SVM and Kernel methods

Primary references:

John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*

Christopher Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2, 121–167 (1998).

Other references:

Aronszajn, Theory of reproducing kernels. Transactions of the American Mathematical Society, 686, 337-404, 1950.

Machine learning: support vector machine

Felipe Cucker and Steve Smale, On the mathematical foundations of learning. Bulletin of the American Mathematical Society, 2002.

Teo Evgeniou, Massimo Pontil and Tomaso Poggio, Regularization Networks and Support Vector Machines Advances in Computational Mathematics, 2000.

Grace Wahba, Spline Models for Observational Data Series in Applied Mathematics, Vol. 59, SIAM, 1990. (Chapter 1)

SVM in cancer

1. SVM illustration in cancer classification

Example 1: Myeloid vs. Lymphoblastic leukemias

ALL: acute lymphoblastic leukemia
AML: acute myeloblastic leukemia

SVM training: leave one out cross-validation

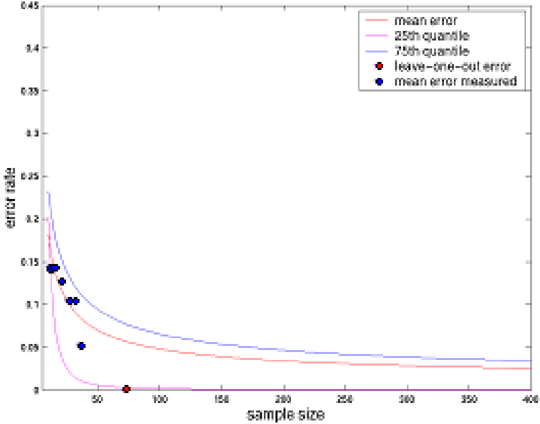
SVM in cancer

Dataset	Algorithm	Total Samples	Total errors	Class 1 errors	Class 0 errors	Number Genes
Leukemia Morphology (trest) AML vs ALL	SVM	35	0/35	0/21	0/14	40
	WV	35	2/35	1/21	1/14	50
	k-NN	35	3/35	1/21	2/14	10
Leukemia Lineage (ALL) B vs T	SVM	23	0/23	0/15	0/8	10
	WV	23	0/23	0/15	0/8	9
	k-NN	23	0/23	0/15	0/8	10
Lymphoma FS vs DLCL	SVM	77	4/77	2/32	2/35	200
	WV	77	6/77	1/32	5/35	30
	k-NN	77	3/77	1/32	2/35	250
Brain MD vs Glioma	SVM	41	1/41	1/27	0/14	100
	WV	41	1/41	1/27	0/14	3
	k-NN	41	0/41	0/27	0/14	5

S. Mukherjee

fig. 1: Myeloid and Lymphoblastic Leukemia classification by SVM, along with other discrimination tasks; k-NN is k-nearest neighbors; WV is weighted voting

SVM in cancer



S. Mukherjee

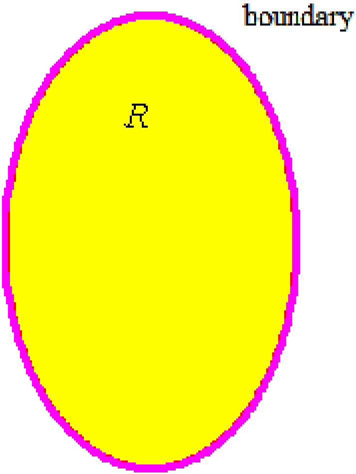
fig 2: AML vs. ALL error rates with increasing sample size;

SVM in cancer

Above curves are error rates with split between training and test sets.

Red dot represents leave one out cross-validation error rate.
Point data are values from selected single experiments.

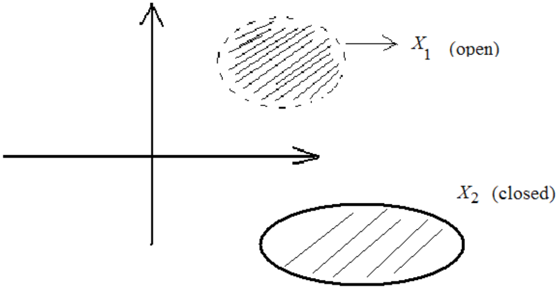
Some topology



Some topology

Def 3: A set $X \subset \mathbb{R}^d$ is *open* if it does not contain its boundary. It is *closed* if it contains its boundary.

Ex 1: in $\mathbb{R}^2$:



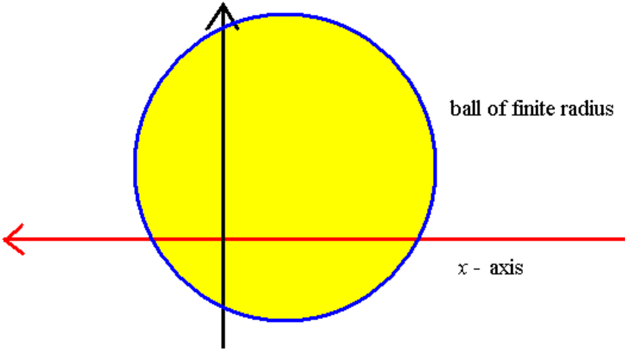
Some topology

Theorem 1: A set $\mathcal{O} \subset \mathbb{R}^d$ is open iff $\sim \mathcal{O}$ (= complement of $\mathcal{O}$) is closed.

Def. 4. $R \subseteq \mathbb{R}^d$ is *bounded* if it is contained in some (sufficiently large) ball $B_M(0)$, i.e., does not extend to ∞

Some topology

Ex.: In $\mathbb{R}^2$, a ball of radius 5 is bounded; x-axis is unbounded:



2. Normed linear spaces - vector spaces with norms

If V $=$ inner product space (i.e. vector space with inner product defined), recall norm of a vector is

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}.$$

Easy to show norm has **3 properties** which follow from those of inner product:

(a) $\|\mathbf{v}\| \geq 0$; $\|\mathbf{v}\| = 0$ iff (*if and only if*) $\mathbf{v} = 0$.
(b) $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$
(c) $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$ if $a \in \mathbb{R}$

Normed spaces

Ex. 2: Let $R = [0,1]$. Consider vector space

$$V = C(R) = \text{all continuous functions on } R,$$

For $f \in V$ let

$$\|f\|_\infty = \max_{x \in R} |f(x)|.$$

Easy to check $\|f\|_\infty$ satisfies properties of norm (exercises).

Norm $\|\mathbf{v}\|$ represents *length* of vector $\mathbf{v}$.

Normed spaces

Even if inner product *not* defined, *any* assignment of length $\|\mathbf{v}\|$ to all vectors **v,** which satisfies properties (a) - (c) is called a *norm.*

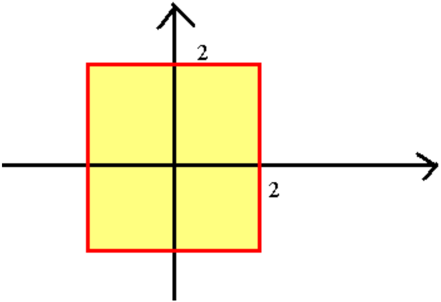
Def. 5: A vector space V is a *normed linear space (NLS)* if for all **v** $\in V$, there is a norm (length) $\|\mathbf{v}\|$ which satisfies **(a) - (c)**.

i.e., V is an NLS if we have notion of length on it

Normed spaces

Ex. 3: L^p norms: if R is the box

$$R = [-2,2] \times [-2,2] \equiv \{(x,y) : \ |x|, |y| \le 2\},$$



(or any other closed bounded subset of $\mathbb{R}^d$).

Normed spaces

Let $\mathbf{x} = (x, y)$ and $d\mathbf{x} = dx\,dy$.

Let $f(x, y) = f(\mathbf{x})$ be a function. Define norm

$$\|f\|_p = \left(\int_R |f(\mathbf{x})|^p d\mathbf{x}\right)^{1/p} \equiv \left(\int_R |f(x, y)|^p dx dy\right)^{1/p}.$$

Can verify this has the properties of a norm (exercises).
We define vector space of functions

$$L^p(R) = \{f : \|f\|_p < \infty\}.$$

Can show this is vector space (i.e., closed under addition and sclalar mult) and a NLS (i.e. $\|f\|_p$ is a norm).

4. Preliminaries

Def. 6. A $n \times n$ matrix M is symmetric if $M_{ij} = M_{ji}$ for all i, j, i.e. is unchanged if reflected about its diagonal.

A matrix M is *positive* if all of its eigenvalues are non-negative.

Equivalently if $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix}$ and $\mathbf{a}^T = [a_1, \ldots, a_d]$, and $\langle \cdot \, , \cdot \rangle$ is standard dot product on $\mathbb{R}^d$, then

$$\langle \mathbf{a}, M\mathbf{a} \rangle \equiv \mathbf{a}^T M \mathbf{a} \geq 0$$

for all $\mathbf{a}$.

5. Reproducing Kernel Hilbert spaces:

Let $R \subseteq \mathbb{R}^d$ be a closed bounded set (e.g. set of possible microarrays **x**).

Let $\mathcal{H}$ be any complete vector space of (classification) functions on R with inner product $\langle f, g \rangle$ defined (recall this makes $\mathcal{H}$ a Hilbert space).

Note this also defines a norm for $f \in \mathcal{H}$:

$$\|f\| = \sqrt{\langle f, f \rangle}$$

RKHS

Motivation: recall we want to find function $f(\mathbf{x})$ which classifies microarrays **x** correctly.

Recall penalty $L(f) = \|f\|^2$, penalizing, e.g. for non-smoothness of f.

The norm $\|f\|$ comes from inner product on some vector space $\mathcal{H}$ of functions on domain R.

This vector space $\mathcal{H}$ (which gives desired penalty norm $\|f\|$) will be a *reproducing kernel Hilbert space.*

RKHS

Definition 7: We say $\mathcal{H}$ is a *reproducing kernel Hilbert space (RKHS)* if whenever we fix an $\mathbf{x} \in R$, then for all functions $f \in \mathcal{H}$

$$|f(\mathbf{x})| \leq C\|f\|$$

for a fixed constant C.

RKHS

Definition 8: A *kernel function* is a function $K(\cdot,\cdot)$ on pairs $\mathbf{x},\mathbf{y} \in R$ which is symmetric, i.e.,

$$K(\mathbf{x},\mathbf{y}) = K(\mathbf{y},\mathbf{x}).$$

and *positive*, i.e. for any fixed collection $\mathbf{x}_1,\ldots,\mathbf{x}_n$ the $n \times n$ matrix

$$\mathbf{K} = (\mathbf{K}_{ij}) \equiv K(\mathbf{x}_i,\mathbf{x}_j)$$

is positive.

K determines $\mathcal{H}$

6. The kernel function $K(\mathbf{x}, \mathbf{y})$ uniquely corresponds to the space $\mathcal{H}$

Theorem 1: *Given an RKHS* $\mathcal{H}$ *of functions on* $R \subset \mathbb{R}^d$, *there exists a unique kernel function* $K(\mathbf{x}, \mathbf{y})$ *such that for all* $f \in \mathcal{H}$,

$$f(\mathbf{x}) = \langle f(\,\cdot\,), K(\,\cdot\,, \mathbf{x})\rangle_{\mathcal{H}}$$

(inner product above is in the variable $\cdot$; **x** is fixed).

Note this means that evaluation of f at **x** is equivalent to taking inner product of f with the fixed function $K(\,\cdot\,, \mathbf{x})$,

K determines $\mathcal{H}$

i.e. $f(\mathbf{x})$ is *reproduced* by using K

We call $K(\mathbf{x}, \mathbf{y})$ the *reproducing kernel* of the space $\mathcal{H}$ of functions.

K determines $\mathcal{H}$

Definition 9: We call the above kernel function $K(\mathbf{x}, \mathbf{y})$ the *reproducing kernel* of the function space $\mathcal{H}$.

Definition 10: A *continuous kernel* is a kernel function $K(\mathbf{x}, \mathbf{y})$ which is also continuous as a function of **x** and **y**.

Recall for continuous function $f(\mathbf{x})$ on R we define

$$\|f\|_\infty = \max_{x \in R} |f(\mathbf{x})|.$$

K determines $\mathcal{H}$

Theorem 2:

(i) For every continuous kernel $K(\cdot, \cdot)$ on R, *there exists a unique RKHS* $\mathcal{H}$ *of functions on* R *such that* K *is its reproducing kernel.*

(ii) Moreover, this $\mathcal{H}$ *consists of continuous functions, and for any* $f \in \mathcal{H}$

$$\|f\|_\infty \leq M_K \|f\|_{\mathcal{H}},$$

where $M_K = \max_{\mathbf{x},\mathbf{y} \in X} \sqrt{K(\mathbf{x}, \mathbf{x})}$.

7. Support vector machines

Recall the *regularization setting*:

Wish to separate classes $\mathcal{C}$ and $\sim\mathcal{C}$ (e.g. cancerous and non-cancerous microarrays)

Have n examples

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\},$$

with feature vector (e.g. microarray) $\mathbf{x}_i \in \mathbb{R}^d$, class $y_i \in \mathbb{B} = \{\pm 1\}$.

SVM

Thus y_i tells whether $\mathbf{x}_i$ is in class $\mathcal{C}$.

Want to find function $f : \mathbb{R}^d \rightarrow \mathbb{B}$ which *generalizes* above data so $f(\mathbf{x}) = y$ can predict class y of novel feature vector $\mathbf{x}$.

In fact we want something more general: function $f(\mathbf{x})$ which will best help us decide the true value of y.

SVM

It may not need to be that we want $f(\mathbf{x}) = y$, but rather we want

$$\begin{cases} f(\mathbf{x}) >> 1 & \text{if } y = 1 \\ f(\mathbf{x}) << 1 & \text{if } y = -1 \end{cases}, \tag{2}$$

i.e., $f(\mathbf{x})$ is large and positive if the correct answer is $y = 1$ (e.g. cancerous) and $f(\mathbf{x})$ is large and negative if the correct answer is $y = -1$ (not cancerous).

Note the larger $f(\mathbf{x})$ is the more certain we are that class $y = 1$.

SVM

Decision rule: conclude whether $y = \pm 1$ based on rule (2).

How to choose the best f?

Need f which works correctly on known samples $D = \{\mathbf{x}_i, y_i\}$ and which is *reasonable*, i.e., satisfies some a priori assumptions (e.g. smoothness).

SVM

Recall:

We can still choose best f by recalling *regularization setting:*

$$f = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2,$$

where $\|f\|_K =$ norm in an RKHS $\mathcal{H}$, e.g.,

$$\|f\|_K = \|Af\|_{L^2} = \int (Af)^2 dx$$

where $Af = \frac{d^2}{dx^2} f - f$ as earlier.

(note $\|f\|_K \equiv \|f\|_{\mathcal{H}}$).

SVM

How do we measure error between $f(\mathbf{x})$ and y?

Hinge function V:

$$V(f(\mathbf{x}), y) = (1 - yf(\mathbf{x}))_+,$$

where

$$(a)_+ \equiv \max(a, 0).$$

(will discuss further)

The Representer Theorem

1. An application: using kernel spaces for regularization

Assume again we have unknown function $f(\mathbf{x})$ on R, with $\mathbf{x} = (x_1, \ldots, x_d)^T$ = microarray values.

Recall

if $f(\mathbf{x}) >> 1$ we are certain $\mathbf{x} \in \mathcal{C}$ (cancer)

if $f(\mathbf{x}) << -1$ we are certain $\mathbf{x} \in \sim\mathcal{C}$ (no cancer)

Motivation: find the classifier $f(\mathbf{x})$

Assume $f \in \mathcal{H} \;=$ vector space of functions on R (more specifically an RKHS with kernel function $K(\mathbf{x}, \mathbf{y})$)

Our data Df

$$Df = \mathbf{y} \equiv (y_1, \ldots, y_n) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$$

$=$ correct classification of samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}_{i=1}^{n}$

Motivation: find the classifier $f(\mathbf{x})$

To find best choice $f = f_0$, approximate it by finding the minimizer

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) + \lambda \| f \|_{\mathcal{H}}^2 \right\}. \tag{1}$$

where $\lambda =$ constant.

Motivation: find the classifier $f(\mathbf{x})$

Note we are finding an f which balances minimizing

$$\text{data error} = \frac{1}{n}\sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2,$$

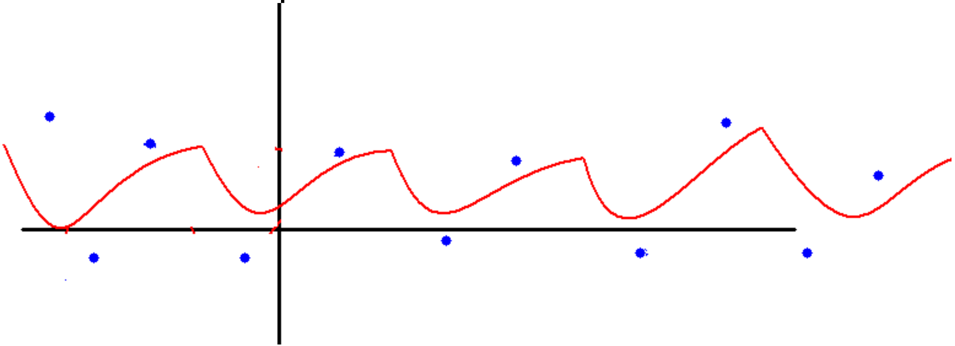
with minimizing

$$L(f) = \|f\|_{\mathcal{H}}^2,$$

i.e., penalty for lack of smoothness.

Motivation: find the classifier $f(\mathbf{x})$

Solution to such a problem will look like:



Will compromise between fitting data (which may have error) and trying to be smooth.

Motivation: find the classifier $f(\mathbf{x})$

A remarkable fact: best choice $\widehat{f}$ can be found *explicitly* using the *reproducing kernel function* $K(\mathbf{x}, \mathbf{y})$ of space $\mathcal{H}$ of allowed choices of f.

2. Solving the minimization

Consider optimization problem (1).

Claim we can solve it explicitly.

Recall want to find

$$f_1 = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2. \tag{1}$$

Note we can have, e.g., $V(f(\mathbf{x}_i), y_i) = (f(\mathbf{x}_i) - y_i)^2$.

Representer theorem

We have the

Representer Theorem: *The solution of the Tikhonov optimization problem* (1) *can be written*

$$f(x) = \sum_{i=1}^{n} a_i K(\mathbf{x}, \mathbf{x}_i), \tag{2}$$

where $\mathbf{x}_i$ *are the examples and* $K(\mathbf{x}, \mathbf{y})$ *is the reproducing kernel of the RKHS* $\mathcal{H}$.

Important theorem: we only need to find n numbers a_i to solve the infinite dimensional problem (1) above.

3. Matrix formulation

Considering again the case where we have information

$$Df = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) = \mathbf{y},$$

We want to find

$$f_1 = \arg\inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \tag{1}$$

Plugging universal solution

Matrix formulation

$$f(\mathbf{x}) = \sum_{j=1}^{n} a_j K(\mathbf{x}, \mathbf{x}_j)$$

into (1) we get:

$$f_1 = \arg\inf_{a_1,\ldots,a_n} \frac{1}{n} \sum_{i=1}^{n} V\left(\sum_{j=1}^{n} a_j K(\mathbf{x}, \mathbf{x}_j), y_j\right) + \lambda \left\| \sum_{j=1}^{n} a_j K(\mathbf{x}, \mathbf{x}_i) \right\|_{\mathcal{H}}^2$$

(1)

Matrix formulation

Note

$$\left\|\sum_{j=1}^{n} a_j K(\mathbf{x}, \mathbf{x}_j)\right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} a_i a_j K_{ij} = \mathbf{a}^T K \mathbf{a}.$$

where $\mathbf{K} = (K_{ij}) = (K(\mathbf{x}_i, \mathbf{x}_j))$, and $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$.

Matrix formulation

Thus

$$f_0 = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} V\left(\sum_{j=1}^{n} a_j K(\mathbf{x}_i, \mathbf{x}_j), y_i\right) + \lambda \mathbf{a}^T \mathbf{K} \mathbf{a}.$$

This now minimizes over $\mathbf{a} = [a_1, \ldots, a_n]^T$ and is now n-dimensional minimization problem.

Can take derivatives wrt a_i and set equal to 0.

Matrix formulation

Special case: $V(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2.$

Here

$$\mathbf{a} = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_j K(\mathbf{x}_i, \mathbf{x}_j) - y_i\right)^2 + \lambda \mathbf{a}^T K \mathbf{a}$$

$$= \arg\min_{\mathbf{a}\in\mathbb{R}^n} \frac{1}{n}(\mathbf{K}\mathbf{a} - \mathbf{y})^2 + \lambda \mathbf{a}^T \mathbf{K}\mathbf{a}.$$

where $\mathbf{a} = [a_1, \ldots, a_n]^T$ and $\mathbf{y} = [y_1, \ldots, y_n]^T$ (known classes of examples $\mathbf{x}_i$).

Matrix formulation

Take the gradient with respect to $\mathbf{a}$ and setting to 0 we get:

$$0 = \frac{2}{n} K(K\mathbf{a} - \mathbf{y}) + 2\lambda K\mathbf{a} = \left(\frac{2K^2}{n} + 2\lambda K\right)\mathbf{a} - \frac{2}{n} K\mathbf{y}.$$

$$= 2K\left(\frac{K}{n} + \lambda\right)\mathbf{a} - 2K\frac{\mathbf{y}}{n}.$$

Matrix formulation

Thus if K is nonsingular:

$$\mathbf{a} = \left(\frac{\mathbf{K}}{n} + \lambda\right)^{-1} \left(\frac{\mathbf{y}}{n}\right) = (\mathbf{K} + \lambda n \mathbf{I})^{-1}\mathbf{y}.$$

where $\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix}$ = identity matrix.

Explicit solution.

Matrix formulation

Thus

$$f_1(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} a_i K(\mathbf{x}, \mathbf{x}_i)$$

= sum of kernel functions.