CAP 6617 ADVANCED MACHINE LEARNING

NAME: SANKET KUMAR   UF ID: 4513-3352

GENERALIZED REPRESENTER THEOREM

AIM: To show that a large class of optimization problems with RKHS regularizers have solutions that can be expressed as kernel expansions in terms of the training data.

All positive definite kernels K are also inner products on some functional space called Reproducing Kernel Hilbert Space(RKHS) associated with the kernel K:

- The RKHS(or **H**) is the set of all functions that are finite linear combinations of kernel functions {K(.,X) : ***X***$\in$input space(say ***X***)}. E.g. $f(.) = \sum_{i=1}^{m} a_i k(.\,, x_i)$ and $g(.) = \sum_{j=1}^{m'} b_j k(.\,, x_j')$ where a, b$\in$**C**. These also follow f, g$\in$**H**; f + g$\in$**H**; a*f$\in$**H** for some a$\in$**R** so **H** is a vector space. The inner product of f and g is defined as $\sum_{i=1}^{m} \sum_{j=1}^{m'} \bar{a}_i b_j k(x_i, x_j')$…(1).
- To show that this is an inner product, we first show that it is well defined: <f, g> = $\sum_{i=1}^{m} a_i \sum_{j=1}^{m'} b_j k(x_i, x_j') = \sum_{i=1}^{m} a_i g(x_i)$ and <f, g> = $\sum_{i=1}^{m'} b_j \sum_{j=1}^{m} a_i k(x_i, x_j') = \sum_{i=1}^{m'} b_j f(x_j')$ which shows that the inner product does not depend upon the specific expansion coefficients and hence is well defined.
- To show this is actually an inner product, it is easy to prove the properties of inner products: <f,g> = $\overline{< g, f >}$ and bi-linearity. Let $f_1 \ldots f_n \in$H and $c_1 \ldots c_n \in$R, then $\sum_{i,j=1}^{p} c_i c_j < f_i, f_j >$ = <$\sum_{i=1}^{m} c_i f_i$,$\sum_{j=1}^{m'} c_j f_j$> through bi-linearity, which is always >=0. <.,.> is a symmetric function that maps **H**x**H** to **R**. This shows that <.> is a positive definite kernel on **H**.
- Now <*k*(.,X), *k*(.,X')> = *k*(X,X') and <*k*(.,X),f> = f(X)…(2) from (1) is known as reproducing kernel property. Since positive definite kernels satisfy Cauchy-Schwartz inequality, for any f$\in$**H**, we have $|f(X)^2| = |<k(.,X),f>|^2$ from (2) <= *k*(X,X)<f,f>. This shows that if <f,f>=0, f=0. This finally proves that what we defined is indeed an inner product.

Representer Theorem:

- Let g be a strictly monotonically increasing function on [0, ∞), f be a minimizer over **H** and regularized risk functional be $c((x_1,y_1,f(x_1))\ldots(x_m,y_m,f(x_m))) + g(||f||)$, where $(x_i,y_i)\ldots(x_n,y_n)$ are training examples$\in$***X******R**. Let ϕ : ***X*** (input space) -> **R**.
- In the space **H**, the span of the functions $k(X_1,.)\ldots k(X_n,.)$ will be a subspace. We can decompose f into this subspace and the subspace orthogonal to it. Thus, $f = f_1+v$ where f_1 is parallel and v is perpendicular. Since f_1 is in the span, we can write $f = \sum_{i=1}^{m} a_i\phi(x_i) + v$. Since v is orthogonal, <v, f_1> = 0 and by bi-linearity <v, $\phi(x_j)$> = 0.
- By reproducing kernel property, $f(X_j)$ = <f,K(X_j,.)> = <f_1+v, K(X_j,.)> = <$\sum_{i=1}^{m} a_i\phi(x_i) + v$, $\phi(x_j)$> = <$\sum_{i=1}^{m} a_i\phi(x_i)$,$\phi(x_j)$> + 0 = $f_1(X_j)$. This shows that $f(x_j)=f_1(x_j)$ and is independent of its orthogonal projection. Since f_1 and v are orthogonal, $||f||^2 = ||f_1||^2 + ||v||^2 >= ||f_1||^2$ …(3). Since g is monotonically increasing, $g(||f||) = g(||\sum_{i=1}^{m} a_i\phi(x_i)+v||) = g((||\sum_{i=1}^{m} a_i\phi(x_i))||^2+||v||^2)^{1/2}) >= g(||a_i\phi(x_i)||)$ from (3).
- This shows that the regularized risk of f_1 can only be less or equal to that of f. Hence the minimizer would be in the subspace spanned by *k*(X_i,.) and hence would have a representation $f(.) = \sum_{i=1}^{m} a_i k(.\,, x_i)$.

Semiparametric Representer Theorem: In addition to the representer theorem, given a set of **M** real valued functions $\{\psi_p\}$ on X such that the **m**x**M** matrix has rank **M**, then any f' = f + h with f$\in$**H** and h$\in$span{ψ(p)} can be expressed as $f(.) = \sum_{i=1}^{m} a_i k(x_i, .) + \sum_{i=1}^{M} b_p \psi_p(.)$ minimizing the regularized risk. Some applications of both the theorems are: (1) SV Regression and classification (2) SVM's actual minimizing risk bounds (3) Bayesian MAP estimates.

Conclusion: A large class of algorithms minimizing a sum of an empirical risk term and a regularization term in a reproducing kernel Hilbert space, the optimal solutions can be written as kernel expansions in terms of training examples. As long as the objective function can be cast into the form considered in the generalized representer theorem, one can recklessly carry out algorithms in infinite dimensional spaces, since the solution will always live in a specific subspace whose dimensionality equals at most the number of training examples.