8.1 Binomial Coefficients

$$C_{(p,N)} + C_{(p,N-1)} = C_{(p+1,N)} \tag{1}$$

$$2\sum_{k=0}^{N-1}\binom{p-1}{k} + 2\sum_{k=1}^{N-1}\binom{p-1}{k-1} = 2\sum_{k=0}^{N-1}\binom{p}{k} \quad \left|\text{extend k range in } 2^{nd} \text{ sum:} \binom{n}{-1} = 0 \;\forall\, n \in \mathbb{N}\right. \tag{2}$$

$$2\sum_{k=0}^{N-1}\binom{p-1}{k} + 2\sum_{k=0}^{N-1}\binom{p-1}{k-1} = 2\sum_{k=0}^{N-1}\binom{p}{k} \tag{3}$$

$$2\sum_{k=0}^{N-1}\left[\binom{p-1}{k} + \binom{p-1}{k-1}\right] = 2\sum_{k=0}^{N-1}\binom{p}{k} \quad \left|\text{ apply (2) from task sheet}\right. \tag{4}$$

$$2\sum_{k=0}^{N-1}\binom{p}{k} = 2\sum_{k=0}^{N-1}\binom{p}{k} \quad \square \tag{5}$$

8.2 Geometry of Linear Classification

a) The decision boundary is a hyperplane defined by $(w^T x - b = 0)$. w decides the orientation, b the offset from the origin. w points in the direction of positive classification values $y(x)$ (1).

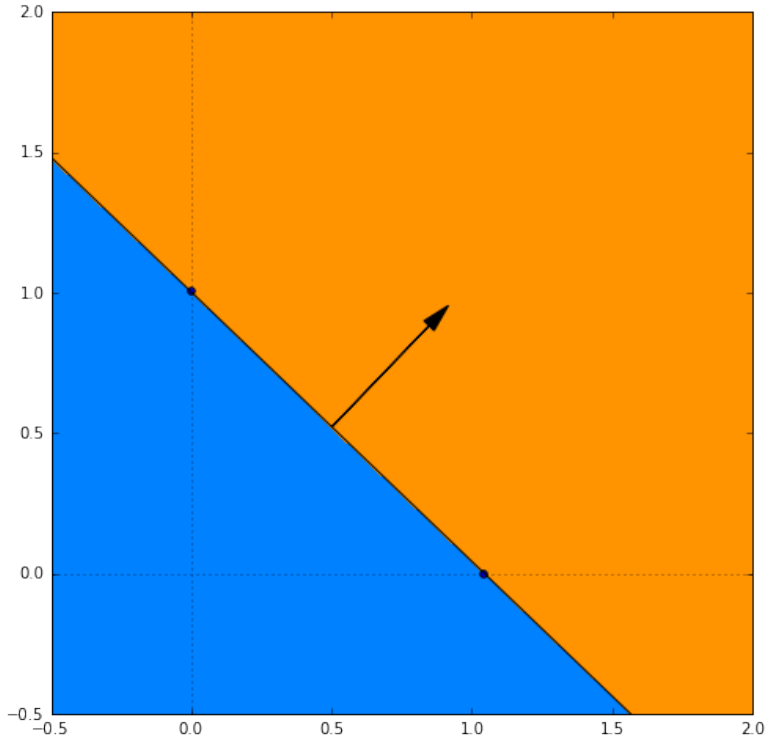

Figure 1:

- the decision boundary is given by $\{x \mid y = 0\}$
- this can be expressed as a linear function: $x_1 = -\frac{w_0}{w_1}x_0 - \frac{b}{w_1}$
- as one can see, the slope of the boundary is $-\frac{w_0}{w_1}$ and the offset is $-\frac{b}{w_1}$
- figure 2: w_0 only affects the slope
- figure 3: varying w_1 also results in different offsets
- figure 4: increasing b leads to a negative offset, scaled by w_1

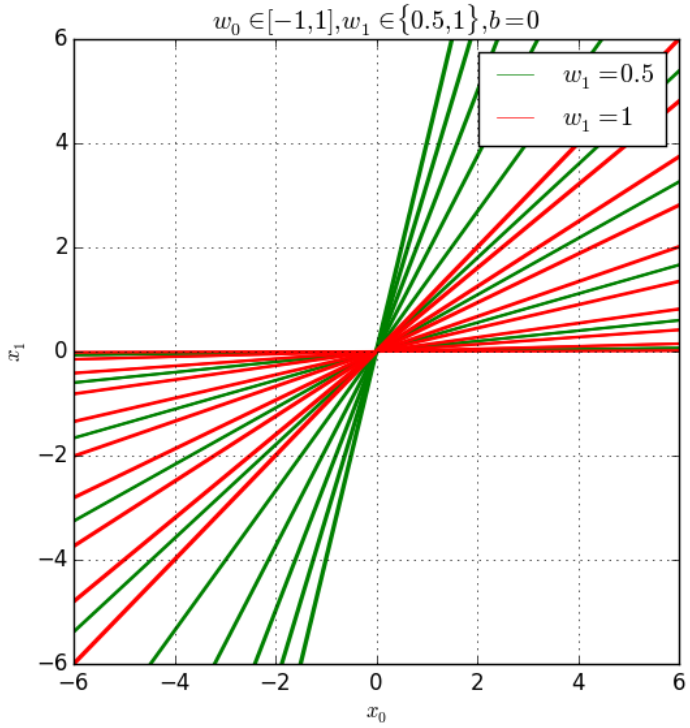

Figure 2:

b) A linear classifier divides the input space into two half-spaces. The *shattering coefficient* $s_{halfspaces}(n)$ then is the largest number of subsets that can be formed by intersecting *any* one input set X of n points using half-spaces. The highest possible shattering coefficient is 2^n.

$s_H(n) = \max_{X \subseteq \Omega | X|=n} |\{X \cap H_i \mid H_i \in H\}|$

The *VC-dimension* d_{VC} provides an upper bound to the shattering coefficient: $ln(s(n)) \leq d_{VC}(1 + ln\frac{p}{d_{VC}})$. It provides a measure of the capacity of the separation capabilities of the separator function class H. If we can find n points, that can be shattered by H (that is, H can separate all possible binary labelings of the points), then VC(H) = n. It is sufficient, that some such example of n points exists (fixed points, but H must shatter them for all possible labelings).

Small VC-dimensions are often applicable for real life data, since close points will have same labels, so separator classes with low VC-dimensions (low complexity models) might actually be preferrable. For a linear classifier in 2 input dimensions, the VC-dimension is 2. Correct linear classification is possible, if all classes can be bounded by non-overlapping convex domains.

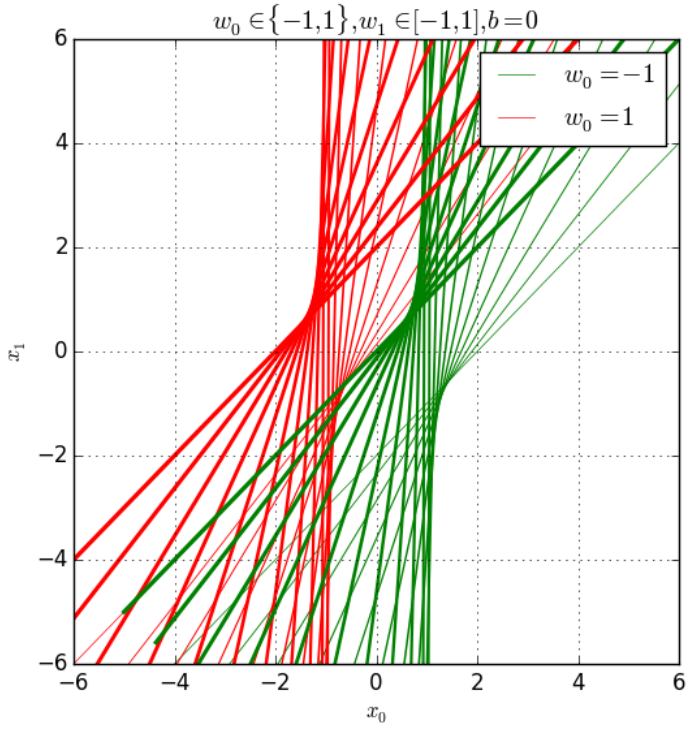

Figure 3:

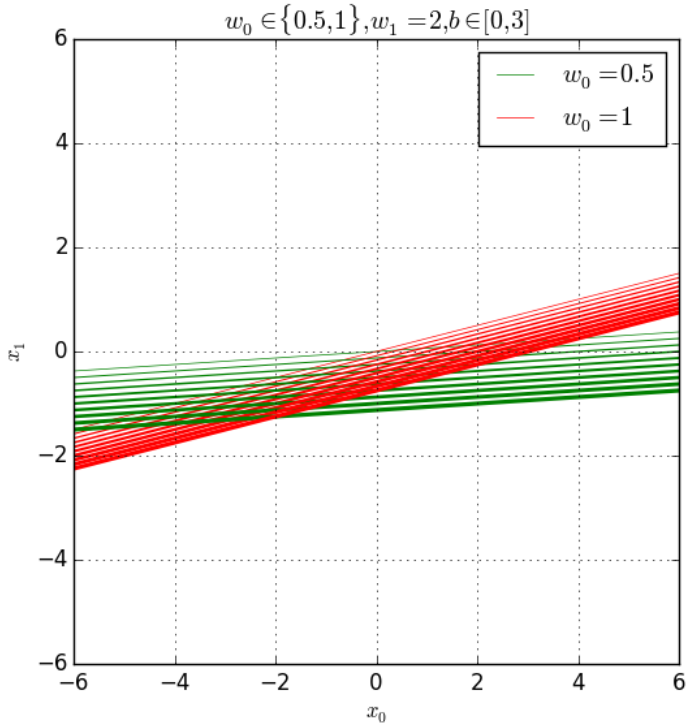

Figure 4:

8.3 The primal SRM problem

a) The margin is the lowest distance between the decision boundary and the closest datapoints. The margin of a canonical hyperplane is set to be $d = \frac{1}{|w|}$ by choosing $|w|$ appropriately. A low margin means higher model complexity C, which means we might get a worse upper bound for the generalization error $E^G < E^T + C$, even if the separate the training set perfectly (overfitting). A high margin reduces model complexity, but increases the training error (underfitting).

b) Idea: Project x on w, which gets us the large vector from the origin to the tip of the dotted line in figure 5. From this we have to subtract the length of the part from the origin to the decision boundary, which will be α.

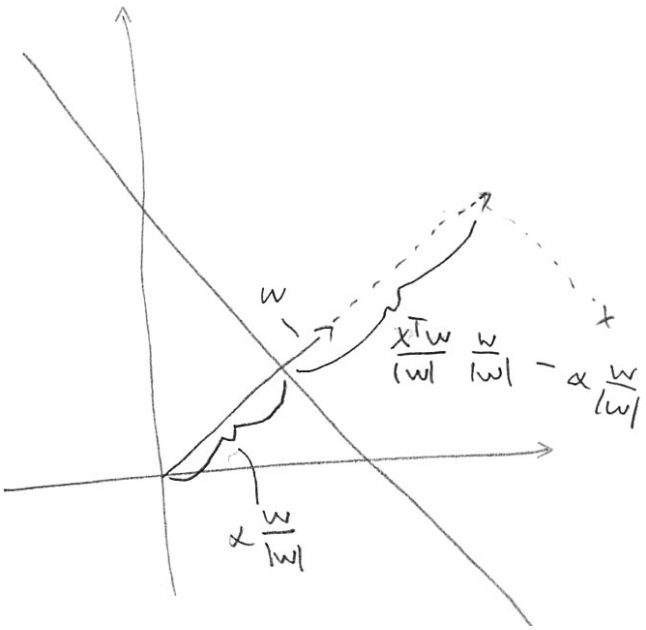

Figure 5: sketch for 8.3b

Let x be any closest point. The projection of x onto w is is $\frac{x^T w}{|w|}\frac{w}{|w|}$, where $\frac{w}{|w|}$ is a unit vector.

Calculate α (the scalar to stretch $\frac{w}{|w|}$ by to reach the boundary):

$$
\begin{aligned}
& w^T(\alpha\frac{w}{|w|}) + b = 0 \\
\Leftrightarrow\ & \frac{\alpha}{|w|} w^T(w) + b = 0 \\
\Leftrightarrow\ & \frac{\alpha}{|w|} \underbrace{w^T w}_{|w|^2} + b = 0 \\
\Leftrightarrow\ & \frac{\alpha}{|w|} |w|^2 + b = 0 \\
\Leftrightarrow\ & \alpha|w| + b = 0 \\
\Leftrightarrow\ & \alpha = \frac{-b}{|w|}
\end{aligned}
$$

Then the distance of x to the decision boundary is:

$$\begin{aligned} d &= \left\| \frac{x^T w}{|w|} \frac{w}{|w|} \right\| - \left\| \alpha \frac{w}{|w|} \right\| \\ &= \frac{x^T w}{|w|} - \frac{-b}{|w|} \\ &= \frac{x^T w + b}{|w|} \qquad \left| \text{for the closest point: } w^T x + b = 1 \right. \\ &= \frac{1}{|w|} \end{aligned}$$

For points farther away than x the term $\frac{x^T w}{|w|} \frac{w}{|w|}$ increases, while $\alpha \frac{w}{|w|}$ stays the same. Hence the distance gets larger, so overall: $d \geq \frac{1}{|w|}$

c) Primal Problem:

minimize $f_0(x)$ subject to $f_k(x) \leq 0$, $k = 1, ..., m$, which in our case is:

minimize $\frac{1}{2}|w|^2$ subject to $-y_T^{(\alpha)}((w^T x^{(\alpha)} + b) - 1) \leq 0$

This means we wish to find the largest margin $\frac{1}{|w|}$, that seperates all training datapoints:

$y_T^{(\alpha)}(w^T x^{(\alpha)} + b) \geq 1 \quad \forall \alpha$

holds for $y_T^{(\alpha)} \in \{-1, 1\}, \quad \forall \alpha$ and

$w^T x^{(\alpha)} + b \geq y_T^{(\alpha)} \quad \forall y_T^{(\alpha)} = 1,$

$w^T x^{(\alpha)} + b \leq y_T^{(\alpha)} \quad \forall y_T^{(\alpha)} = -1$

The constraint term at the top is just a concise rewrite of these 2 inequalities.

Since the objective function $\frac{1}{2}|w|^2$ is convex, it only has one optimal solution. And since this is a standard formulation, we can use solvers to solve it. The solution should classify all training points perfectly ($E^T = 0$), while minimizing the complexity C.

8.5 Kernel Construction

a) kernel matrix is symmetric:

$$K(x, x') \;=\; \phi(x)^T \, \phi(x') \;=\; \phi(x')^T \, \phi(x) \;=\; K(x', x) \tag{6}$$

b) kernel matrix is positive semidefinite:

$$a^T K a \;=\; \sum_{i,j=0}^{P} a_i a_j K_{ij} \;=\; \sum_{i,j=0}^{P} a_i a_j \langle \phi(x_i), \phi(x_j) \rangle \tag{7}$$

$$= \langle \sum_{i=0}^{P} a_i \phi(x_i) \, , \, \sum_{j=0}^{P} a_j \phi(x_j) \, \rangle \;=\; || \sum_{i=0}^{P} a_i \phi(x_i) ||^2 \;\geq\; 0 \tag{8}$$

c) sum of two kernels is also a valid kernel:

- the sum of two symmetric matrices is (obviously) also symmetric and the sum of two positive semidefinite matrices is positive semidefinte as well:

$$a^T (K_1 \;+\; K_2) a \;=\; a^T K_1 \, a \;+\; a^T K_2 \, a \;=\; c_1 \;+\; c_2 \;\geq\; 0 \tag{9}$$

- this makes the sum of those matrices also a valid kernel, as Mercer's theorem states (Haykin - *"Neural Networks - A Comprehensive Foundation"*, 2nd Ed., p. 331)