Lecture 11C (Optional).

MA 751 Part 6
Support Vector Machines

3. An example: Gene expression arrays

Assume we are given a tissue sample s, and a *feature vector*

$$\mathbf{x} = \Phi(s) \in \mathbb{R}^{30,000}$$

consisting of 30,000 gene expression levels as read by a gene expression array.

We wish to determine whether the tissue is cancerous or not.

For an **x** which in fact corresponds to cancerous tissue, we will set the corresponding output variable $y = 1$; otherwise $y = -1$.

Consider a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, consisting of pairs of feature vectors $\mathbf{x}_i$ and corresponding (correct) diagnoses $y_i \in \{-1, 1\}$.

Can we find the right function $f_1 : F \longrightarrow \mathcal{B}$ which generalizes the above examples, so that $f_1(\mathbf{x}) = y$ for all feature vectors?

Easier (see below):

Find a $f : F \longrightarrow \mathbb{R}$, where

$f(\mathbf{x}) > 0$ if $f_1(\mathbf{x}) = 1$; $f(\mathbf{x}) < 0$ if $f_1(\mathbf{x}) = -1$.

4. Support vector machine framework

Recall the *regularization setting*: we have n examples

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

with $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{B} = \{\pm 1\}$.

As mentioned above, we want to find a function $f_1 : \mathbb{R}^d \longrightarrow \mathbb{B}$ which *generalizes* the above data so that $f(\mathbf{x}) = y$ generalizes the data D.

As mentioned there, we will actually want here something more general: a function $f(\mathbf{x})$ which will best help us decide the true value of y.

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

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

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

Then the decision rule will be to conclude the value of y based on the rule (2). This is made precise as follows. We have the following

optimization criterion for the 'right' f:

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

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

$$\|f\|_K = \|Af\|_{L^2} = \int_{\mathbb{R}^n} (Af)^2 dx.$$

Above 'arg min' denotes the f which minimizes the above expression.

Loss function: hinge loss

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

where

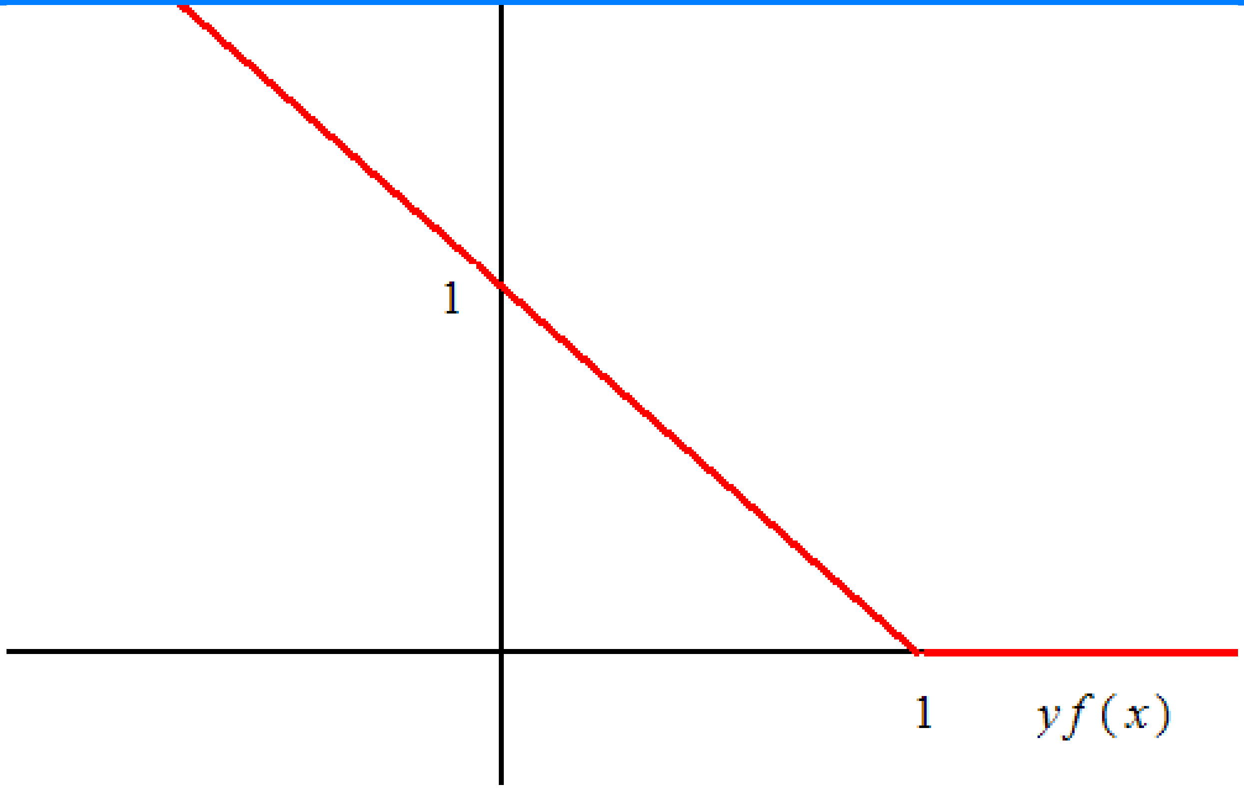
$$(a)_+ \equiv \max(a, 0).$$

5. More about the hinge loss

Consider the error function

$$L(f(\mathbf{x}), y) = (1 - yf(\mathbf{x}))_+ \equiv \text{max}(1 - yf(\mathbf{x}), 0).$$

$$= \begin{cases} \text{small} & \text{if } y, f(\mathbf{x}) \text{ have same sign} \\ \text{large} & \text{otherwise} \end{cases}.$$



This is called the *hinge loss function*.

[Notice *margin* built in: error 0 only if $yf(\mathbf{x}) \geq 1$ (more stringent requirement than just $yf(\mathbf{x}) \geq 0)$]

Thus data error is

$$e_d = \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{x}_j), y_j)$$

What is a priori information?

Note surface $H : f = 0$ will separate "positive" **x** with $f(\mathbf{x}) > 0$, and "negative" **x** with $f(\mathbf{x}) < 0$:

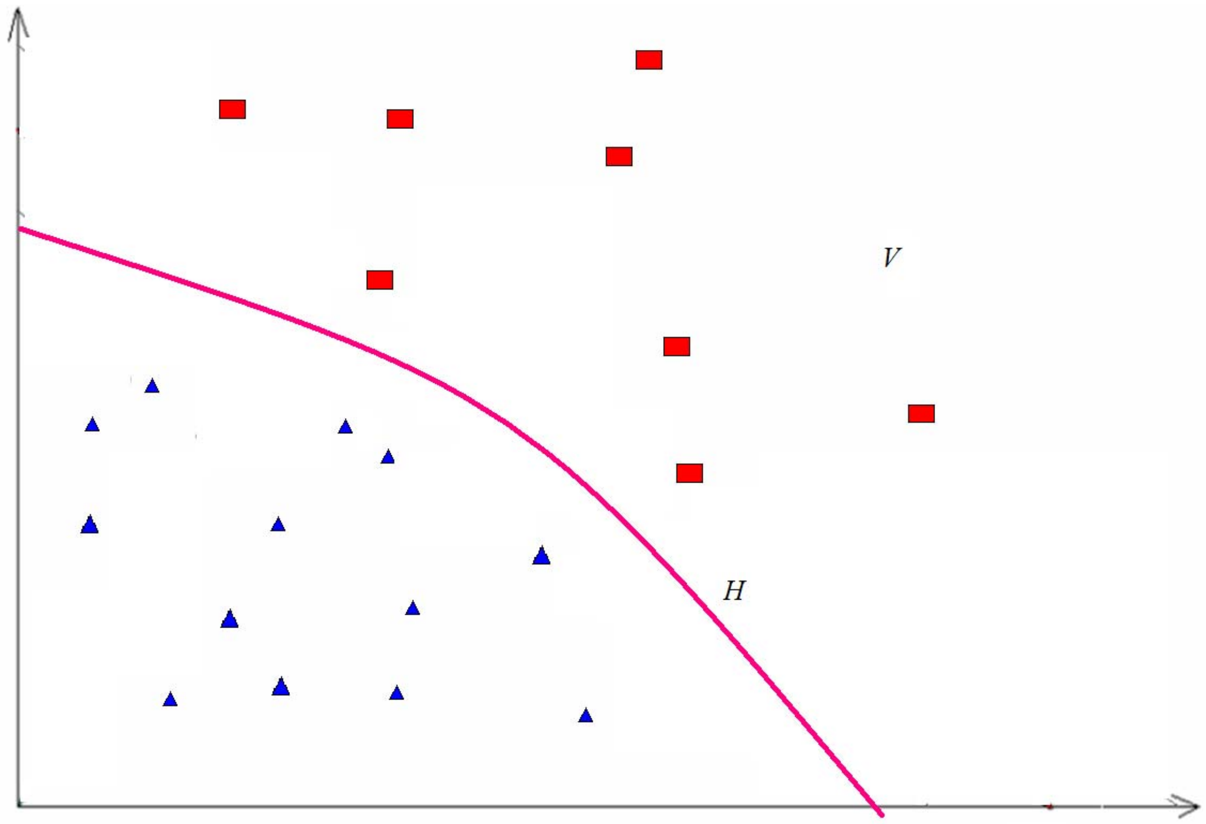


Fig. 1. Red points have $y = +1$ and blue have $y = -1$ in the space F. $H : f(\mathbf{x}) = 0$ is the separating surface.

Assume some a priori information defined in terms of an RKHS norm $\|\cdot\|_K$ so $\|f\|_K$ is small if a priori assumption is satisfied.

Let $\mathcal{H}$ be corresponding RHKS.

Will specify desirable norm $\|\cdot\|_K$ later...

Now solve regularization problem for the above norm and loss V:

$$f_0 = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{j=1}^{n} (1 - y_j f(\mathbf{x}_j))_+ + \lambda \|f\|_K^2. \quad (1)$$

6. Introduction of slack variables

Define new variables ξ_j, and note if we find the min over $f \in \mathcal{H}$ and ξ_j of

$$\underset{f \in \mathcal{H},\, \xi_j}{\arg\min} \frac{1}{n} \sum_{j=1}^{n} \xi_j + \lambda \| f \|_K^2 \qquad (1a)$$

with the constraint

$$y_j f(\mathbf{x}_j) \geq 1 - \xi_j$$

$$\xi_j \geq 0,$$

we get the same solution f.

`

To see this, note the constraints are

$$\xi_j \geq \max\left(0, 1 - y_j f(\mathbf{x}_j)\right) = \left(1 - y_j f(\mathbf{x}_j)\right)_+ \quad \text{(1b)}$$

which yields the claim.

(Clearly in fact in minimizing sum we will end up with $\xi_j = \left(1 - y_j f(\mathbf{x}_j)\right)_+$).

From form (1) above by representer theorem:

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

To find $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ (see above material): let

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

Then

$$\|f\|_K^2 = \left\langle \sum_j a_j K(\mathbf{x}, \mathbf{x}_j), \sum_i a_i K(\mathbf{x}, \mathbf{x}_i) \right\rangle$$

$$= \sum_{i,j} a_i a_j \langle K(\mathbf{x}, \mathbf{x}_j), K(\mathbf{x}, \mathbf{x}_i)\rangle$$

$$= \sum_{i,j} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j} a_i a_j K_{ij} = \mathbf{a}^T K \mathbf{a}.$$

Thus:

$$\mathbf{a} = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \frac{1}{n}\sum_{j=1}^{n}\xi_j + \lambda \mathbf{a}^T K \mathbf{a} \qquad (2a)$$

with constraint:

$$y_j \sum_{i=1}^{n} a_i K(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_j \qquad (2b)$$

$$\xi_j \geq 0. \qquad (2c)$$

5. Bias

Given choice of $\mathcal{H}$, K we have concluded

$$f(\mathbf{x}) = \sum_{j=1}^{n} a_j K(\mathbf{x}, \mathbf{x}_j) \qquad (3)$$

which optimizes (1), equivalently (2).

Now can *expand* class (2) of allowable f *ad hoc*. We may feel larger class than $\mathcal{H}$ is appropriate.

Often adding a constant b is useful.

Thus change $f(\mathbf{x})$ by adding a bias term b:

$$f(\mathbf{x}) = \sum_{j=1}^{n} a_j K(\mathbf{x}, \mathbf{x}_j) + b. \tag{4}$$

The effect: regularization term unchanged (i.e., we ignore b in the norm $\|f\|_K$; remember any a priori assumption is valid if it is useful).

Note this is still a norm on the expanded space of functions of the form (4), but may not be positive definite, i.e., $\|f\| = 0$ for some f of the form (4).

For example we may have $\|b\|_K = 0$.

But: minimization of (1) using (4) still makes sense and allows possibly richer set of functions than $\mathcal{H}$, as long as the regularization term $\|f\|_K$ still makes sense for such a richer set.

In terms of slack variables ξ_i, new optimization problem:

Find $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ which minimizes:

$$\frac{1}{n}\sum_{j=1}^{n}\xi_j + \lambda \mathbf{a}^T K \mathbf{a}$$

with constraints:

$$y_j\left(\sum_{i=1}^{n} a_i K(\mathbf{x}_i, \mathbf{x}_j) + b\right) \geq 1 - \xi_j \quad \text{(4a)}$$

$$\xi_i \geq 0$$

(*quadratic programming* problem).