Chapter 8

Support Vector Machines

Support vector machines are universal tools in machine learning, where they are used for almost every task imaginable. Still the most prominent application for support vector machines is discriminant analysis. In discriminant analysis we are given labeled training data, where the label indicates to which class a datum belongs. The task is to compute a classifier from the labeled training data that allows to categorize new data, i.e., attach a label to it. In a first phase, we want to focus on geometric aspects of this task.

8.1 Maximum Margin Hyperplane

Here we study a version of the discriminant problem, where we assume that the data are points in the Euclidean space $\mathbb{E}^d$ and that there are only two classes P and Q with labels 1 and -1, respectively. We assume that $P \cup Q = \{x_1, \ldots, x_n\}$. At first we want to further assume that the two classes are linearly separable, i.e., that there exists a hyperplane that has all points with negative label strictly on one side and all points with positive label strictly on the other side. Such a hyperplane is given by two parameters, a unit normal $\tilde{w} \in \mathbb{E}^d$ and an offset $\tilde{b} \in \mathbb{R}$. Thus we have

$$\begin{aligned} \tilde{w}^T p_i - \tilde{b} &> 0, \quad p_i \in P \\ \tilde{w}^T p_i - \tilde{b} &< 0, \quad p_i \in Q. \end{aligned}$$

The classifier associated with a hyperplane $h = \{x \in \mathbb{E}^d \,|\, w^T x = b\}$ is the function $\operatorname{sign}(w^T x - b)$, where $\operatorname{sign}(z) = 1$ if $z > 0$, $\operatorname{sign}(z) = -1$ if $z < 0$ and $\operatorname{sign}(0) = 0$. Among hyperplanes that separate P and Q we are looking for one that has a *maximal margin*, i.e., a hyperplane that we can move the farthest in both directions along the normal $\tilde{w}$ before we meet a point from either P or Q. The naive intuition that the separation by such a hyperplane has good generalization properties, i.e., unseen data are likely to fall on the right side of the hyperplane, can be made more precise in statistical learning theory. Here we simply postulate that it is desirable to have a separating hyperplane with large margin, see also Figure 8.1. In order to compute the margin of a hyperplane

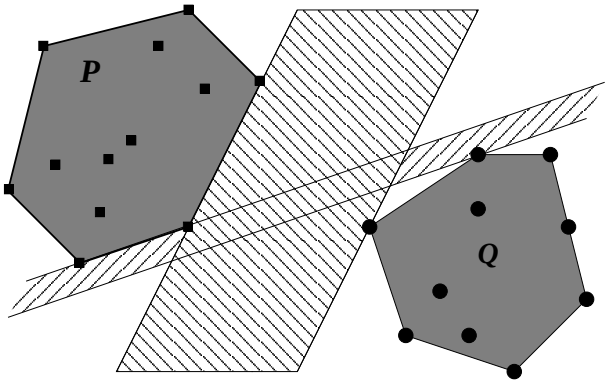

Figure 8.1: A large (good) and and a small (bad) margin.

let

$$c := \min_{p_i \in P \cup Q} |\tilde{w}^T p_i - \tilde{b}| > 0.$$

That is,

$$\begin{array}{rcll} \tilde{w}^T p_i - \tilde{b} & \geq & c, & p_i \in P \\ \tilde{w}^T p_i - \tilde{b} & \leq & -c, & p_i \in Q. \end{array}$$

By scaling the normal $\tilde{w}$ and the offset $\tilde{b}$ by $1/|c|$ we get

$$\begin{array}{rcll} w^T p_i - b & \geq & 1, & p_i \in P \\ w^T p_i - b & \leq & -1, & p_i \in Q, \end{array}$$

where $w := \tilde{w}/|c|$ and $b := \tilde{b}/|c|$. The margin of the hyperplane described by $\tilde{w}$ and $\tilde{b}$ is defined as the distance between the hyperplanes $h = \{x \in \mathbb{E}^d \,|\, w^T x = b + 1\}$ and $h' = \{x \in \mathbb{E}^d \,|\, w^T x = b - 1\}$. From the material in the introductory chapter, it is easy to deduce that this distance is $2/\|w\|$. In order to maximize the margin we can therefore minimize $\|w\|/2$. This leads to the convex quadratic program

$$\begin{array}{lrcll} \min_{w,b} & \frac{1}{2} w^T w & & & \\ \text{s.t.} & w^T p_i - b & \geq & 1, & p_i \in P, \\ & w^T p_i - b & \leq & -1, & p_i \in Q. \end{array} \tag{8.1}$$

Defining class labels y_i with $y_i = 1$ if $p_i \in P$ and $y_i = -1$ if $p_i \in Q$, the constraints of (8.1) become

$$y_i(w^T p_i - b) - 1 \geq 0, \quad i = 1, \ldots, n.$$

We could solve the optimization problem directly but for reasons that become apparent later we want to move to a dual formulation.

8.2 Lagrangian and dual problem

Lagrangian. Consider a (primal) optimization problem of the form

$$\begin{array}{ll} \min_x & f(x) \\ \text{s.t.} & c_i(x) \leq 0, \qquad i = 1, \ldots, n, \end{array} \tag{8.2}$$

where $f, c_i : \mathbb{R}^d \to \mathbb{R}$. The *Lagrangian* of (8.2) is the function $L : \mathbb{R}^d \times \mathbb{R}^n_+ \to \mathbb{R}$, defined by

$$L(x, \alpha_i) = f(x) + \sum_{i=1}^{n} \alpha_i c_i(x), \quad x \in \mathbb{R}^d, \alpha \in \mathbb{R}^n_+.$$

The Dual problem. Let us assume that there exist $\hat{x} \in \mathbb{R}^d, \hat{\alpha} \geq 0$ such that

$$L(\hat{x}, \alpha) \leq L(\hat{x}, \hat{\alpha}) \leq L(x, \hat{\alpha}), \quad \forall x \in \mathbb{R}^d, \forall \alpha \geq 0, \tag{8.3}$$

meaning that $(\hat{x}, \hat{\alpha})$ is a *saddle point* of the Lagrangian. We get that in this situation, $\hat{x}$ is an optimal solution to (8.2), with

$$f(\hat{x}) = L(\hat{x}, \hat{\alpha}).$$

This is seen as follows. The first saddle point inequality implies $\hat{\alpha}_i c_i(\hat{x}) = 0$ for all i. Indeed, this product is always nonpositive for any feasible solution, and if it were strictly negative, we could increase $L(\hat{x}, \hat{\alpha})$ by setting $\hat{\alpha}_i = 0$.
This means, we have established

$$f(\hat{x}) = L(\hat{x}, \hat{\alpha}) \leq L(x, \hat{\alpha}) \leq f(x), \quad \forall x \in \mathbb{R}^d : c_i(x) \leq 0 \forall i,$$

using the second saddle point inequality as well as $\hat{\alpha}_i c_i(x) \leq 0$ for all i.
We further get

$$\max_{\alpha \geq 0} \min_{x \in \mathbb{R}^d} L(x, \alpha) \leq \max_{\alpha \geq 0} L(\hat{x}, \alpha) \overset{(8.3)}{=} L(\hat{x}, \hat{\alpha}) \overset{(8.3)}{=} \min_{x \in \mathbb{R}^d} L(x, \hat{\alpha}) \leq \max_{\alpha \geq 0} \min_{x \in \mathbb{R}^d} L(x, \alpha),$$

where the first and last inequality have nothing to do with (8.3). This means, the optimum value of the primal problem (8.2) coincides with the optium value of the *dual problem*

$$\begin{array}{ll} \max_\alpha & \min_x L(x, \alpha) \\ \text{s.t.} & \alpha \geq 0 \end{array} \tag{8.4}$$

This dual problem has the nice feature that the constraints $c_i(x) \leq 0$ have 'disappeared'. In return, the objective function looks more complicated now than in the primal, because it has a nested minimum.
If for any fixed α, the Lagrangian is a *convex* differentiable function in x, with continuous partial derivatives. From highsool we know that this implies

$$L(x^*, \alpha) = \min_x L(x, \alpha) \quad \Leftrightarrow \quad \partial_x L(x^*, \alpha) = 0,$$

where ∂_x is the vector of partial derivatives with respect to the variables $x_1, \ldots, x_d$. In this case, we can get rid of the nested minimum in the dual problem, by simply stipulating the additional constraint $\partial_x L(x, \alpha) = 0$. This leads to the following equivalent formulation of (8.4).

$$\begin{array}{ll} \max_\alpha & L(x, \alpha) \\ \text{s.t.} & \alpha \geq 0 \\ & \partial_x L(x, \alpha) = 0. \end{array} \tag{8.5}$$

In the maximum margin hyperplane problem, the Lagrangian is the convex quadratic function (in $x = (w, b)$)

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_i \alpha_i (y_i (w^T p_i - b) - 1),$$

so we can apply the previous machinery. The condition $\partial_x L(x, \alpha) = 0$ reads as

$$\begin{aligned} \frac{\partial L}{\partial w} = 0 \quad &\Leftrightarrow \quad w = \sum_i \alpha_i y_i p_i \\ \frac{\partial L}{\partial b} = 0 \quad &\Leftrightarrow \quad \sum_i \alpha_i y_i = 0. \end{aligned}$$

We can use the first equation to eliminate w and b from the objective function of the dual problem, i.e., from the Lagrangian L, and get the dual problem

$$\begin{array}{ll} \max_\alpha & -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j p_i^T p_j + \sum_i \alpha_i \\ \text{s.t.} & \alpha \geq 0 \\ & \sum_{i=1}^n \alpha_i y_i = 0. \end{array} \tag{8.6}$$

The constraint $w = \sum_i \alpha_i y_i p_i$ can be omitted from the problem; later, we are going to use it, of course, to derive w from an optimal solution $\hat{\alpha}$ to (8.6).
The important fact is that in our case, there is a saddle point of the Lagrangian according to (8.3), so that the maximum value of the dual problem (8.6) indeed coincides with the minimum value of the primal problem (8.1): there is no *duality gap*. This is implied by the following very general result.

Theorem 8.2.1 (Karush-Kuhn-Tucker) *Given a (primal) optimization problem (8.2) with convex objective function f and convex constraints c_i. Under some mild additional conditions (no conditions are needed if the c_i are linear), $\hat{x}$ is an optimal solution to (8.2) if and only if there exists $\hat{\alpha} \geq 0$ such that $(\hat{x}, \hat{\alpha})$ is a saddle point of the Lagrangian.*

If P and Q can be linearly separated, there is a feasible and therefore also an optimal solution $\hat{w}, \hat{b}$ to the primal problem (8.1).[1]

[1]This follows from a standard compactness argument: if there is a feasible solution w_0, then we only need to look for an optimal solution within the ball $\{w \mid w^T w \leq w_0^T w_0\}$. This ball is compact, and so is its intersection with the closed halfspaces induced by the constraints. Within this compact intersection, the continuos objective function $w^T w/2$ assumes a minimum.

By the Karush-Kuhn-Tucker conditions, there is a saddle point of the Lagrangian which in turn implies that the dual problem (8.6) has an optimal solution whose value coincides with the optimum value of the primal. We can use the latter solution to compute the normal $\hat{w}$ of a maximum margin hyperplane as $\sum_i \hat{\alpha}_i y_i p_i$, where $\hat{\alpha}_i$ is an optimal solution of the dual problem. To compute the offset of a maximum margin hyperplane we can use the *complementarity conditions* implied by the saddle point, see above: given that $\alpha_i > 0$ for some index i, we find that $y_i(\hat{w}^T p_i - \hat{b}) - 1 = 0$. From this we get

$$\hat{b} = \hat{w}^T p_i - y_i = \sum_j \hat{\alpha}_j y_j p_j^T p_i - y_i.$$

The data points p_i for which $\alpha_i > 0$ are called *support vectors*.
Let us now have a closer look at the solution $\hat{w}$ of the dual problem.

Lemma 8.2.2 *Let $\hat{w}$ be a normal of a maximum margin hyperplane that separates point sets P and Q and let $\|p - q\|$ with $p \in \mathrm{conv}(P)$ and $q \in \mathrm{conv}(Q)$ the minimal distance between $\mathrm{conv}(P)$ and $\mathrm{conv}(Q)$. Then $p - q = \hat{w}/\sum_{\{i \,|\, p_i \in P\}} \hat{\alpha}_i$, where the $\hat{\alpha}_i$ are optimal for (8.6).*

Proof. Consider the dual of the maximum margin hyperplane problem. The second constraint can also be written as

$$\sum_{\{i \,|\, p_i \in P\}} \alpha_i = \sum_{\{i \,|\, p_i \in Q\}} \alpha_i.$$

Let $c := \sum_{\{i \,|\, p_i \in P\}} \hat{\alpha}_i$. We must have $c \neq 0$ (why?), so if we re-scale the vector $\hat{w} = \sum_i \hat{\alpha}_i y_i p_i$ by the factor $1/c$ we get $\hat{w}/c = p - q$, where

$$p = \sum_{\{i \,|\, p_i \in P\}} \bar{\alpha}_i p_i \in \mathrm{conv}(P) \quad \text{and} \quad q = \sum_{\{i \,|\, p_i \in Q\}} \bar{\alpha}_i p_i \in \mathrm{conv}(Q),$$

and $\bar{\alpha}_i := \hat{\alpha}_i / c$. We have

$$\|p - q\|^2 = \sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i y_j p_i^T p_j$$

and claim that this is the squared distance of the polytopes $\mathrm{conv}(P)$ and $\mathrm{conv}(Q)$. To see this assume the contrary, i.e., $\|p - q\|$ is larger than the distance between $\mathrm{conv}(P)$ and $\mathrm{conv}(Q)$. Consider the following quadratic programming formulation of the polytope distance problem:

$$\begin{array}{ll} \min\limits_{\beta} & \sum_{i,j} \beta_i \beta_j y_i y_j p_i^T p_j \\ \text{s.t.} & \beta \geq 0 \\ & \sum_{\{i \,|\, p_i \in P\}} \beta_i = 1 \\ & \sum_{\{i \,|\, p_i \in Q\}} \beta_i = 1. \end{array}$$

Note that a solution $\hat{\beta}$ for this problem is feasible for the dual of the maximum margin hyperplane problem. The same holds for $\tilde{\beta} := c\hat{\beta}$. By our assumption, we have

$$\|p - q\|^2 = \sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i y_j p_i^T p_j > \sum_{i,j} \hat{\beta}_i \hat{\beta}_j y_i y_j p_i^T p_j.$$

This implies

$$\begin{aligned} -\frac{1}{2} \sum_{i,j} \hat{\alpha}_i \hat{\alpha}_j y_i y_j p_i^T p_j + \sum_i \hat{\alpha}_i &= -\frac{c^2}{2} \sum_{i,j} \bar{\alpha}_i \bar{\alpha}_j y_i y_j p_i^T p_j + 2c \\ &< -\frac{c^2}{2} \sum_{i,j} \hat{\beta}_i \hat{\beta}_j y_i y_j p_i^T p_j + 2c \\ &= -\frac{1}{2} \sum_{i,j} \tilde{\beta}_i \tilde{\beta}_j y_i y_j p_i^T p_j + \sum_i \tilde{\beta}_i, \end{aligned}$$

which is not possible since $-\frac{1}{2} \sum_{i,j} \hat{\alpha}_i \hat{\alpha}_j y_i y_j p_i^T p_j + \sum_i \hat{\alpha}_i$ is the optimum value of (8.6). Thus, $\|p - q\|$ is the optimal value for the objective function of the polytope distance problem and the statement of the lemma follows through uniqueness of $p - q$ (an easy exercise). □

That is, the maximum margin problem is essentially the polytope distance problem, see also Figure 8.2.

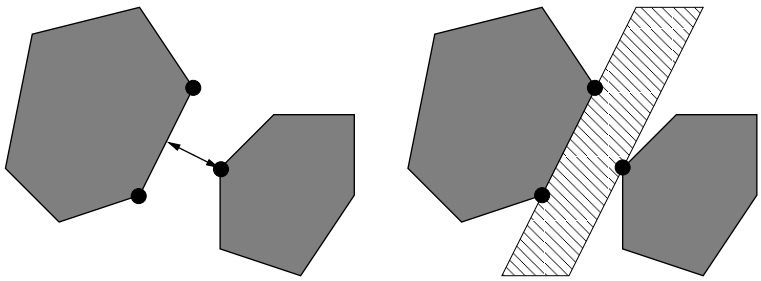

Figure 8.2: The maximum margin problem and the polytope problem are related. The highlighted vertices are the support vectors for both problems.

8.3 Relaxed Maximum Margin Hyperplane

The assumption of linearly separable data sets is not realistic for most applications. Here we want to deal with the case that though the data are not linearly separable a linear separation still makes sense, because it classifies most of the data correctly. Figure 8.3 depicts an example where the data are not linearly separable, but a linear separation still makes sense.

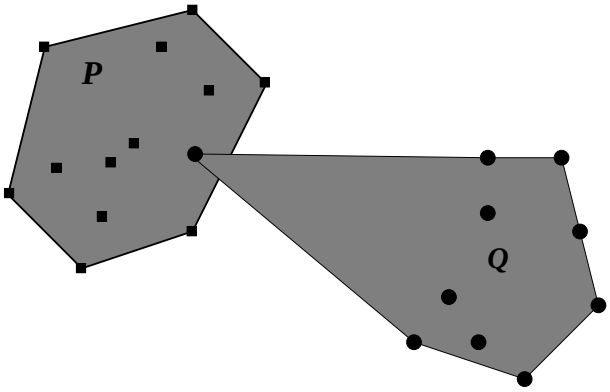

Figure 8.3: Inseparable data set for which a linear separation still is meaningful.

For linearly inseparable data sets (this includes the case where the convex hulls of the two data sets just touch), any pair (w, b) will violate at least one of the constraints

$$y_i(w^T p_i - b) - 1 \geq 0$$

of the primal problem (8.1). The plan is now to relax these constraints by adding positive slack variables z_i. The i-th relaxed constraints now reads as

$$y_i(w^T p_i - b) + z_i - 1 \geq 0, \quad z_i \geq 0.$$

Relaxing the constraints means allowing outliers. We penalize outliers by adding another term to the objective function of the maximum margin hyperplane problem that contains the slack variables. The relaxed maximum margin hyperplane problem becomes

$$\begin{array}{lll} \min_{w,b} & \frac{1}{2}w^T w + C\sum_i z_i & \\ \text{s.t.} & y_i(w^T p_i - b) + z_i - 1 \;\geq\; 0, & i = 1, \ldots, n \\ & z_i \;\geq\; 0, & i = 1, \ldots, n. \end{array} \tag{8.7}$$

Here $C \geq 0$ is a parameter that controls the trade-off between maximizing the margin and penalizing the outliers. The problem still is a convex quadratic optimization problem that we can dualize as we did with the non-relaxed problem. We find that the dual problem to (8.7) is the problem

$$\begin{array}{ll} \max_{\alpha} & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j p_i^T p_j + \sum_i \alpha_i \\ \text{s.t.} & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \\ & \sum_i \alpha_i y_i = 0. \end{array} \tag{8.8}$$

That is, the only difference to the non-relaxed dual problem (8.6) is that the coefficients α_i are also upper-bounded by the trade-off parameter C. The geometric interpretation of this situation is that instead of separating the convex hulls of the data, *reduced convex hulls* get separated, see Figure 8.4 for an example. To see this, we can argue as before that the vector $\hat{w} = \sum_i \hat{\alpha}_i y_i p_i$ resulting from an optimal solution to (8.8) satisfies

$$\hat{w}/c = p - q,$$

where $c = \sum_{\{i|x_i \in P\}} \hat{\alpha}_i$ and $p - q$ is the shortest vector with $p \in \text{conv}_{C/c}(P), q \in \text{conv}_{C/c}(Q)$,

$$\text{conv}_t(X) := \{\sum_{x \in X} \lambda_x x \mid \sum_{x \in X} \lambda_x = 1, 0 \leq \lambda_x \leq t \ \ \forall x \in X\}.$$

Note that $\text{conv}_t(X)$ becomes empty for $t < 1/|X|$.

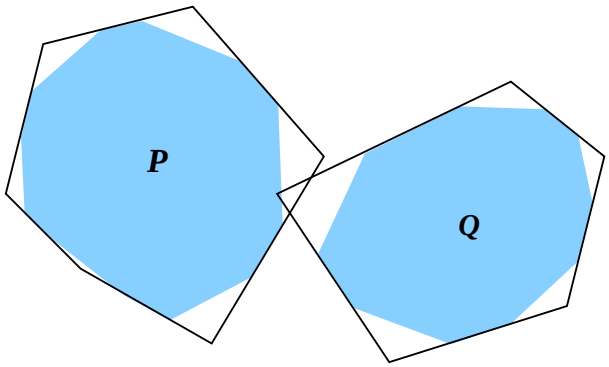

Figure 8.4: Reduced convex hulls.

8.4 Kernel trick

In many cases a linear classifier simply does not do the job even if we allow outliers. For example the data in Figure 8.5 can be separated meaningfully only with a non-linear discriminant function.

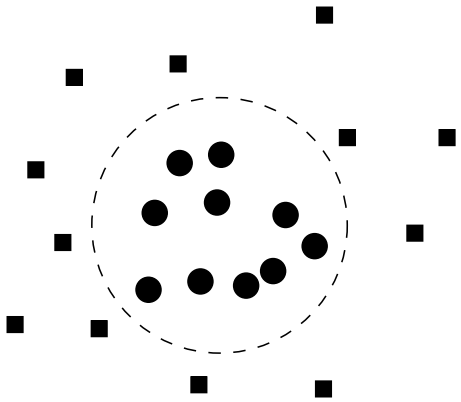

Figure 8.5: Data set is "best" separated by a non-linear classifier.

The key idea behind support vector machines is to map the data points non-linearly into some higher dimensional space, where they (hopefully) can be separated linearly.

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^{d'}$ be such a mapping. The dual of the relaxed maximum margin problem in $\mathbb{R}^{d'}$ looks as follows

$$\begin{array}{ll} \max\limits_{\alpha} & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \varphi(p_i)^T\varphi(p_j) + \sum_i \alpha_i \\ \text{s.t.} & 0 \le \alpha_i \le C, \quad i = 1,\dots,n \\ & \sum_i \alpha_i y_i = 0. \end{array}$$

That is, the constraints remain exactly the same, only the objective function changes. The classifier that we get from a solution $\hat{\alpha}_i$ of this problem is

$$f(x) := \operatorname{sign}\left(\underbrace{\sum_i \hat{\alpha}_i y_i \varphi(p_i)^T}_{\hat{w}}\varphi(x) - \hat{b}\right).$$

Note that (depending on φ) this is now a non-linear classifier on $\mathbb{E}^d$ though it is linear on $\mathbb{E}^{d'}$. In Figure 8.6 it is schematically shown how this non-linear classification works.

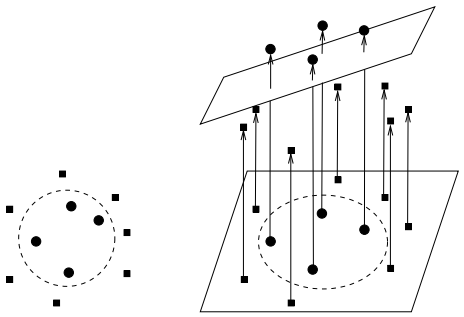

Figure 8.6: Linear separation of lifted data and the resulting non-linear classifier in input space.

In practice it is often infeasible to map the data explicitly to some higher dimensional space. Instead it is done implicitly using a *kernel* function. A kernel function is a positive semi-definite function $k : \mathbb{E}^d \times \mathbb{E}^d \to \mathbb{R}$, and we will use $k(x, y)$ to replace the scalar product $\varphi(x)^T\varphi(y)$ in d'-dimensional space.
It can be shown that for certain kernel functions, this actually works, meaning that there is a mapping φ from $\mathbb{R}^d$ to some higher (possibly infinite) dimensional Hilbert space such that

$$k(x, y) = \varphi(x)^T\varphi(y), \quad x, y \in \mathbb{R}^d.$$

This is Mercer's theorem, also known as the *kernel trick*, Popular kernel function are

(1) Polynomials of degree d': $k(x, y) = (x^T y + c)^{d'}$,

(2) Gaussian functions: $k(x, y) = e^{-c\|x-y\|^2}$,

(3) Sigmoid functions: $k(x, y) := \tanh(x^T y + c)$.

The resulting classifier when using the kernel trick is

$$\operatorname{sign}\left(\sum_i \hat{\alpha}_i y_i k(p_i, x) - b\right).$$

Note that the kernels in the sum of the classifier only have to be evaluated for support vectors, i.e., for data points p_i with $\hat{\alpha}_i > 0$.

Bibliography