1 Support Vector Machines

We have a set of points $\{\mathbf{x}_i\}_{i=1}^{l}, \mathbf{x}_i \in \mathbb{R}^d$ with associated labels $\{y_i\}$, where $y_i \in \{-1, 1\}$. The classification labels $\{-1, 1\}$ simplify some of the math.

We wish to find a hyperplane in $\mathbb{R}^d$ which separates the set $\mathcal{X}_1 = \{\mathbf{x}_i | y_i = 1\}$ from the set $\mathcal{X}_0 = \{\mathbf{x}_i | y_i = -1\}$, and which will generalize to new data in some optimal way. Such a plane can be defined as the set $\mathcal{P}_{\mathbf{w},b} = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ for some vector $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$. Our problem will boil down to finding $\{\mathbf{w}, b\}$ that optimize a certain metric of "generelizeability" to new data points.

1.1 Separable Case

In this case, at least one separating hyperplane $\{\mathbf{w}, b\}$ exists. All $x \in \mathcal{X}_0$ will lie on one side of the plane and all $x \in \mathcal{X}_1$ will lie on the other.

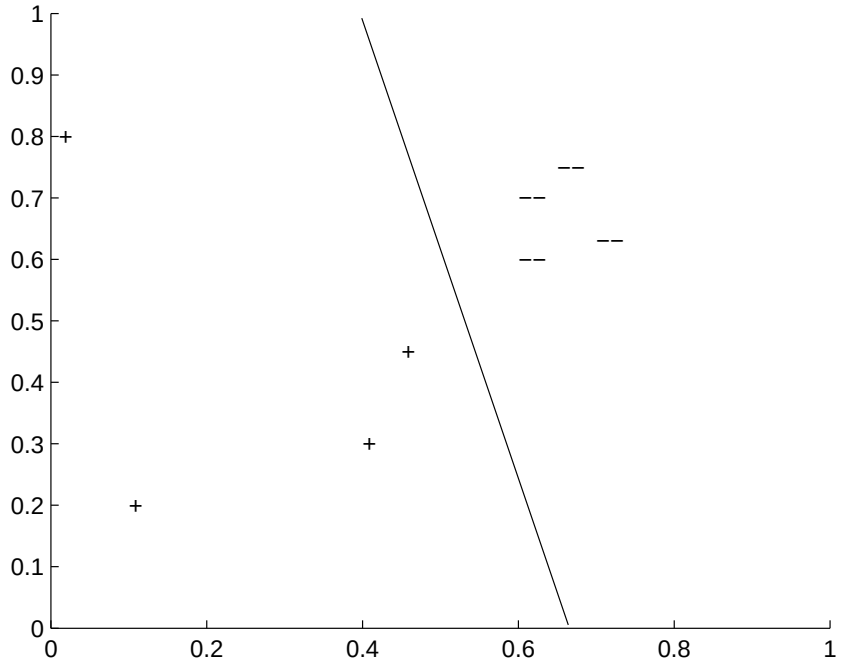

Figure 1: Separating Hyperplane

1.1.1 Shortest distance to origin

The perpendicular (shortest) distance from some plane $\mathcal{P}_{\mathbf{w},b} = \{\mathbf{x} | \langle \mathbf{x}, \mathbf{w} \rangle + b = 0\}$ to the origin can be calculated.

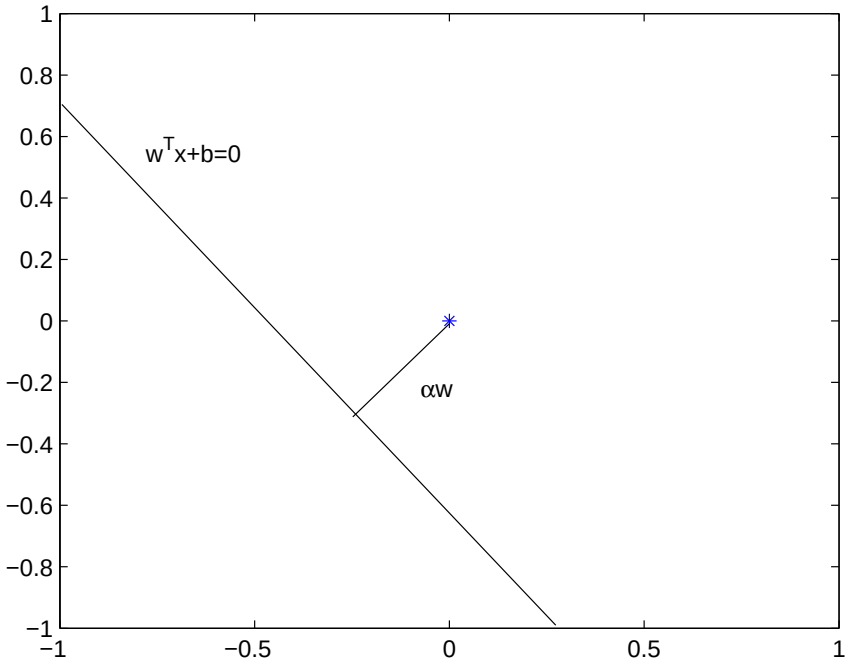

Figure 2: Shortest distance from $\mathcal{P}_{\mathbf{w},b}$ to the origin

The vector from the origin to $\mathcal{P}_{\mathbf{w},b}$ which is perpendicular to $\mathcal{P}_{\mathbf{w},b}$ must be of the form $\mathbf{s} = \alpha\mathbf{w}$. Furthermore, $\alpha\mathbf{w} \in \mathcal{P}_{\mathbf{w},b}$ because it sits on the plane. Thus

$$\langle \mathbf{s}, \mathbf{w} \rangle + b = \langle \alpha\mathbf{w}, \mathbf{w} \rangle + b = 0 \Rightarrow \alpha = \frac{-b}{\|\mathbf{w}\|^2}$$

and the shortest distance from $\mathcal{P}_{\mathbf{w},b}$ to the origin,

$$\|\alpha\mathbf{w}\| = \frac{|b|}{\|\mathbf{w}\|}$$

Thus for fixed $\mathbf{w}$, we see that b is a linear measure of distance from the hyperplane to the origin.

1.1.2 Partitioning $\mathbb{R}^d$ and distance to the hyperplane

Suppose that for some plane $\mathcal{P}_{\mathbf{w},b} = \{\mathbf{x} | \langle \mathbf{x}, \mathbf{w} \rangle + b = 0\}$ that partitions $\mathbb{R}^d$ into two half spaces, we have $b > 0$. Then for some point $\mathbf{y} \in \mathbb{R}^d$, the test

$$\langle \mathbf{y}, \mathbf{w} \rangle + b > 0$$

is a test to see whether y is in the half-space with the origin (to see this, simply set $\mathbf{y} = 0$ to get the test $b > 0$). For $b < 0$ the test is reversed.

Now consider some point $\mathbf{y}$ such that

$$\langle \mathbf{y}, \mathbf{w} \rangle + b = \epsilon > 0$$

we can see that ϵ measures the perpendicular distance from that point to $\mathcal{P}_{\mathbf{w},b}$ in the following way. $\mathbf{y}$ sits on $\mathcal{P}_{\mathbf{w},b-\epsilon}$.

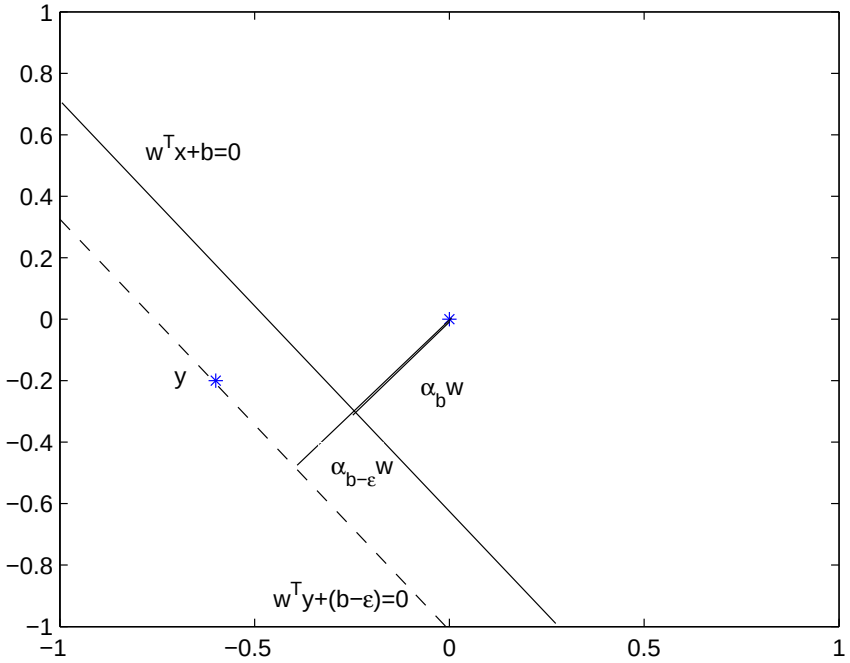

Figure 3: Distance from point $\mathbf{y}$ to $\mathcal{P}_{\mathbf{w},b}$

We can now see that the distance from $\mathcal{P}_{\mathbf{w},b}$ to the origin is $\frac{|b|}{\|\mathbf{w}\|}$ and the distance from $\mathcal{P}_{\mathbf{w},b-\epsilon}$ to the origin is $\frac{|b-\epsilon|}{\|\mathbf{w}\|}$ (in the image above, $b < 0$). Any point $\mathbf{y}$ for which

$$\langle \mathbf{w}, \mathbf{y} \rangle + b = \epsilon > 0$$

is a distance of $\frac{\epsilon}{\|\mathbf{w}\|}$ away from its closest point on $\mathcal{P}_{\mathbf{w},b}$.

Now suppose that we have a set of points $\{\mathbf{y}_i\}_{i=1}^{l}$ all on one side of a hyperplane $\mathcal{P}_{\mathbf{w},b}$. Then we know that the following holds:

$$\begin{aligned} \langle \mathbf{y}_1, \mathbf{w} \rangle + b &= \epsilon_1 > 0 \\ &\vdots \\ \langle \mathbf{y}_l, \mathbf{w} \rangle + b &= \epsilon_l > 0 \end{aligned}$$

(we have arbitrarily chosen $\epsilon_i > 0$). From the above discussion, we can choose the $\mathbf{y}^*$, the closest point to $\mathcal{P}_{\mathbf{w},b}$, by finding $\epsilon^* = \arg\min_i |\epsilon_i|$ (more than one may exist; choose one). Then we can normalize the equations above; e.g., we will have

$$\begin{gathered}\left\langle \mathbf{y}_1, \underbrace{\frac{\mathbf{w}}{\epsilon^*}}_{\bar{\mathbf{w}}} \right\rangle + \underbrace{\frac{b}{\epsilon^*}}_{\bar{b}} = \tfrac{\epsilon_1}{\epsilon^*} \geq 1 \\ \vdots \\ \langle \mathbf{y}^*, \bar{\mathbf{w}} \rangle + \bar{b} = 1 \\ \vdots \\ \langle \mathbf{y}_l, \bar{\mathbf{w}} \rangle + \bar{b} \geq 1\end{gathered}$$

The hyperplanes induced by $\mathbf{y}^*$, these "closest" hyperplanes to our separating plane $\mathcal{P}_{\mathbf{w},b}$ are called support vectors.

1.1.3 A separating hyperplane that maximizes margins

Going back to the separable case, where we have points $\mathcal{X} = \mathcal{X}_0 \sqcup \mathcal{X}_1$ which can be separated by hyperplanes of the form $\mathcal{P}_{\mathbf{w},b}$. Then we arbitrarily choose sides:

$$\begin{aligned}\langle \mathbf{x}_i, \mathbf{w} \rangle + b \geq 1 \quad & \text{for } \mathbf{x}_i \in \mathcal{X}_1 \quad (y_i = +1) \\ \langle \mathbf{x}_i, \mathbf{w} \rangle + b \leq -1 \quad & \text{for } \mathbf{x}_i \in \mathcal{X}_0 \quad (y_i = -1)\end{aligned}$$

we can combine these restrictions on $\{\mathbf{w}, b\}$ into one set of inequalities:

$$y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) - 1 \geq 0 \quad \forall i$$

For any such separating plane $\mathcal{P}_{\mathbf{w},b}$, we will have some closest point or group of points (in terms of perpendicular distance to $\mathcal{P}_{\mathbf{w},b}$) on both sides of $\mathcal{P}_{\mathbf{w},b}$; these points will lie on the support vectors of $\mathcal{P}_{\mathbf{w},b}$.

The closest $\mathbf{x}_i \in \mathcal{X}_0$, for which $\langle \mathbf{x}^*, w \rangle + b = -1$ will sit on a support vector for which the perpendicular distance to the origin is $\frac{|b+1|}{\|\mathbf{w}\|}$. Similarly the closest $\mathbf{x}_i \in \mathcal{X}_1$, for which $\langle \mathbf{x}^*, w \rangle + b = 1$ will sit on a support vector with distance $\frac{|b-1|}{\|\mathbf{w}\|}$ to the origin.

Both of these support vectors will be distance $\frac{1}{\|\mathbf{w}\|}$ from $\mathcal{P}_{\mathbf{w},b}$. We call the sum of these distances, $\frac{2}{\|\mathbf{w}\|}$, the margin of $\mathcal{P}_{\mathbf{w},b}$, and we wish to maximize this value. The optimization problem thus becomes

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1 \geq 0 \quad \forall i \end{aligned}$$

note that this is a convex criterion in a convex space.

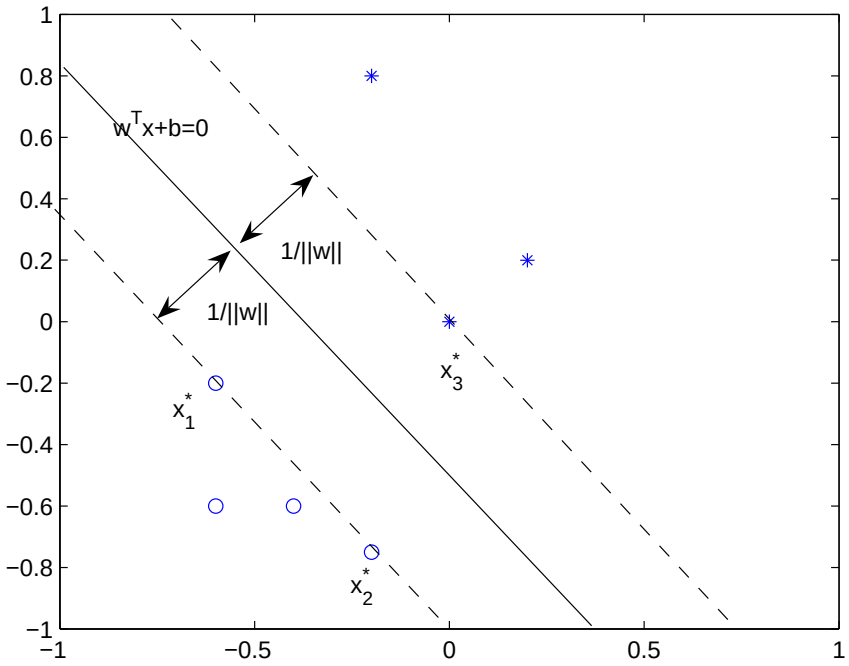

Figure 4: Separating hyperplane for $\mathcal{X}_0$ and $\mathcal{X}_1$ and support vectors

1.1.4 The Lagrangian and dual problem

We form the lagrangian

$$\mathcal{L}(\mathbf{w}, b, a) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i a_i \left(y_i(\mathbf{x}_i^T\mathbf{w} + b) - 1\right)$$

in which the $a_i \geq 0$. We know that $\nabla_{\mathbf{w},b}\mathcal{L} = 0$ and $\nabla_a\mathcal{L} = 0$. Solving the former, we get

$$\begin{aligned} \mathbf{w} &= \sum_i a_i y_i \mathbf{x}_i \\ \sum_i a_i y_i &= 0 \end{aligned}$$

Substituting these values back into the original problem, and using the Wolfe dual (strong dual) formulation, we get the following formulation:

$$\begin{array}{lll} \max\limits_{\alpha,b} & -\frac{1}{2}\alpha^T\mathbf{H}\alpha + 1^T\alpha & \\ \text{s.t.} & \alpha_i \geq 0 & \forall i \\ & \sum_i \alpha_i y_i = 0 & \forall i \end{array}$$

where $(\mathbf{H})_{ij} = y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Once we solve for α, we can plug it back into $\mathbf{w}$ and calculate b using the support vector points, i.e. where $y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) = 1$.

Note that the data, $\{\mathbf{x}_i\}$ only appear as inner products in the optimization, which allows us to replace inner products of the form $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with some kernel function $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathbb{R}^?} = k(\mathbf{x}_i, \mathbf{x}_j)$ for some $k(\cdot, \cdot)$ that we define.

1.2 Nonseparable case

In this case, it may not be possible to find a hyperplane that separates $\mathcal{X}_0$ and $\mathcal{X}_1$ into disjoint regions. In this case, the optimization problem above will have no solution.

We want to allow bad outliers across $\mathcal{P}_{\mathbf{w},b}$ (by loosening our constraints), but we want to penalize this behavior (by adding a parameter to the cost function).

1.2.1 Easing constraints

We want to introduce slack variables, for each data point we ease the constraints by adding $\xi_i \geq 0$. The new constraints become: $y_i(\langle \mathbf{x_i}, w \rangle + b) \geq 1 - \xi_i$.

Consider a point for which $\xi_i > 1$. Then $y_i(\langle \mathbf{x_i}, w \rangle + b) < 0$, and the point is on the wrong side of the hyperplane. This is what we want to allow, but only when absolutely necessary; thus we must penalize it.

1.2.2 Penalizing the slack variables

In the separable case, we wanted to maximize the margin by minimizing $\frac{1}{2}\|\mathbf{w}\|^2$. Now we want to add a term that minimizes a function which is monotonic in $\{\xi_i\}$ (we want to minimize the number of incorrectly labeled training data).

One function particularly well suited to the Wolfe dual formulation is $C\sum_i \xi_i$ for some $C \geq 0$ chosen by the user. The primal Lagrangian becomes

$$\mathcal{L}(\mathbf{w}, b, a) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i - \sum_i a_i \left(y_i(\mathbf{x}_i^T\mathbf{w} + b) - 1 + \xi_i\right) - \sum_i \mu_i\xi_i$$

with $\alpha_i \geq 0$ and $\mu_i \geq 0$.

1.2.3 The simplified dual formulation

The Wolfe dual formulation can be shown to be:

$$\begin{array}{lll} \max\limits_{\alpha,b} & -\frac{1}{2}\alpha^T\mathbf{H}\alpha + 1^T\alpha & \\ \text{s.t.} & 0 \leq \alpha_i \leq C & \forall i \\ & \sum_i \alpha_i y_i = 0 & \forall i \end{array}$$

from which $\mathbf{w}$ and b may be calculated as before. Note that the only difference is now an upper bound on the α_i's.

1.3 Kernel Trick

1.3.1 Motivating(?) Example

Consider our data space of sample points, $\{\mathbf{x}_i\} \in \mathbb{R}^d$. If $d = 2$, then distances in the data space are roots of $\|\mathbf{x} - \mathbf{y}\|^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2$. Suppose now that we wish to generalize this idea of distance so that we may include terms of the form $x_i y_j$. The usefulness of doing this will become apparent when we look at the applications of this idea to SVM. Suppose we choose the mapping

$$\Phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

Where we choose the regular inner (dot) product; we have $\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_2^2 = (x_1^2 - y_1^2)^2 + 2(x_1x_2 - y_1y_2)^2 + (x_2^2 - y_2^2)^2$. But notice; the image of Φ is a Hilbert space, and

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|_2^2 = \langle\Phi(\mathbf{x}), \Phi(\mathbf{x})\rangle - 2\langle\Phi(\mathbf{x}), \Phi(\mathbf{y})\rangle + \langle\Phi(\mathbf{y}), \Phi(\mathbf{y})\rangle$$

where it's easy to see that $\langle\Phi(\mathbf{x}), \Phi(\mathbf{y})\rangle = \langle\mathbf{x}, \mathbf{y}\rangle^2$ (this inner product taken in the data space $\mathbb{R}^2$). Denote $k_P(\mathbf{x}, \mathbf{y}) = \langle\mathbf{x}, \mathbf{y}\rangle^2$, and our distance calculation can be done only using k_P; we don't need to calculate anything using Φ.

Let's answer two questions:

- Can we find SVM decision rules ("linear" separating hyperplanes) in the new space without explicitly calculating Φ? And is this useful?

- If we have some "kernel" function $k(\mathbf{x}, \mathbf{y})$, under what conditions is this really an inner product in some Hilbert space? e.g., when is $k(\mathbf{x}, \mathbf{y}) = \sum_i \Phi_i(\mathbf{x})\Phi_j(\mathbf{y})$, for some $\Phi : \mathbb{R}^d \to \mathcal{H}$?

1.3.2 Revisiting SVM

We said last time that, using the dual Lagrangian formulation, we can solve for the separating hyperplane parameters $\{\mathbf{w}, b\}$ in the nonseparable case by solving the following quadratic optimization problem

$$\begin{array}{ll} \max\limits_{\alpha,b} & -\frac{1}{2}\alpha^T \mathbf{H}\alpha + 1^T\alpha \\ \text{s.t.} & 0 \le \alpha_i \le C \qquad \forall i \\ & \sum_i \alpha_i y_i = 0 \qquad \forall i \end{array}$$

where $\mathbf{H}_{ij} = y_i y_j \langle \mathbf{x_i}, \mathbf{x_j} \rangle$. $\{\mathbf{x_i}\}_i$ are our training observations, $y_i \in \{-1, 1\}$ are our observations, and C describes how many incorrect "crossovers" of the training data we wish to allow in the final hyperplane. We can then find $\mathbf{w}$ from the Lagrangian parameters α:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

and we can find b by using a point on one of the support vectors, i.e., where $y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + b) = 1$. This will maximize the margin $\frac{1}{\|\mathbf{w}\|}$ of our separating hyperplane (under certain slack conditions).

If we were to construct some new set of points in $\mathcal{H}$, $\{\Phi_i\} = \{\Phi(\mathbf{x_i})\}_i$, it's straightforward to see that we would solve a slightly different optimization problem (the derivation is identical, just with different data points):

$$\begin{array}{ll} \max\limits_{\alpha,b} & -\frac{1}{2}\alpha^T \tilde{\mathbf{H}}\alpha + 1^T\alpha \\ \text{s.t.} & 0 \le \alpha_i \le C \qquad \forall i \\ & \sum_i \alpha_i y_i = 0 \qquad \forall i \end{array}$$

where $\tilde{\mathbf{H}}_{ij} = y_i y_j k(\mathbf{x_i}, \mathbf{x_j}) = y_i y_j K_{ij}$; K_{ij} is the kernel gram matrix.

We now get $\mathbf{w} = \sum_i \alpha_i y_i \Phi(\mathbf{x}_i)$ (which we do not calculate explicitly); and we find b at a support vector point $\mathbf{x_i}$, where

$$y_i(\langle \Phi(\mathbf{x_i}), \mathbf{w} \rangle + b) = y_i \left(\left\langle \Phi(\mathbf{x_i}), \sum_j \alpha_j y_j \Phi(\mathbf{x}_j) \right\rangle + b \right) = y_i \left(\sum_j \alpha_j y_j k(\mathbf{x_i}, \mathbf{x_j}) + b \right) = 1$$

Finally, to classify a new datapoint $\mathbf{y}$, we find the

$$\operatorname{sgn}\left(\langle \Phi(\mathbf{y}), \mathbf{w} \rangle + b\right) = \operatorname{sgn}\left(\sum_j \alpha_j y_j k(\mathbf{y}, \mathbf{x_j}) + b \right)$$

1.3.3 Usefulness of the Kernel Trick

What does the preimage of the "decision rule" $\langle \Phi(\mathbf{y}), \mathbf{w} \rangle + b$ look like for data points $\mathbf{y} \in \mathbb{R}^2$? For $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle^3$ (all such "polynomial" kernels can be shown to be inner products in some space $\mathcal{H}$), here are two examples:

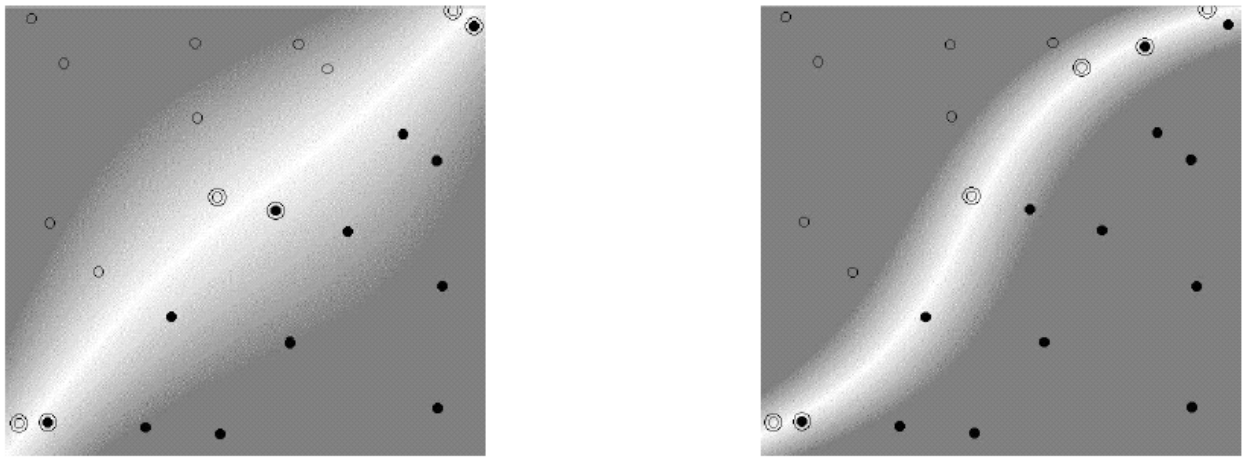

Figure 5: Degree 3 Polynomial Kernel k_P. The background color shows the shape of the decision surface. Ref: Burges 99

Notice the "support vector" points (under mapping Φ) are circled.

1.3.4 When is $k(\mathbf{x}, \mathbf{y})$ really an inner product?

Suppose we choose some kernel function k, e.g., $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$. When can we show that there exists some Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathbb{R}^d \to \mathcal{H}$ such that $k(\mathbf{x}, \mathbf{y}) = \sum_i \Phi(\mathbf{x})_i \Phi(\mathbf{y})_i$? Hilbert spaces allow infinite dimensional data points; under certain conditions of "niceness".

We can use Mercer's Theorem (Functional Analysis, Riesz / Integral Equations, Hochstadt) [I couldn't find a self-contained proof], which states:

Theorem Mercer's Theorem

Let X be a compact Hausdorff space with μ a countably additive measure on the Borel algebra of X. Suppose $K : X \times X \to \mathbb{R}$ be a bicontinuous symmetric nonnegative definite kernel on X; e.g., it is defined for all $(x, y) \in \mathcal{X}^2$; nonnegative definite means $\forall g \in \mathcal{L}^2_\mu(X)$,

$$\int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} K(x, y) g(x) g(y) d[\mu \otimes \mu](x, y) \geq 0$$

then $K(x, y)$, being a linear operator, has an eigenvalue decomposition. Further, there is an orthonormal basis $\{e_i\}_i$ of $L^2_\mu(X)$ consisting of eigenfunctions with an associated sequence of nonnegative eigenvalues. The eigenfunctions corresponding to the nonnegative eigenvalues are continuous. That is, we can write

$$K(x, y) = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(y)$$

where the convergence is absolute and uniform on X. By nonnegativity of the eigenvalues ($\lambda_i \geq 0$), can write $\Phi(x)_j = \sqrt{\lambda_j} e_j(x)$. This specific generalization of Mercer's theorem proves the actual nonnegativity of the eigenvalues, and the convergence properties of the eigenfunctions associated with them.

1.3.5 Heat Kernel SVM at work

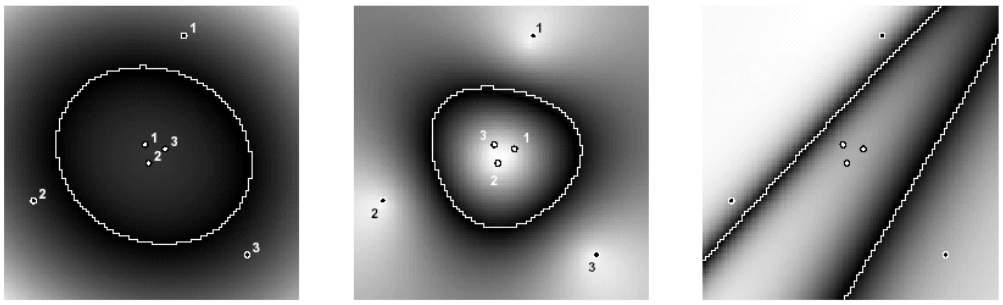

Figure 4: Decision boundaries for maximum margin classifiers with second order polynomial decision rule $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^2$ (left) and an exponential RBF $K(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||/2)$ (middle). The rightmost picture shows the decision boundary of a two layer neural network with two hidden units trained with backpropagation.

Figure 6: Source: Boser, Guyon, Vapnik