13.1 Review of Last Lecture

Review of primal and dual of SVM.

Insights:

- Dual only depends on inner products \((x_i^T x_j)\). This inner product can be replaced by a kernel function \(k(x_i, x_j)\) which takes the inner product in a high dimensional space: \(k(x_i, x_j) = \phi(x_i)^T \phi(x_j)\)

- Representation property: at optimal solution, the weight vector \(w\) is a linear combination of the data points; that is, the optimal weight vector lives in the span of the data. \(w^* = \sum_i \alpha_i y_i x_i\) with kernels \(w^* = \sum_i \alpha_i y_i \phi(x_i)\). Note that \(w^*\) can be an infinite dimensional vector, that is, a function.

- In some sense, we can treat our problem as a parameter estimation problem; the dual problem is non-parametric (one parameter / dual variable per data point)

What about noise?

We introduce Slack variables. In the primal formulation we have:

\[
\min_w \frac{1}{2} w^T w + C \sum_i \xi_i \text{ such that } y_i w^T x_i \geq 1 - \xi_i
\]

which is equivalent to

\[
\min_w \frac{1}{2} w^T w + C \sum_i \max(0, 1 - y_i w^T x_i)
\]

The first term above serves to keep the weights small, while the second term is a sum of hinge loss functions, which are high for poor fit. The two terms balance against one another in the minimization.

13.2 Kernelization

Naive approach to Kernelization: see what happens if we just assume that

\[
w = \sum_i \alpha_i y_i x_i.
\]
Then the optimization problem becomes equivalent to

$$\min_{\alpha} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j + C \sum_i \max(0, 1 - y_i \sum_j \alpha_j y_j x_j^T x_i).$$

To kernelize, replace $x_i^T x_j$ terms with $k(x_i, x_j)$ When is this appropriate? The key assumption is that $w \in Span\{x_i, \forall i\}$ (which we derived last lecture in the case of no-noise).

Let $\tilde{\alpha}_i = \alpha_i y_i$ Note that we’re unconstrained here: we can flip signs arbitrarily.

Then the problem is equivalent to

$$\min_{\tilde{\alpha}} \frac{1}{2} \sum_{i,j} \tilde{\alpha}_i \tilde{\alpha}_j k(x_i, x_j) + C \sum_i \max(0, 1 - y_i \sum_j \tilde{\alpha}_j k(x_i, x_j)).$$

Recall:

$$K = \begin{pmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{pmatrix}$$

This matrix is called the "Gram Matrix", and so the above is equivalent to

$$\min_{\tilde{\alpha}} \frac{1}{2} \tilde{\alpha}^T K \tilde{\alpha} + C \sum_i \max(0, 1 - y_i f(x_i))$$

where the first term is the complexity penalty and the second term represents the penalty for poor fit, where we use the notation $f(x) = f_\alpha(x) = \sum_j \alpha_j k(x_j, x)$.

Suppose we want to learn a non-linear classifier for the unit interval: one way to do this is to learn a non-linear function $f$ which takes values roughly the labels, st. $y_i \approx \text{sign}(f(x_i))$ This function could fit this condition only at the datapoints and so look sort of like a comb (with the teeth at the datapoints, and the function otherwise near zero) or it could be a much more smoothly varying function which takes a value in between the datapoints which is similar to close by datapoints. These functions are sketched in Figure 13.2.

The complicated, comb-like, high-order function would work, but we would prefer the simpler, smoother function: To ensure goodness-of-fit, we want to have correct prediction with a good margin: $|f(x_i)| > 1$. To control complexity, we prefer simpler functions.

How can we mathematically express this preference? In general, we want to solve:

$$f^* = \min_{f \in F} \frac{1}{2} ||f||^2 + C \sum_i l(f(x_i), y_i)$$

where $l$ is an arbitrary loss function, for example, the hinge loss used above.

Questions: what is $F$? What is the right norm/complexity of the function?
In the following, we will answer these questions. For the definition of $||f||$ that we will derive, it will, for functions $f = f_\alpha = \sum_i \alpha k(x_i, \cdot)$, it will hold that $||f||^2 = \alpha^T K \alpha$, i.e., the same penalty term as introduced above.

In the following, we will assume that $l$ is an arbitrary loss function, i.e., we require that $l(f(x_i), y_i) \geq 0$ and if $f(x_i) = y_i$ then $l(f(x_i), y_i) = 0$.

13.3 Reproducing Kernel Hilbert Spaces

Definition 13.3.1 (Hilbert space) Let $X$ be a set ("index set")

A Hilbert space $H = H(X)$ is a linear space of functions $H : \{f : X \rightarrow \mathbb{R}\}$ along with an inner product $< f, g >$ (which implies a norm $||f|| = \sqrt{< f, f >}$) which is complete: all Cauchy sequences in $H$ converge to a limit in $H$.

Definition 13.3.2 (Cauchy Sequence) $f_1, \ldots, f_n$ is a Cauchy sequence if $\forall \epsilon, \exists n_0$ such that $\forall n, n' \geq n_0 ||f_n - f_n'|| < \epsilon$. The Cauchy sequence $f_1, \ldots, f_n$ converges to $f$ if $||f_n - f|| \rightarrow 0$ as $n \rightarrow \infty$.

Definition 13.3.3 (RKHS) A Hilbert space is called a Reproducing Kernel Hilbert Space (RKHS) for kernel function $k$ if both of the following conditions are true:

1) any function $f \in H$ can be written as an infinite linear combination of kernel evaluations: $f = \sum_{i=1}^{\infty} a_i k(x_i, \cdot)$ for $x_1, \ldots, x_n \in X$

Note that for any fixed $x_i$, $k(x_i, \cdot)$ maps $X \rightarrow \mathbb{R}$

2) $H_k$ satisfies the reproducing property:

$< f, k(x_i, \cdot) > = f(x_i)$ that is, the kernel function clamped to one $x_i$ is the evaluating functional for that point.

The above definition implies that $< k(x_i, \cdot), k(x_j, \cdot) > = k(x_i, x_j)$ ← entries in the Gram matrix
Example: $X = \mathbb{R}^n$

$H = \{ f : f(x) = w^T x \text{ for some } w \in \mathbb{R}^n \}$

For functions $f(x) = w^T x$, $g(x) = v^T x$, define $< f, g > = w^T v$

Define kernel function (over $X$): $k(x, x') = x^T x'$

Verify (1) and (2)

(1) Consider $f(x) = w^T x = \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i k(e_i, x)$ where $e_i$ is the indicator vector: the unit vector in the $i$th direction. So (1) is verified.

(2) Let $f(x) = w^T x < f, k(x_i, \cdot) >= w^T x_i = f(x_i)$ so (2) is verified. Note that the first equality holds because $k(x_i, x) = x_i^T x$ and $k(x_i, x)$ is a function on $X \rightarrow \mathbb{R}$ because $k : X \times X \rightarrow \mathbb{R}$.

13.4 Important points on RKHS

Questions:

(a) Does every kernel $k$ have an associated RKHS?

(b) Does every RKHS have a unique kernel?

(c) Why is this useful?

Answers (a) Yes:

let $H_k' = \{ f = \sum_{i=1}^{n} k(x_i, \cdot) \}$ and

$< f, g > = \sum_{i,j} \alpha_i \beta_j k(x_i, x_j)$ for $f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$, $g = \sum_{j=1}^{m} \beta_j k(x_j, \cdot)$

Check if this satisfies the reproducing property: $< f, k(x', \cdot) > = \sum_{i=1}^{n} \alpha_i k(x, x')$

Space $H_k'$ is not yet complete: add all limits of all cauchy sequences to it to complete it. Then this is an RKHS.

Define $\phi : x \rightarrow k(x, \cdot)$ consider $< \phi(x), \phi(x') > = < k(x, \cdot), k(x', \cdot) > = k(x, x')$ Can think of $k$ as an inner product in that RKHS. The above is an explicit way of constructing the high dimensional space for which the kernel function is the inner product.

(b) Consider $k$ and $k'$, two positive definite kernel functions which produce the same RKHS. Does $k = k'$? Yes $\rightarrow$ next homework assignment.

(c) Why is this useful? Return to our original problem:

$$f^* = \min_{f \in F} \frac{1}{2} ||f||^2 + \sum_{i} l(f(x_i), y_i)$$

Let $F$ be an RKHS: $F = H_k$

Theorem 13.4.1 For arbitrary (not even convex) loss functions of form above, any optimal solution to the problem can be written as a linear combination of these kernel evaluations: for all datasets $x_i, y_i \exists \alpha_1, \ldots, \alpha_n$ such that $f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$
**Proof:** Next lecture.

The above is from [Kimmeldorf and Wahba].

Representer Theorem: For convex loss functions under strong convexity conditions, the solution is unique. If not strongly convex, but convex, the set of solutions is a convex set.