CSCI 590: Machine Learning

Lecture 13:

VC theory and support vector machines (SVMs)

Instructor: Murat Dundar

Acknowledgements:
1. SVM Tutorial by C. Burges (http://research.microsoft.com/pubs/67119/svmtutorial.pdf)

Structural Risk Minimization and VC Theory

Structural Risk Minimization:

For a given learning task, with a given finite amount of training data, the best generalization performance will be achieved if the right balance can be established between the accuracy on a particular training set, and the capacity of the machine.

Bound on the Test Error (1)

Suppose we are given n observations $\{x_i, y_i\}$.

Assume binary labels (y_i=1 for positive cases, y_i=-1 for negative cases)

There exists an unknown p(x,y) from which these data are i.i.d. drawn

Suppose we have a machine that maps x_i onto y_i .

The machine is defined by a set of possible mappings $\mathrm{x} \rightarrow f(x, w)$.

The machine is deterministic: for a given input x, and a choice of w, it will always give the same output f(x,w).

Bound on the Test Error (2)

Suppose there exists a trained machine w.

The expectation of the test error for a trained machine is:

$$R(w) = \int \frac{1}{2} |y - f(x, w)| dP(x, y)$$

The quantity R(w) is called the expected risk, or the actual risk.

Bound on the Test Error (3)

$$R_{emp}(w) = \frac{1}{2N}\sum_{i=1} |y_i - f(x_i, w)|$$

The empirical $R_{emp}(\alpha)$ is defined to be just the measured error rate on the training set.

Bound on the Test Error (4)

Now choose η such that $0 \leq \eta \leq 1$.

Then, with probability $1-\eta$, the following bound holds (Vapnik, 1995)

$$R(w) \leq R_{emp}(w) + \sqrt{\left(\frac{h\left(\log\left(\frac{2N}{h}\right)+1\right)-\log\left(\frac{\eta}{4}\right)}{N}\right)}$$

where h is a non-negative integer called the Vapnik Chervonenkis (VC) dimension, and is the measure of "capacity".

Bound on the Test Error (5)

$$R(w) \leq R_{emp}(w) + \sqrt{\left(\frac{h\left(\log\left(\frac{2N}{h}\right)+1\right)-\log\left(\frac{\eta}{4}\right)}{N}\right)}$$

Three key points:

1. Independent of P(x,y)
2. Usually not possible to compute R(w)
3. If we know h, we can compute the right side

Bound on the Test Error (6)

Given several different learning machines, i.e. family of functions, choose a sufficiently small η, the machine which minimizes the right hand side, gives the lowest upper bound on the actual risk.

This is the essential idea of structural risk minimization.

VC Dimension (1)

VC dimension is a property of a set of functions $\mathrm{f}\{(w)\}$.

Let's consider functions that correspond to the two-class classification problem.

A given set of n points can be labeled in 2^n possible ways.

If for each labeling we can find a member of the set $\mathrm{f}\{(w)\}$ that can correctly assign those labels, then we say the set of points is *shattered* by that set of functions.

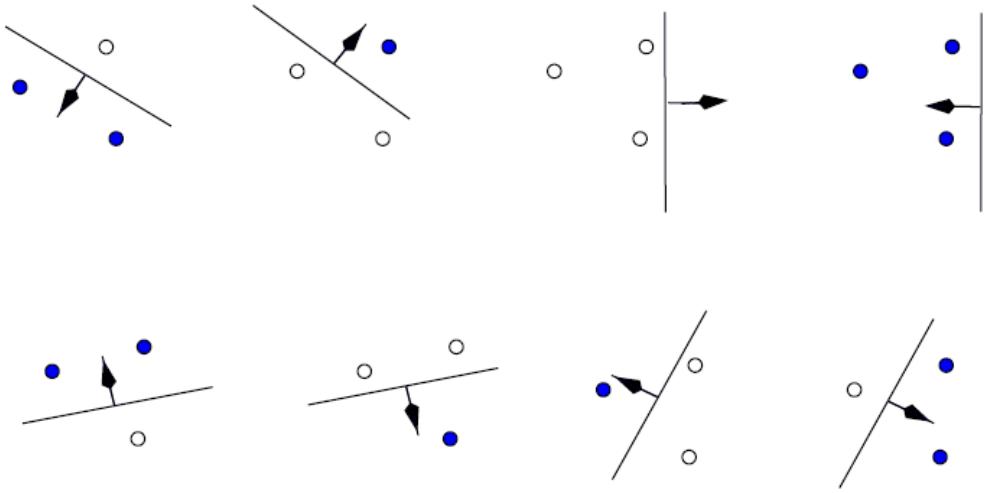
VC Dimension (2)

The VC dimension for a set of functions $f\{(w)\}$ is defined as the maximum number of points that can be shattered by $f\{(w)\}$.

If the VC dimension is h, then there exists at least one set of h points that can be shattered.

VC Dimension (3)

Suppose $x \in R^2$, and the set f$\{(w)\}$ consists of oriented straight lines.



Source: A tutorial on SVMs by C. Burges

VC Dimension (4)

While it is possible to shatter three points by this set of functions, it is not possible to find four points that can be shattered by oriented straight lines.

The VC dimension of the class of oriented straight lines is 3.

VC Dimension (5)

Theorem: Consider some set of m points in R^d. Choose any one of these points as the origin. Then the m points can be shattered by oriented hyperplanes iff the position vectors of the remaining points are linearly independent.

Corollary: The VC dimension of the set of oriented hyperplanes in R^d is d+1, since we can always choose d+1 points, and then choose one of the points as origin, such that the position vectors of the remaining n points are linearly independent.

Support vector machines

Suppose we have a seperating hyperplane $w^T x + b = 0$.

Let $d_+(d_-)$ be the shortest distance from the separating hyperplane to the closest positive (negative) example.

We define the "margin" of the separating hyperplane as $d_+ + d_-$

For the linearly separable case, SVM looks for the separating hyperplane with largest margin.

Defining the margin (1)

Let us suppose that all training data satisfy the following

$$w^T x_i + w_0 \geq 1 \text{ for } y_i = 1$$
$$w^T x_i + w_0 \leq -1 \text{ for } y_i = -1$$

More compactly

$$y_i(w^T x_i + w_0) \geq 1 \; \forall i$$

Defining the margin (2)

Pick a point x_1 on $H_+: w^T x_1 + w_0$ and pick another point x_2 perpendicular to x_1 on $H_-: w^T x_2 + w_0$.

Note that w is orthonormal to both H_+ and H_- and thus, the inner product of the vector x_1- x_2 with w is

$$(x_1 - x_2)^T w = \|x_1 - x_2\| \|w\| \cos(0)$$

We know that the perpendicular distance between H_+ and H_- is

$$w^T (x_1 - x_2) = 2$$

Thus, the margin between H_+ and H_- is $\|x_1 - x_2\| = 2/\|w\|$