1 Kernel-based Learning Algorithms

Kernel-based Learning Algorithms are used in data analysis and machine learning. There are several types of learning mechanism:

- Unsupervised Learning - No teacher/labels
- Supervised Learning - Teachers/labels
- Semi-supervised Learning - The labels might be expensive and only some data point has labels.
- Online Learning - Time Series Data

2 Kernel-based Method

- A way to model a much larger class of data using a vector space model.
- A lot more descriptive flexibility without much additional computational cost.

The kernel method involves a mapping into a high (possibly infinite) dimensional space.

\[ \phi(X) : X \rightarrow F \]

Given a set of vector \( x_i \in \mathbb{R} \), we define the Gram matrix \( G \):

\[ G_{ij} = x_i^T x_j \]

which is a symmetric matrix of inner products.

**Definition** Given a matrix \( G \), we say \( G \) is positive semi-definite if for all vectors \( x \), we have \( x^T G x \geq 0 \).

We can also generalize the concept of positive number to a partial ordering on matrices. To compare two matrices \( A, B \), we can check if \( A - B \) is positive semi-definite.

In \( \mathbb{R}^n \), any Gram matrix is positive semi-definite. Also, any positive semi-definite matrix is a Gram matrix for some set of vectors.

**Note** The set of vectors that generates a certain Gram matrix is not unique.
3 Supervised Learning - Classification

There are different ways to formalize this. One way is to say the data are \((x_i, y_i)_{i=1}^n \in \mathbb{R}^N \times Y\), where \(Y = \{-1, 1\}\). Then the goal is to find a function \(f : \mathbb{R}^N \rightarrow Y\) such that if given a new example, it will classify it correctly. For example, we can say

\[
\begin{align*}
  f(x) > 0 & \rightarrow \text{assign 1} \\
  f(x) < 0 & \rightarrow \text{assign -1}
\end{align*}
\]

Question: What if the data is more representable as a graph?

3.1 Risk Minimization

Given some training data \((x_i, y_i)_{i=1}^n\) and also test data drawn from the same distribution \(P(x, y)\), our goal is to find the best function \(f\) from what we already know:

- \((x_i, y_i)_{i=1}^n\)
- a function class \(I\) to optimize over

We want to minimize the risk/error defined by

\[
R[f] = \mathbb{E} [L(f(x), y)] dP(x, y)
\]

where \(L\) denotes some loss function.

Our goal is to minimize \(R[f]\) subject to bias/variance trade-off while having flexibility generally.

3.2 Empirical Risk Minimization (ERM)

The empirical risk is defined on the test data set:

\[
R_{\text{emp}}[f] = \frac{1}{n} \sum_{i=1}^n L(f(x_i), y_i)
\]

and we hope that if \(n \rightarrow \infty\), we would have \(R_{\text{emp}}[f] \rightarrow R[f]\).

3.3 Structural Risk Minimization

The idea is to restrict ourselves to some nice function class and do ERM with the following procedures:

1. Construct a nested family of function class

   \[
   F_1 \subset F_2 \subset \cdots \subset F_k
   \]

2. Let \(f_1, \ldots, f_k\) be the ERM solutions in \(F_k\)

3. Choose \((k^*, F_{k^*}, f_{k^*})\) such that upper bound on generalization error is minimized.
**Theorem**  Let $h$ be the VC dimension of $I$. Then $\forall \delta > 0, \ f \in I$

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h(\ln(\frac{2n}{h} + 1)) - \ln(\delta/4)}{n}}$$

with probability $(1 - \delta), \ \forall n > h$

**Note**  The above bound represents a bias/variance trade-off. It doesn’t not depend on $P(x,y)$. The main reason to use this bound is that we cannot compute the LHS, but given any $h$, we can compute the RHS.

### 3.4 VC dimension

**Definition**  A *dichotomy* of set $S$ is a partition of $S$ into two disjoint pieces.

**Definition**  A set of points $S$ is *shattered* by a hypothesis space $H$ if for all dichotomy of $S$, $\exists$ a hypothesis $h \in H$ consistent with the dichotomy.

**Definition**  The *VC dimension* of $H$ over given set of points $S$ is the size of the largest subset of $S$ shattered by $H$.

The point is that the complexity of a classifier does not depend on the size of $H$, but on how it performs on $S$.

**Note**  To show that the VC dimension of $H$ is $\geq d$. View it as a game:

1. I choose $d$ points.
2. The adversary chooses labels from $\{-1, 1\}$.
3. I produce a hypothesis $h \in H$.

The VC dimension is the maximum of such $d$.

**Note**  The VC dimension is powerful to bound certain things, but

1. it can be hard to work with.
2. it is suboptimal bound.
3. it is a distribution-independent bound.

### 3.5 Hyperplane

**Definition**  *Hyperplane* is a set of $H$ in $\mathbb{R}^n$ that is “nice” and has the following form: $\langle w, x \rangle + b = 0$. The decision boundary of a hyperplane is sign ($\langle w, x \rangle + b$).

**Claim**  In $\mathbb{R}^2$, I can find 3 points such that I can shatter with a hyperplane, but I can not find 4. The general result is that given $m$ points in $\mathbb{R}^n$, they can be shattered by oriented hyperplane if and only if the points we have are linearly independent.

**Claim**  The VC dimension of oriented hyperplane in $\mathbb{R}^n$ is $n + 1$. 

3
**Definition** We say the data \( \{x_i, y_i\}_{i=1}^n \) are linearly separable if there exist \( w, b \) such that \( \langle w, x \rangle + b = 0 \) separates the data, i.e.

\[
x_i \cdot w + b \geq 1 \quad \text{if} \quad y_i = 1
\]
\[
x_i \cdot w + b \leq -1 \quad \text{if} \quad y_i = -1
\]

**Definition** Let \( d_+ (d_-) \) be the shortest distance from the separating hyperplane to a data point with \((+)(-))\) label. The margin of the separating hyperplane w.r.t. the data is \( d_+ + d_- \). If \( w \) is the weight vector, then \( d_+ = d_- = \frac{1}{||w||} \), so the margin \( \gamma = 2/||w|| \).

**Fact** Let \( \mathcal{H} \) be the set of linear classifiers, and \( \mathcal{H}_\gamma \) be the set of linear classifiers with margin \( \gamma \). Intuitively, \( \mathcal{H}_\gamma \) is smaller than \( \mathcal{H} \). Let \( R \) be the radius of the smallest inclosing ball of the data, then

\[
VC(\mathcal{H}_\gamma) \leq R^2 w \cdot w + 1,
\]

independent of dimension.

### 3.6 Support Vector Machines (SVM)

The fact in (1) suggests optimizing the margin (SVM). Given \( \{x_i, y_i\} \in \mathbb{R}^N \times \{-1, 1\} \), find a good classification hyperplane given by the following optimization problem:

\[
\min_{w, b} \quad \frac{1}{2} ||w||^2
\]
\[
\text{Subject to} \quad y_i (\langle w, x \rangle + b) \geq 1.
\]

We can write (2) as an unconstrained problem with the Lagrange multipliers. Define

\[
L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_i \alpha_i (y_i (\langle w, x \rangle + b) - 1),
\]

then (2) becomes

\[
\min_{w, b} \max_{\alpha > 0} \quad L(w, b, \alpha)
\]

We can view (3) as a two player game

1. If player A violates the constraint in (2), then player B can choose \( \alpha \) such that the maximum goes to \( \infty \).
2. If player A satisfies the constraint in (2), then player B chooses \( \alpha_i = 0 \).