L = \frac{1}{2}W^TW - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\eta_i\xi_i + \sum_{i=1}^{n}\mu_i(\rho - \xi_i - y_i[W^T\phi(X_i) + b]$$

The Kuhn-Tucker conditions give us

- $\nabla_W L = 0 \Rightarrow W = \sum_i \ \mu_i y_i \phi(X_i)$
- $\frac{\partial L}{\partial b} = 0 \Rightarrow \sum \mu_i y_i = 0$
- $\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \mu_i + \eta_i = \frac{1}{n}, \ \forall i$
- $\frac{\partial L}{\partial \rho} = 0 \Rightarrow \sum \mu_i = \nu$
- $\rho - \xi_i - y_i(W^T\phi(X_i) + b) \leq 0; \ \ \xi_i \geq 0; \ \forall i$
- $\mu_i \geq 0; \ \eta_i \geq 0, \ \forall i$
- $\mu_i(\rho - \xi_i - y_i(W^T\phi(X_i) + b)) = 0; \ \ \eta_i\xi_i = 0, \ \forall i$

- Suppose $\xi_i > 0$ for some i. Then we have $\eta_i = 0$ and hence $\mu_i = \frac{1}{n}$. Hence

$$\begin{aligned} \nu &= \sum_{i=1}^{n} \mu_i = \sum_{i\,:\,\xi_i>0} \mu_i + \sum_{i\,:\,\xi_i=0} \mu_i \\ &\geq \sum_{i\,:\,\xi_i>0} \mu_i = \frac{|\{i : \xi_i > 0\}|}{n} \end{aligned}$$

- Hence we have:
  ν is an upper bound on the fraction of ‘margin errors’.

- We also have, because $0 \leq \mu_i \leq \frac{1}{n}$,

$$\begin{aligned} \nu &= \sum_{i=1}^{n} \mu_i = \sum_{i\,:\,\mu_i>0} \mu_i + \sum_{i\,:\,\mu_i=0} \mu_i \\ &\leq \sum_{i\,:\,\mu_i>0} \mu_i \leq \frac{|\{i : \mu_i > 0\}|}{n} \end{aligned}$$

- Hence we have:
  ν is a lower bound on the fraction of support vectors.

- In the ν-SVM formulation, the ν is the user chosen constant.
- Unlike the parameter C, the ν has an interesting interpretation.
- It is simultaneously the upperbound on fraction of errors and lower bound on fraction of support vectors.
- If for the chosen ν, the problem has a solution with $\rho > 0$, then the bounds would be met.
- This gives us a good way to choose this 'penalty constant'.

- The dual for the ν-SVM turns out to be

$$\max_{\boldsymbol{\mu}} \qquad q(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i,j=1}^{n} \mu_i\mu_j y_i y_j K(X_i, X_j)$$

$$\text{subject to} \qquad 0 \leq \mu_i \leq \frac{1}{n}, \forall i;\ \ \sum_{i=1}^{n} y_i\mu_i = 0;\ \ \sum_{i=1}^{n}\mu_i = \nu$$

- This a simple optimization problem similar to that of 'C-SVM'.
- One can show that if we have a solution for ν-SVM then if we choose $C = 1/\rho n$, we get the same solution with 'C-SVM'.

ν SVR

- This idea can be extended to the regression problem also.
- In support vector regression, we had two user defined constants: ϵ and C.
- The ϵ specifies the 'tolerable error' and it is difficult to know what value to choose for it.
- We can reformulate SVR so that we can optimize on ϵ also.
- This will be very similar to the ν-SVM formulation.

- Recall the optimization problem in SVR:

$$
\begin{array}{ll}
\min\limits_{W,b,\boldsymbol{\xi},\boldsymbol{\xi'}} & \frac{1}{2}W^TW + C\left(\sum\limits_{i=1}^{n}\xi_i + \sum\limits_{i=1}^{n}\xi_i'\right) \\
\text{subject to} & y_i - W^T\Phi(X_i) - b \leq \epsilon + \xi_i, \;\; i = 1, \dots, n \\
 & W^T\Phi(X_i) + b - y_i \leq \epsilon + \xi_i', \;\; i = 1, \dots, n \\
 & \xi_i \geq 0, \; \xi_i' \geq 0 \;\; i = 1, \dots, n
\end{array}
$$

- We change the optimization problem to the following:

$$\begin{array}{ll}
\min\limits_{W,b,\epsilon,\boldsymbol{\xi},\boldsymbol{\xi'}} & \frac{1}{2}W^TW + C\left(\nu\epsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i')\right) \\
\text{subject to} & y_i - W^T\phi(X_i) - b \leq \epsilon + \xi_i, \;\; i = 1, \ldots, n \\
 & W^T\phi(X_i) + b - y_i \leq \epsilon + \xi_i', \;\; i = 1, \ldots, n \\
 & \xi_i \geq 0,\; \xi_i' \geq 0 \; \epsilon \geq 0,\; i = 1, \ldots, n
\end{array}$$

  where ν is a user-chosen constant.
- We get similar results as in ν-SVM.

Risk minimization view of SVM

- We posed the support vector regression problem as a (regularized) risk minimization under a special loss function.
- It was then reformulated into an (equivalent) constrained optimization problem.
- In contrast, we formulated the SVM directly as a constrained optimization problem.
- However, it can also be seen to be minimization of (regularized) empirical risk under a special loss function.

- The optimization problem for SVM is

$$\begin{aligned} \min_{W,b,\boldsymbol{\xi}} \quad & \frac{1}{2}W^TW \; + \; C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(W^TX_i + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, n \\ & \xi_i \geq 0, \;\; i = 1, \ldots, n \end{aligned}$$

- Given any W, b, the ξ_i have to satisfy

$$\xi_i \geq \max(0,\; 1 - y_i(W^TX_i + b))$$

- Since we need to minimize $\sum \xi_i$, we need to take the value above for each ξ_i.

- Hence we can find SVM by solving the following unconstrained optimization problem:

$$\min_{W,b} \ \frac{1}{2}W^TW \ + \ C\sum_{i=1}^{n} \max(0,\ 1 - y_i(W^TX_i + b))$$

- Consider the loss function defined by

$$L_{\textsf{hinge}}(y, f(X)) = \max(0, 1 - yf(X))$$

- Then the optimization problem is same as

$$\min_{W,b} \ \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(X_i)) \ + \ C'\frac{1}{2}W^TW$$

- Then the optimization problem is same as

$$\min_{W,b} \; \frac{1}{n} \sum_{i=1}^{n} \, L(y_i, f(X_i)) \; + \; C' \frac{1}{2} W^T W$$

- The model (or classifier) we are learning is $f(X) = W^T X + b$.
- For this model, we already saw $W^T W$ is a good regularization term.
- Thus, our SVM formulation is empirical risk minimization under hinge-loss along with a regularization term.

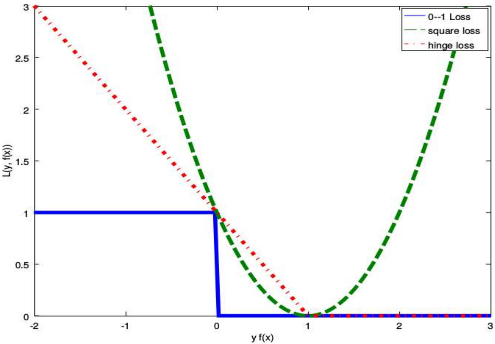
- As we saw earlier, the hinge-loss and square-loss are good convex approximations of the 0–1 loss.
- For 0–1 loss $L(y,\ f(X))$ is one if $yf(X)$ is negative and zero otherwise.
- The squared error loss can be written as

$$L_{\textsf{square}}(y,\ h(X)) = (1 - yf(X))^2$$

- The hinge loss is given by

$$L_{\textsf{hinge}}(y, h(X)) = \max(0,\ 1 - yf(X))$$

- We can plot all the functions as follows.



(Here we plot $yf(X)$ on x-axis and $L(y, f(X))$ on y-axis).

- Hinge loss is also called soft-margin loss.
- Supoose we want to minimize, over all f,

$$E[\max(0,\ 1 - yf(X))],\quad y \in \{+1,\ -1\}$$

- Intuitively the best we can do is to make sign of $f(X)$ to be same as sign of the corresponding y.
- Hence, intuitively, the best f is

$$f(X) > 0,\quad \text{if}\ \ P[y = +1|X] > 0.5;\quad \text{else}\ \ f(X) < 0$$

- This is indeed a good classifier.

- In SVM method, there are two important ingradients.
- One is the Kernel function.
- Kernel functions allow us to learn nonlinear models using essentially linear techniques.
- Second is the ‘support vector’ expansion – the final model is expressed as a (‘sparse’) linear combination of some of the data vectors.
- Kernels are a good way to capture ‘similarity’ and are useful in general.
- The support vector expansion is also a general property of Kernel based methods.
- We look at this general view of Kernels next.

- Often, in pattern recognition, we use distance between pattern vectors as a means to assess similarity (e.g. nearest neighbour classifier).
- Kernels allow us to generalize such notions of distance or similarity between patterns.

- Consider a 2-class classification problem with training data

$$\{(X_i, y_i),\ i = 1, \cdots, n\},\ X_i \in \Re^m,\ y_i \in \{+1,\ -1\}$$

- Suppose we implement a nearest neighbour classifier, by computing distance of a new pattern to a set of prototypes.
- Keeping with the viewpoint of SVM, suppose we want to transform the patterns into a new space using ϕ and find the distances there.

- Suppose we use two prototypes given by

$$C_+ = \frac{1}{n_+} \sum_{i:\, y_i = +1} \phi(X_i) \quad \text{and} \quad C_- = \frac{1}{n_-} \sum_{i:\, y_i = -1} \phi(X_i)$$

  where n_+ is the number of examples in class $+1$ and n_- is that in class -1.
- The prototypes are 'centers' of the two classes.
- Given a new X, we would put it in class $+1$ if

$$||\phi(X) - C_+||^2 \, < \, ||\phi(X) - C_-||^2$$

- We can implement this using kernels. We have

$$||\phi(X) - C_+||^2 = \phi(X)^T\phi(X) - 2\phi(X)^T C_+ + C_+^T C_+$$

- Thus we would put X in class $+1$ if

$$\phi(X)^T C_+ - \phi(X)^T C_- + \frac{1}{2}\left(C_-^T C_- - C_+^T C_+\right) > 0$$

- All these inner products are now easily done using kernel functions.

- By the definition of C_+, we get

$$\begin{aligned}\phi(X)^T C_+ &= \phi(X)^T \left(\frac{1}{n_+} \sum_{i:\, y_i=+1} \phi(X_i) \right) \\ &= \frac{1}{n_+} \sum_{i:\, y_i=+1} K(X_i, X)\end{aligned}$$

- Similarly we get

$$C_+^T C_+ = \frac{1}{n_+^2} \sum_{i,j\ : y_i = y_j = +1} K(X_i, X_j)$$

- Thus, our classifier is $sgn(h(X))$ where

$$h(X) = \frac{1}{n_+} \sum_{i:\, y_i=+1} K(X_i, X) - \frac{1}{n_-} \sum_{i:\, y_i=-1} K(X_i, X) + b$$

where

$$b = \frac{1}{2} \left(\frac{1}{n_-^2} \sum_{y_i, y_j=-1} K(X_i, X_j) - \frac{1}{n_+^2} \sum_{y_i, y_j=+1} K(X_i, X_j) \right)$$

- Thus we can implement such nearest neighbour classifiers by implicitly transforming the feature space and using kernel function for the inner product in the transformed space.
- The kernel function allows us to formulate the right kind of similarity measure in the original space.

- Define

$$\begin{aligned} P_+(X) &= \frac{1}{n_+} \sum_{i:\, y_i=+1} K(X_i, X), \\ P_-(X) &= \frac{1}{n_-} \sum_{i:\, y_i=-1} K(X_i, X) \end{aligned}$$

- With a proper normalization, these are essentially non-parametric estimators for the class conditional densities – the kernel density estimates.

- We could, for example, use a Gaussian kernel and then it is the nonparametric density estimators we studied earlier.
- Thus, our nearest neighbour classifier is essentially a Bayes classifier using nonparametric estimators for class conditional densities.

- We next look at positive definite kernels in some detail.
- We show that for any such kernel, there is one vector space with an innerproduct such that the kernel realizes an innerproduct in that space.
- This is called the Reproducing Kernel Hilbert Space (RKHS) associated with the Kernel.
- We also show that if we are doing regularized empirical risk minimization on this space, then the final solution would have the 'support vector expansion' form.

Positive definite kernels

- Let $\mathcal{X}$ be the original feature space.
- Let $K : \mathcal{X} \times \mathcal{X} \to \Re$ be a positive definite kernel.
- Given any n points, $X_1, \cdots, X_n \in \mathcal{X}$, the $n \times n$ matrix with $(i,\ j)$ element as $K(X_i, X_j)$ is called the Gram matrix of K.
- Recall that K is positive definite if the Gram matrix is positive semi-definite for all n and all $X_1, \cdots, X_n$.

- Positive definiteness of Kernel means, for all n,

$$\sum_{i,j=1}^{n} c_i c_j K(X_i, X_j) \geq 0, \;\; \forall c_i \in \Re, \, \forall X_i \in \mathcal{X}$$

- Taking $n = 1$, we get $K(X, X) \geq 0, \; \forall X \in \mathcal{X}$.
- Taking $n = 2$ and remembering that K is symmetric, we get

$$K(X_1, X_2)^2 \leq K(X_1, X_1)\, K(X_2, X_2), \;\; \forall X_1, X_2 \in \mathcal{X}$$

Thus, K satisfies Cauchy-Schwartz inequality

- Suppose $K(X, X') = \phi(X)^T\phi(X')$.
- Then K is a positive definite kernel:

$$\begin{aligned}\sum_{i,j} c_i c_j \phi(X_i)^T\phi(X_j) &= \left(\sum_i c_i\phi(X_i)\right)^T \left(\sum_j c_j\phi(X_j)\right) \\ &= \Big|\Big|\sum_i c_i\phi(X_i)\Big|\Big|^2 \geq 0\end{aligned}$$

- Thus, e.g., if K satisfies Mercer theorem, then it is a positive definite kernel.

- We now show that all positive definite kernels are also innerproducts on some appropriate space.
- Given a kernel K, we will construct a space endowed with an inner product and show how any positive definite kernel is essentially implementing inner product in this space.
- This space is called the Reproducing Kernel Hilbert Space associated with the Kernel, K.

- Let $\Re^{\mathcal{X}}$ be the set of all real-valued functions on $\mathcal{X}$.
- Let K be a positive definite kernel.
- For any $X \in \mathcal{X}$, let $K(\cdot\,, X) \in \Re^{\mathcal{X}}$ denote the function that maps $X' \in \mathcal{X}$ to $K(X', X) \in \Re$.
- That is, $K(\cdot\,, X)(X') = K(X, X')$.
  If the notation is confusing, think of $K(\cdot\,, X)$ as a function $g_X(\cdot)$ with $g_X(X') = K(X, X'), \forall X' \in \mathcal{X}$.
- Consider the set of functions

  $\mathcal{H}_1 = \{K(\cdot\,, X) \,:\, X \in \mathcal{X}\}$.
- Let $\mathcal{H}$ be the set of all functions that are finite linear combinations of functions in $\mathcal{H}_1$.

- Note that elements of $\mathcal{H}$ are certain real-valued functions on $\mathcal{X}$
- Any $f(\cdot) \in \mathcal{H}$ can be written as

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i \, K(\cdot \, , X_i), \quad \text{for some } \; n, X_i \in \mathcal{X}, \; \alpha_i \in \Re$$

- It is easy to see that if $f, g \in \mathcal{H}$ then $f + g \in \mathcal{H}$ and $\alpha f \in \mathcal{H}$ for $\alpha \in \Re$.
- Thus, $\mathcal{H}$ is a vector space. (The scalars are reals)
- We now define an inner product on $\mathcal{H}$.

- Let $f, g \in \mathcal{H}$ with

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i \, K(\cdot\,, X_i), \quad g(\cdot) = \sum_{j=1}^{n'} \beta_j \, K(\cdot\,, X_j')$$

- We define the inner product as

$$< f \ , \ g > \ = \ \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j \, K(X_i, X_j')$$

- We first show this is well defined.
- That is, we show that the inner product does not depend on the specific representation used for f and g.

- Note that

$$< f\ ,\ g > \ = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n'} \beta_j \, K(X_i, X_j') = \sum_{i=1}^{n} \alpha_i \, g(X_i)$$

  Thus the innerproduct does not depend on β_j or X_j'.
- Similarly we have

$$< f\ ,\ g > \ = \sum_{j=1}^{n'} \beta_j \sum_{i=1}^{n} \alpha_i \, K(X_i, X_j') = \sum_{j=1}^{n'} \beta_j \, f(X_j')$$

- Thus our inner product does not depend on the α_i, β_j or the specific representation used and hence is well defined.

$< f, g >$ is an Inner Product

$$< f \, , \, g > \; = \; \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j \; K(X_i, X'_j)$$

- By definition, $< f, g > = < g, f >$. (Symmetric)
- It is easily verified that it is bilinear:

$$< f \, , \, g_1 + g_2 > = < f \, , \, g_1 > + < f \, , \, g_2 >$$

$$< f_1 + f_2 \, , \, g > = < f_1 \, , \, g > + < f_2 \, , \, g >$$

- It is also easy to see that $< cf \, , \, g > = \; c \; < f \, , \, g >$.
- We have, by the positive definiteness of K, $< f \, , \, f > = \; \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(X_i, X_j) \; \geq \; 0$

- Finally, we have to show $< f \, , \, f > = 0 \;\; \Rightarrow \;\; f = 0$.
- Let $f_1, \cdots, f_p \in \mathcal{H}$ and let $\gamma_1, \cdots, \gamma_p \in \Re$.
- Let $g_1 = \sum_{i=1}^{p} \gamma_i f_i \;\; \in \mathcal{H}$.
- Now we get

$$\begin{aligned} \sum_{i,j=1}^{p} \gamma_i \gamma_j < f_i \, , \, f_j > &= \; < \sum_{i=1}^{p} \gamma_i f_i \, , \, \sum_{j=1}^{p} \gamma_j f_j > \\ &= \; < g_1 \, , \, g_1 > \\ &\geq \; 0 \end{aligned}$$

for any scalaras γ_i and any f_i and any p.

- Note that $< \cdot\,, \cdot >$ is a symmetric function that maps $\mathcal{H} \times \mathcal{H}$ to $\Re$.
- Thus what we have shown is that $< \cdot\,, \cdot >$ is a positive definite kernel on $\mathcal{H}$
- Since positive definite kernels satisfy Cauchy-Schwartz inequality, we have

$$|<g_1, g_2>|^2 \quad \leq \quad <g_1, g_1><g_2, g_2>$$

- In particular, for any $f \in \mathcal{H}$, we have

$$|<K(\cdot, X), f>|^2 \leq <K(\cdot, X), K(\cdot, X)><f, f>$$

for all $X \in \mathcal{X}$.

- Recall

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i \, K(\cdot\, , X_i), \;\; g(\cdot) = \sum_{j=1}^{n'} \beta_j \, K(\cdot\, , X'_j)$$

$$< f \, , \, g > \; = \; \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j \, K(X_i, X'_j)$$

- Hence We have

$$\begin{aligned} < K(\cdot\, , X) \, , \, K(\cdot\, , X') > \quad &= \quad K(X, X') \quad \text{and} \\ < K(\cdot\, , X) \, , \, f > \quad &= \quad \sum_{i=1}^{n} \alpha_i \, K(X \, , X_i) = f(X) \end{aligned}$$

- This is called the reproducing Kernel property.

- We have

$$\begin{aligned} < K(\cdot\,, X)\,,\; K(\cdot\,, X') > \quad &= \quad K(X, X') \quad \text{and} \\ < K(\cdot\,, X)\,,\; f > \quad &= \quad f(X) \end{aligned}$$

- Now, $\forall X$,

$$|f(X)|^2 = |< K(\cdot\,, X)\,,\; f >|^2 \;\leq\; K(X, X)\; < f\,,\; f >$$

- This shows $< f\,,\; f > \; = 0 \;\; \Rightarrow \;\; f = 0$.
- This shows what we defined is indeed an inner product.

- Given any positive definite kernel, we can construct this inner product space $\mathcal{H}$ as explained here.
- We can complete it in the norm induced by the inner product.
- It is called the Reproducing Kernel Hilbert Space (RKHS) associated with K.
- The reproducing kernel property is

$$< K(\cdot\, , X)\, ,\; f > \; = f(X),\; \forall f \in \mathcal{H}$$

- Note that elements of RKHS are certain real-valued functions on $\mathcal{X}$. Essentially, a kind of generalization of linear functionals on $\mathcal{X}$.

- Given this RKHS $\mathcal{H}$ associated with K, define $\phi : \mathcal{X} \to \mathcal{H}$ by

$$\phi(X) = K(\cdot\,, X)$$

- Now we have

$$K(X, X') = \; < \phi(X)\,,\; \phi(X') >$$

- This shows that any positive definite kernel gives us the inner product in some other space as needed.

- As a simple example, let $\mathcal{X} = \Re^m$ and $K(X, X') = X^T X'$.
- Now, $K(\cdot\,, X)$ is the function that takes dot product of its argument with X.
- Let $X = [x_1, \, \cdots \, , \, x_m]^T$. Let e_i, $i = 1, \cdots, m$, be the coordinate unit vectors. Thus, $X = \sum_i x_i e_i$.
- For any $X' \in \Re^m$,

$$K(X', X) = X^T X' = \sum_{i=1}^{m} x_i e_i^T X' = \sum_{i=1}^{m} x_i K(X', e_i)$$

- This gives us $K(\cdot\,, X) = \sum_{i=1}^{m} x_i K(\cdot\,, e_i)$.

- This means all functions in $\mathcal{H}$ are linear combinations of $K(\cdot\,, e_i)$.
- Thus any $f \in \mathcal{H}$ can be written as $f = \sum_{i=1}^{m} w_i K(\cdot\,, e_i)$.
- Each $f \in \mathcal{H}$ is characterized by $W = (w_1, \cdots, w_m)^T$.
- We have, $f(X) = W^T X$. So, $\mathcal{H}$ is the space of all linear functionals over $\Re^m$.
- So, the RKHS is isomorphic to $\Re^m$ and it represents hyperplanes on $\mathcal{X}$.

- We can see the reproducing kernel property: Let

$$f = \sum_{j=1}^{m} w_j K(\cdot\, , e_j) \;\; \Rightarrow \;\; f(X) = \sum_{i=1}^{m} w_i x_i$$

- We have $K(\cdot\, , X) = \sum_{i=1}^{m} x_i K(\cdot\, , e_i)$
- Hence we get

$$< K(\cdot\, , X)\, ,\; f >= \sum_{i,j} x_i w_j K(e_i, e_j) = \sum_{i} x_i w_i = f(X)$$

- So, the RKHS is isomorphic to $\Re^m$ and thus it represents hyperplanes on $\mathcal{X}$.
- The inner product in this $\mathcal{H}$ would be simply the usual dot product.
- Learning hyperplanes is same as searching over this $\mathcal{H}$ for minimizer of empirical risk with the usual norm as a regularizer.

- What we have shown is the following.
- Given a positive definite kernel, there is a vector space with an inner product, namely, the RKHS associated with K, and a mapping ϕ from $\mathcal{X}$ to $\mathcal{H}$ such that the kernel is an inner product in $\mathcal{H}$.
- This RKHS represents a space of functions where we can search for the empirical risk minimizer.
- An important insight gained by this view point is the Representer theorem.

Representer Theorem

- Let K be a positive definite Kernel and let $\mathcal{H}$ be the RKHS associated with it.
- Let $\{(X_i, y_i),\ i = 1, \cdots, n\}$ be the training set.
- For any function f, the empirical risk, under any loss function can be represented as a function

$$\hat{R}_n(f) = C(\ (X_i, y_i, f(X_i)),\ i = 1, \cdots, n)$$

- We search over $\mathcal{H}$ for a minimizer of empirical risk.
- Let $||f||^2 = \ < f\ ,\ f >$ be the norm under our inner product.

- **Theorem**: Let $\Omega : [0, \infty) \to \Re^+$ be a strictly monotonically increasing function. Consider minimization of empirical risk over $\mathcal{H}$. Then any minimizer of the regularized risk

$$C(\,(X_i, y_i, g(X_i)),\ i = 1, \cdots, n)\ +\ \Omega(||g||^2)$$

admits a representation

$$g(X) = \sum_{i=1}^{n} \alpha_i\, K(X_i, X)$$

- What this means is the following.
- Functions in $\mathcal{H}$ are linear combinations of kernels centered at all points of $\mathcal{X}$.
- Though we are searching over this space, the minimizer can always be expressed as a linear combinations of kernels centered on data points only.
- Thus, irrespective of the dimension of $\mathcal{H}$, we can solve the optimization problem by searching for only n real numbers α_i.
- This is essentially what we have done in solving the dual for SVM.

Proof of Representer Theorem

- In the vector space $\mathcal{H}$, consider the span of the functions $K(X_1, \cdot), \cdots, K(X_n, \cdot)$. ($X_i$ are training data)
- This will be a subspace.
- Given any $f \in \mathcal{H}$, we can decompose it into two components – one in this subspace and one in the subspace orthogonal to it.
- Let us call these two components as $f_{\|}$ and $f_{\perp}$.

- Thus, For any $f \in \mathcal{H}$ and any $X \in \mathcal{X}$, we have

$$f(X) = f_{\|}(X) + f_{\perp}(X) = \sum_{i=1}^{n} \alpha_i \, K(X_i, X) + f_{\perp}(X)$$

  where $\alpha_i \in \Re$, $f_{\perp}(X) \in \mathcal{H}$ and
  $< f_{\perp}(X) \, , \, K(X_i, \, \cdot) > \; = 0, \; i = 1, \cdots, n$.
- Since $\mathcal{H}$ is the RKHS of K, the reproducing kernel property gives us

$$f(X') = \; < f \, , \, K(X', \, \cdot) >$$

- Hence for any of the data points, X_j, $j = 1, \cdots, n$,

- Hence for any of the data points, X_j, $j = 1, \cdots, n$,

$$\begin{aligned} f(X_j) &= \; < f \, , \, K(X_j, \, \cdot) > \\ &= \; < f_{\|} + f_{\perp} \, , \, K(X_j, \, \cdot) > \\ &= \; < f_{\|} \, , \, K(X_j, \, \cdot) > + < f_{\perp} \, , \, K(X_j, \, \cdot) > \\ &= \sum_{i=1}^{n} \alpha_i \, K(X_i, X_j) + \; < f_{\perp} \, , \, K(X_j, \, \cdot) > \\ &= \sum_{i=1}^{n} \alpha_i \, K(X_i, X_j) \; = \; f_{\|}(X_j) \end{aligned}$$

- This is true for any $f \in \mathcal{H}$.

- Now let $g \in \mathcal{H}$ be a minimizer of the regularized risk.
- We can write $g = g_{||} + g_{\perp}$
- $g(X_j) = g_{||}(X_j)$ for all data vectors, X_j.
- Hence the empirical risk of g,

$$C(\,(X_i, y_i, g(X_i)),\ i = 1, \cdots, n)$$

  would be same as empirical risk of $g_{||}$.
- Since $g_{||}$ and $g_{\perp}$ are orthogonal,

$$||g||^2 = ||g_{||}||^2 + ||g_{\perp}||^2 \geq ||g_{||}||^2$$

- Since Ω is strictly monotone increasing, $\Omega(||g||^2) \geq \Omega(||g_{\|}||^2)$.
- Hence we have

$$\begin{aligned} & C(\,(X_i, y_i, g(X_i)),\; i = 1, \cdots, n) \;+\; \Omega(||g||^2) \\ = \;& C(\,(X_i, y_i, g_{\|}(X_i)),\; i = 1, \cdots, n) \;+\; \Omega(||g||^2) \\ \geq \;& C(\,(X_i, y_i, g_{\|}(X_i)),\; i = 1, \cdots, n) \;+\; \Omega(||g_{\|}||^2) \end{aligned}$$

- This shows that the regularized risk of $g_{\|}$ can only be less than or equal to that of g.
- Hence any minimizer would be in the subspace spanned by $K(X_i, \cdot)$ and hence would have a representation

$$g(X) = \sum_{i=1}^{n} \alpha_i \, K(X_i, X)$$

- This completes proof of the theorem.