Master of Technology

U2/6: Computational Intelligence I

Support Vector Machines
Day 5 Statistical Machine Learning

Dr TIAN Jing
Institute of Systems Science,
National University of Singapore
Email: tianjing@nus.edu.sg

Statistical machine learning

- Learning process can be mathematically described as a system that
  - » Receives data (observations) as input and
  - » Outputs a function that can be used to predict some features of future data

- Statistical machine learning models this as a function estimation problem
  - » Measures Generalization Performance as accuracy in labeling test data

Empirical risk minimization (ERM)

- In statistical learning theory, enumerative induction is called empirical risk minimization
  - » In general, the risk (expected error) cannot be computed
    - ♦ because the distribution is unknown to the learning algorithm.
  - » However, we can compute an approximation, called empirical risk,
    - ♦ by averaging the loss function on the training set

ERM: Over-fitting

- Empirical risk minimization (ERM) principle states that the learning algorithm should choose a hypothesis which minimizes the empirical risk
  - » Commonly in machine learning, a generalized model must be selected based on a finite data set
    - ♦ with the consequent problem of over-fitting – the model becoming too strongly tailored to the particularities of the training set and generalizing poorly to new data.

Learning machine

- For a given learning task (e.g., classification, recognition, regression, …), with a given finite amount of training data
  - » a best generalization performance will be achieved if the right balance is struck between
    - ♦ the accuracy on that particular training set
      and
    - ♦ the "capacity" of the machine, i.e.: the ability of the machine to learn any training set without error

Learning machine

- Suppose we have a machine whose task is to learn the mapping $\mathbf{x}_i \ \alpha \ y_i \quad \forall_i$
- The machine is actually defined by a set of possible mapping: $\mathbf{x} \ \alpha \ f(\mathbf{x}, \alpha)$
  - » Where the functions themselves are labeled by the adjustable parameters α
  - » A particular choice α generates what we call a "trained machine", or a model learned
    - ♦ E.g.: a neural network with fixed architecture, with α corresponding to the weights and biases

Expectation of error

- The expectation of the test error for a trained machine is therefore

$$R(\alpha) = \int \frac{1}{2}\left|y - f(\mathbf{x}, \alpha)\right| dP(\mathbf{x}, y)$$

$R(\alpha)$ is the expected risk, here we call it the actual risk (to emphasize that it is the quantity we are ultimately interested in, sometimes also called true risk in literature)

$\frac{1}{2}\left|y_i - f(\mathbf{x}_i, \alpha)\right|$ is called the loss, it can only take the values 0 and 1 (because $y \in \{-1,1\}$)

$P(\mathbf{x}, y)$ is the (unknown) joint distribution function of $\mathbf{x}$ and y

Empirical error

- The "empirical risk" is defined to be just the measured error rate on the training set (for a fixed, finite number of observations)

$$R_{emp}(\alpha) = \frac{1}{2l}\sum_{i=1}^{l}\left|y_i - f(\mathbf{x}_i, \alpha)\right|$$

where l is the number of data points

- Note
  - » There is no probability distribution involved
  - » $R_{emp}(\alpha)$ is a fixed number for a particular choice of α and for a particular training set $\{\mathbf{x}_i, y_i\}$

Empirical risk and true risk

- Vapnik & Chervonenkis showed that an upper bound on the true risk can be given by the empirical risk + an additional term
  - » With probability $1 - \eta$ such that $0 \leq \eta \leq 1$, the following bound holds [Vapnik, 1995]

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

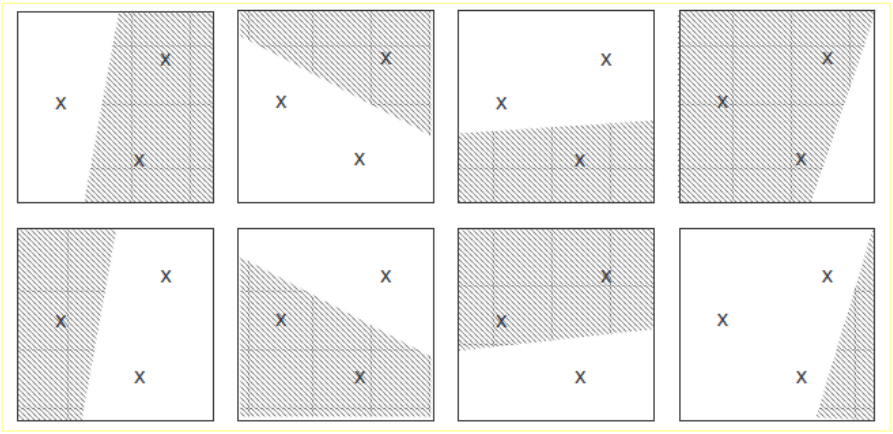
VC confidence

Risk bound

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

VC dimension: Basic idea

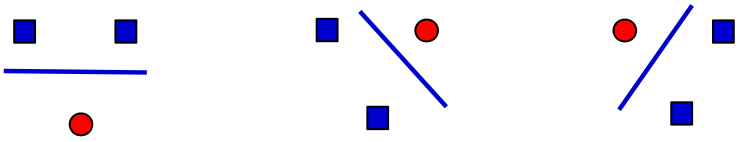
- VC Dimension is a measure of machine capacity or complexity
  - » The VC dimension of a set of functions is the maximum number of points that can be separated in all possible ways by that set of functions
  - » For hyperplane in R^n, the VC dimension can be shown to be n + 1



(source: J. Weston, "Support Vector Machine Tutorial")

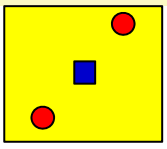
VC dimension: Example

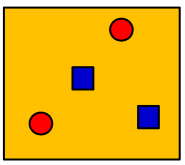
- Consider the set of all linear classifications of points in the plane where YESes and Nos are separated by a straight line.
  - » Some set of three points (not collinear) in the plane can be separated

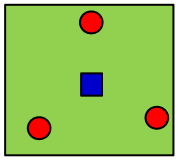


  - » So the VC dimension for these linear separations is <u>at least 3</u>

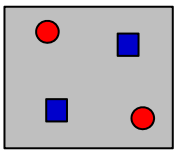
VC dimension: Example

- Three collinear points cannot be separated 
- No four points can be separated by this class of rule:

any three of the four are collinear; 



either one of the points is within the triangle defined by the other three;

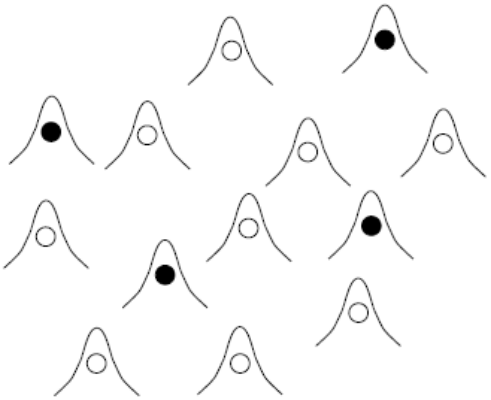
Or none of them is 

The VC dimension of the set of all linear separations in D-dimensional space is D+1

- So the VC dimension for these linear separations is <u>exactly 3</u>

VC dimension of SVM RBF classifiers

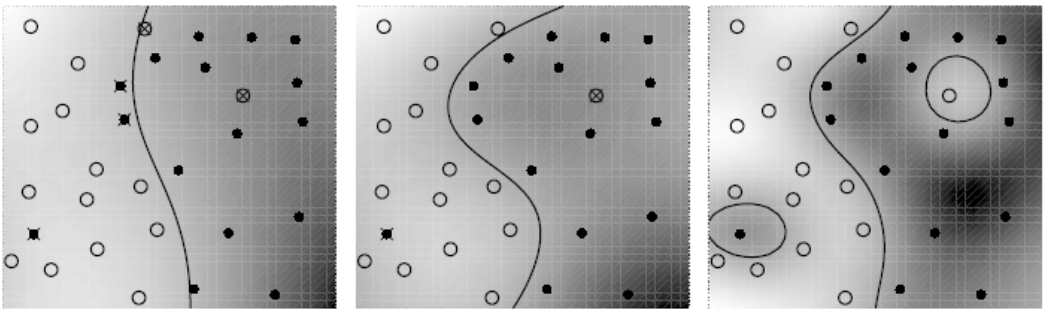
- Some theoretical works related to VC dimension of SVM with RBF kernels
  - » Even though VC dimension of these classifiers is infinite, SVM RBF can have excellent performance [Schölkopf et al, 1997]
    - ♦ A similar story holds for polynomial SVMs !



Gaussian RBF SVMs of sufficiently small width can classify an arbitrarily large number of training points correctly, and thus have infinite VC dimension
[Burges. 1998]

Training error and capacity of functions

- A high capacity set of functions (explain a lot)
  - » may lead a low training error, but might "overfit"
- A very simple model
  - » gives low complexity, but won't get low training error
- Capacity of a set of functions (classification)



(source: J. Weston, "Support Vector Machine Tutorial")

Risk bound

- Three key points about the risk bound
  - » It is independent of $P(\mathbf{x}, y)$
    - ♦ Only assume both the training data and the test data are drawn independently according to some $P(\mathbf{x}, y)$
  - » It is usually not possible to compute the left hand side (the true risk). However,
  - » If we know *h*, we can minimizes the right hand side
    - ♦ given several different learning machines, choosing a fixed, sufficiently small η, taking the machine which gives the lowest upper bound on the actual risk

$\Rightarrow$ the essential idea of structural risk minimization

Structural risk minimization (SRM)

- Structural risk minimization (SRM) is an inductive principle of use in machine learning
  - » SRM addresses the generalization problem by balancing the model's complexity against its success at fitting the training data.

SVM follows the principle of structural risk minimization that is rooted in VC dimension theory

- » Solve the generalization problem with high dimensionality

Summary

- Empirical risk minimization VS Structural risk minimization

SVM Day 5 Statistical Machine Learning

Thank you!

Dr TIAN Jing
Email: tianjing@nus.edu.sg