ASSIGNMENT 3: CONSTRAINED OPTIMIZATION, SVMS

Johannes Kofler

JOHANNES KEPLER UNIVERSITY LINZ

Institute for Machine Learning

Copyright statement:

SVMs for linear separability: The formal setup: Part 5: Abstract problem formulation

- Original Problem: For given linearly separable data set $\mathbf{Z}$, $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

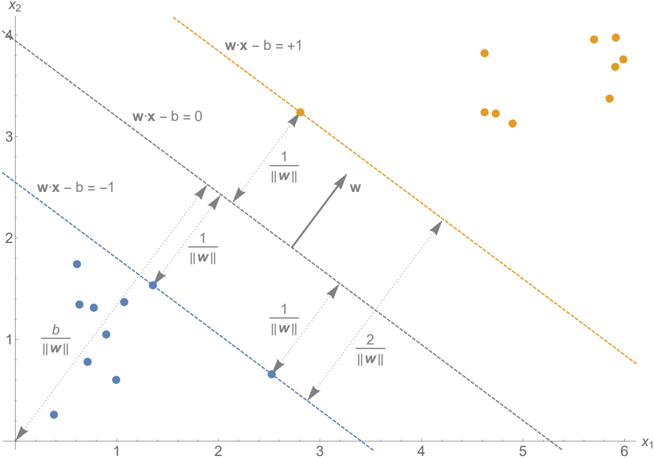
$$\begin{array}{lll} \text{Maximize} & \dfrac{2}{\|\mathbf{w}\|} & \\ \text{subject to} & y_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1 \geq 0 & \text{for } i = 1, \ldots, l. \end{array}$$

- Equivalent: Primal Problem:

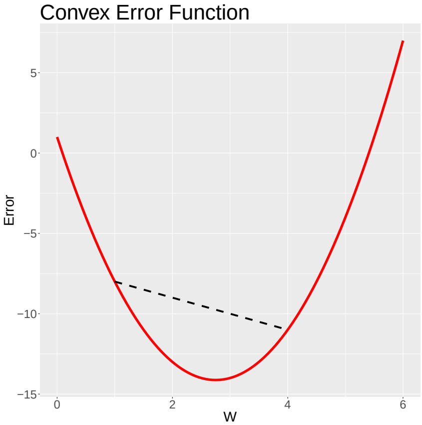
$$\begin{array}{lll} \text{Minimize} & \dfrac{1}{2}\|\mathbf{w}\|^2 = \dfrac{1}{2}\sum_{i=1}^{d} w_i^2 & \\ \text{subject to} & -(y_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1) \leq 0 & \text{for } i = 1, \ldots, l. \end{array}$$

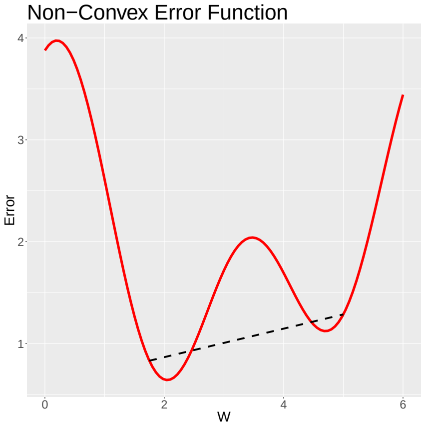
- $\rightarrow$ Convex quadratic optimization problem with linear constraints. Next: recall basic mathematical tools.

SVMs for linear separability: The formal setup: Part 3

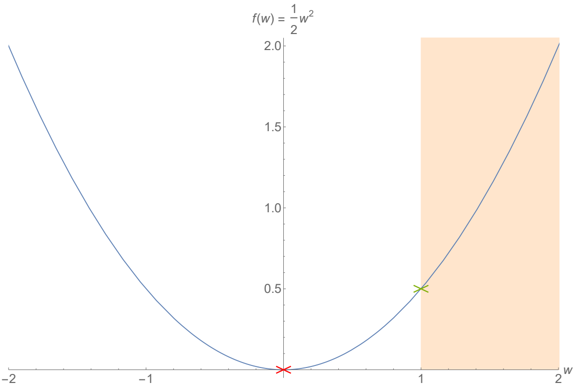


Interlude: constrained convex optimization: Part 1





Interlude: constrained convex optimization: Part 2



- Red: Global minimum
- Green: Constrained minimum under constraint $w > 1$

Interlude: constrained convex optimization: Part 3

- Assumptions: we have functions f and h_i ($i = 1, \ldots, l$) from $\mathbb{R}^d$ to $\mathbb{R}$ with following requirements:
  1. Convex
  2. Continuously Differentiable
  3. Slater condition: there exists $\mathbf{w}'$ with $h_i(\mathbf{w}') < 0$ for all $i = 1, \ldots, l$.
- Primal Problem:

$$\begin{array}{rll} \text{Minimize} & f(\mathbf{w}) & \\ \text{subject to} & h_i(\mathbf{w}) \leq 0 & \text{for } i = 1, \ldots, l. \end{array}$$

- For simplicity: we don't deal with equality constraints here
- Next slide: strategy how to solve this problem

Interlude: constrained convex optimization: Part 4

- Lagrange function:

$$L(\mathbf{w}; \alpha_1, \ldots, \alpha_l) = f(\mathbf{w}) + \sum_{i=1}^{l} \alpha_i h_i(\mathbf{w})$$

$\alpha_1, \ldots, \alpha_l$: Lagrange multipliers.

- Dual Problem:

Maximize $\quad \mathcal{L}(\alpha_1, \ldots, \alpha_l) = \inf_{\mathbf{w}} L(\mathbf{w}; \alpha_1, \ldots, \alpha_l)$ wrt. $\alpha_1, \cdots, \alpha_l$

subject to $\quad \alpha_i \geq 0 \quad$ for $i = 1, \ldots, l$.

- Karush-Kuhn-Tucker (KKT): $\mathbf{w}^*$ solves primal problem $\Leftrightarrow$ there exist non-negative Lagrange multipliers with:
  1. $\mathcal{L}(\alpha_1, \ldots, \alpha_l) = L(\mathbf{w}^*; \alpha_1, \ldots, \alpha_l)$
  2. $\alpha_1, ..., \alpha_l$ solve dual problem, i.e. they maximize $\mathcal{L}$.
  3. $\alpha_i h_i(\mathbf{w}^*) = 0$ for all $i = 1, \ldots, l$.

Interlude: constrained convex optimization: Part 5

- Illustrative example (toy problem):

$$\begin{aligned} \text{Minimize} \quad & f(w_1, w_2) = (w_1 - 2)^2 + w_2^2 \\ \text{subject to} \quad & h_1(w_1, w_2) = -w_1 \le 0 \\ \text{and} \quad & h_2(w_1, w_2) = w_1 - w_2 \le 0. \end{aligned}$$

- Lagrange function:

$$L(w_1, w_2, \alpha_1, \alpha_2) = (w_1 - 2)^2 + w_2^2 - \alpha_1 w_1 + \alpha_2 (w_1 - w_2)$$

- To find $\inf_{(w_1,w_2)} L(w_1, w_2, \alpha_1, \alpha_2)$ for given (α_1, α_2), consider:

$$\frac{\partial L}{\partial w_1}(w_1, w_2, \alpha_1, \alpha_2) = 2\,w_1 - 4 - \alpha_1 + \alpha_2$$

$$\frac{\partial L}{\partial w_2}(w_1, w_2, \alpha_1, \alpha_2) = 2\,w_2 - \alpha_2$$

Interlude: constrained convex optimization: Part 6

- Setting the two derivatives to 0, we obtain $w_1^* = 2 + \frac{\alpha_1 - \alpha_2}{2}$ and $w_2^* = \frac{\alpha_2}{2}$. (This must be a minimum as L is convex and no maximum exists.)
- Furthermore:

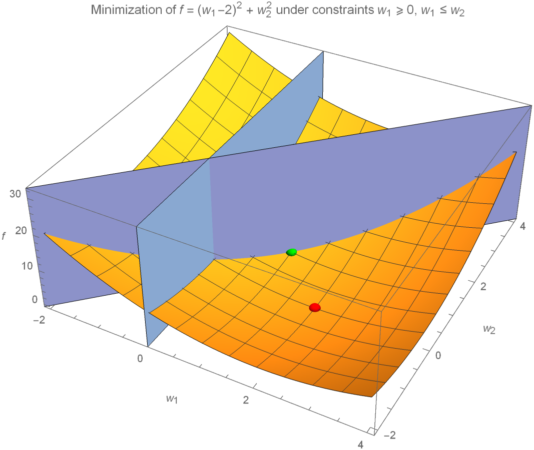
$$\begin{aligned}\mathcal{L}(\alpha_1, \alpha_2) &= L\Big(w_1^*, w_2^*, \alpha_1, \alpha_2\Big) \\ &= -\frac{1}{4}\left(\alpha_1^2 + 8\alpha_1 - 2\alpha_1\alpha_2 - 8\alpha_2 + 2\alpha_2^2\right)\end{aligned}$$

- Dual problem: maximize $\mathcal{L}(\alpha_1, \alpha_2)$ subject to $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$. Yields $\alpha_1 = 0$ and $\alpha_2 = 2$ and thus $w_1^* = 1$ and $w_2^* = 1$. Can be deduced by KKT-conditions.
- In many practical situations: algorithms for solving dual problem exist, especially for SVMs (as we will see later). Here: argument for toy example using KKT conditions.

Interlude: constrained convex optimization: Part 7

- By the third KKT condition: $\alpha_1 h_1(w_1^*, w_2^*) = 0$.
- Assume $h_1(w_1^*, w_2^*) < 0$. Then $\alpha_1 = 0$. Now look at the second KKT-condition $\alpha_2 h_2(w_1^*, w_2^*) = 0$.
  - If $\alpha_2 = 0$: $h_2(w_1^*, w_2^*) = 2 > 0$ which contradicts constraint on h_2 $\rightarrow$ can be ruled out.
  - If $h_2(w_1^*, w_2^*) = 0$ we can deduce $w_1^* = w_2^*$ and thus $\alpha_2 = 2$. (recall: $\alpha_1 = 0$). All the constraints are satisfied, thus we are done, since we have unique solution because of convexity. For sake of completeness, we will also provide arguments ruling out other possibilities:
- Assume $h_1(w_1^*, w_2^*) = 0$. Then $\alpha_2 = 4 + \alpha_1$.
  - If $\alpha_2 = 0$ then $\alpha_1 = -4 \rightarrow$ can be ruled out as $\alpha_i \geq 0$.
  - If $\alpha_2 \neq 0$ then $h_2(w_1^*, w_2^*) = 0$, which implies $\alpha_2 = \frac{\alpha_1}{2} + 2$, i.e. $4 + \alpha_1 = \frac{\alpha_1}{2} + 2$, i.e. $\alpha_1 = -4 \rightarrow$ can be ruled out again.
- Solution: $\alpha_1 = 0, \alpha_2 = 2$, yielding: $w_1^* = 1, w_2^* = 1$ and $f(w_1^*, w_2^*) = 2$.

Interlude: constrained convex optimization: Part 8



- Red: global minimum (not obeying the constraints)
- Green: solution for minimization under constraints

Back to linear SVMs: Part 1

- Associated Lagrange function is given as

$$L(\mathbf{w}, b; \alpha_1, \ldots, \alpha_l) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i \big(y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1\big)$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}\cdot\sum_{i=1}^{l} \alpha_i\, y_i\, \mathbf{x}_i + b\sum_{i=1}^{l} \alpha_i\, y_i + \sum_{i=1}^{l} \alpha_i$$

- Solving the dual problem enforces the conditions

$$\tfrac{\partial L}{\partial \mathbf{w}}(\mathbf{w}, b; \alpha_1, \ldots, \alpha_l) = 0 \qquad \tfrac{\partial L}{\partial b}(\mathbf{w}, b; \alpha_1, \ldots, \alpha_l) = 0,$$

- This implies:

$$\mathbf{w}^* = \sum_{i=1}^{l} \alpha_i\, y_i\, \mathbf{x}_i \qquad \sum_{i=1}^{l} \alpha_i\, y_i = 0$$

- Furthermore:

$$\mathcal{L}(\alpha_1, \ldots, \alpha_l) = \sum_{i=1}^{l} \alpha_i - \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\, \alpha_j\, y_i\, y_j\, \mathbf{x}_i\cdot\mathbf{x}_j\,.$$

- Final solution: maximize $\mathcal{L}$ with respect to α_i subject to $\alpha_i \geq 0$ (for all $i = 1, \ldots, l$) and $\sum_{i=1}^{l} \alpha_i\, y_i = 0$.

Back to linear SVMs: Part 2

- Introduce

$$\mathbf{0} = (\overbrace{0, \ldots, 0}^{l \text{ times}})^T, \qquad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^T,$$

$$\mathbf{1} = (\underbrace{1, \ldots, 1}_{l \text{ times}})^T, \qquad \mathbf{Q} = \left(y_i\, y_j\, \mathbf{x}_i \cdot \mathbf{x}_j\right)_{i=1,\ldots,l}^{j=1,\ldots,l},$$

- The dual problem can be written as follows:

$$\text{Minimize} \qquad -\mathcal{L} = \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \ \text{ wrt. } \boldsymbol{\alpha}$$

$$\text{subject to} \qquad \boldsymbol{\alpha} \geq \mathbf{0} \text{ and } \boldsymbol{\alpha}^T \mathbf{y} = 0.$$

- Easy observation: $\mathbf{Q}$ is positive semi-definite.
- $\rightarrow$ convex quadratic optimization problem with linear constraints: only global minima: uniqueness if positive definite.

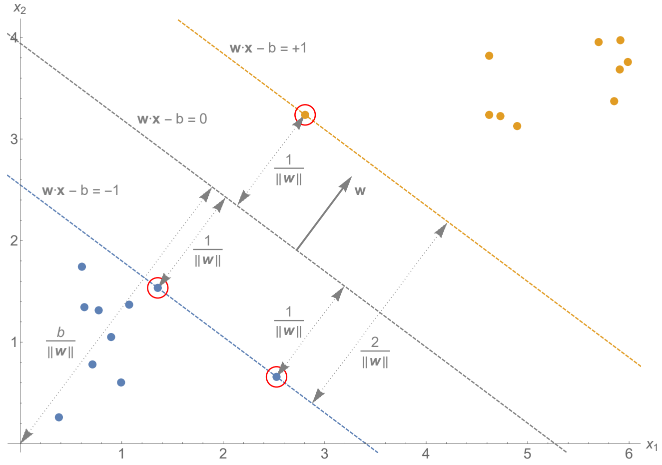
Back to linear SVMs: Part 3

- Once we solved the dual problem $\rightarrow$ obtain $\alpha_1, \ldots, \alpha_l$ which also solve primal problem. By the KKT-conditions:

$$\alpha_i\big(y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1\big) = 0 \quad \text{for} \quad i = 1, \ldots, l.$$

- Thus, for $i = 1, \ldots, l$:
  - either $\alpha_i = 0$
  - or $y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 = 0$
  - or both.
- Samples $\mathbf{x}_i$ for which $\alpha_i > 0$ holds (i.e. $y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 = 0$) lie on the margin border and are called support vectors (encircled in red in next slide).

Back to linear SVMs: Part 4



Back to linear SVMs: Part 5

- Given Lagrange multipliers $\alpha_1, \ldots, \alpha_l$ solving the primal problem, we can construct $\mathbf{w}$ as noted above already:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i \, y_i \, \mathbf{x}_i$$

- $\rightarrow$ final classification function, i.e. the linear Support Vector Machine (SVM), is given as

$$g(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - b) = \mathrm{sign}\Big(\underbrace{\sum_{i=1}^{l} \alpha_i \, y_i \, \mathbf{x}_i \cdot \mathbf{x} - b}_{\text{discriminant function } \bar{g}(\mathbf{x})} \Big).$$

Back to linear SVMs: Part 6

- For arbitrary support vector $\mathbf{x}_j$ (with $\alpha_j > 0$), the KKT condition implies $y_j(\mathbf{w} \cdot \mathbf{x}_j - b) = 1$, and thus

$$b = -y_j + \mathbf{w} \cdot \mathbf{x}_j = -y_j + \sum_{i=1}^{l} \alpha_i \, y_i \, \mathbf{x}_i \cdot \mathbf{x}_j$$

- Recommended: don't base the computation of b on only one support vector (for reasons of numerical precision), but compute a b value for each support vector and use the average finally.
- Under specific conditions: it may be useful to adjust b according to some other quality measure after training.

C-SVMs: Non-linear separability: Part 1

- If positive and negative samples are not linearly separable, the constraints
$$y_i\,(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1 \quad (i = 1, \ldots, l)$$
cannot be all fulfilled simultaneously.
- Introduce non-negative slack variables $\xi_i \geq 0$:
$$y_i\,(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1 - \xi_i \quad (i = 1, \ldots, l)$$
- Require slack variables to be as small as possible. Are scaled by factor $C > 0$.
- The adapted primal problem (called C-SVM) is given as:

$$\begin{aligned}
\text{Minimize} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{l}\xi_i \\
\text{subject to} \quad & -\left(y_i\,(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 + \xi_i\right) \leq 0 \\
\text{and} \quad & -\xi_i \leq 0
\end{aligned}$$

C-SVMs: Non-linear separability: Part 2

- Using the same techniques as for linear case (i.e. formulate dual problem and apply KKT-Theorem), the problem can be cast into the framework of convex quadratic optimization again.
- Classification function g also has similar structure, however, KKT-conditions are a bit more involved
- Overview of calculations: next slides.
- As meaning of C is not very intuitive: different variant called ν-SVM also exists (see later slides). KKT Theorem applies again.
- Connection to hinge loss: $L_h(y_i, \bar{g}(\mathbf{x}_i)) = \max(0, 1 - y_i\bar{g}(\mathbf{x}_i))$. In case of SVMs:
  - L_h is zero $\Leftrightarrow$ data point lies on correct side of margin.
  - If not: loss value is proportional to distance from margin.

C-SVMs: Mathematical details: Part 1

- Again introduce $\alpha_1, \ldots, \alpha_l$ and $\lambda_1, \ldots, \lambda_l$. Then:

$$L(\mathbf{w}, b, \xi_1, \ldots, \xi_l; \alpha_1, \ldots, \alpha_l, \lambda_1, \ldots, \lambda_l)$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\big(y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 + \xi_i\big) - \sum_{i=1}^{l}\lambda_i\xi_i$$

- For dual problem: minimize L for $\mathbf{w}$, b and $\xi_1, \ldots, \xi_l$.
- This enforces:

$$\frac{\partial L}{\partial \mathbf{w}}(\mathbf{w}, b, \xi_1, \ldots, \xi_l; \alpha_1, \ldots, \alpha_l, \lambda_1, \ldots, \lambda_l) = 0,$$
$$\frac{\partial L}{\partial b}(\mathbf{w}, b, \xi_1, \ldots, \xi_l; \alpha_1, \ldots, \alpha_l, \lambda_1, \ldots, \lambda_l) = 0,$$
$$\frac{\partial L}{\partial \xi_j}(\mathbf{w}, b, \xi_1, \ldots, \xi_l; \alpha_1, \ldots, \alpha_l, \lambda_1, \ldots, \lambda_l) = 0, \quad \text{for all } j = 1, \ldots, l$$

- Which implies (the first two conditions are the same as for linear SVMs):

$$\mathbf{w} = \sum_{i=1}^{l}\alpha_i\, y_i\, \mathbf{x}_i,$$
$$\sum_{i=1}^{l}\alpha_i\, y_i = 0,$$
$$C - \alpha_j - \lambda_j = 0 \;\; \text{for all } j = 1, \ldots, l$$

C-SVMs: Mathematical details: Part 2

- Equalities $C - \alpha_j - \lambda_j = 0$ imply $\lambda_j = C - \alpha_j$.
- Constraints $\lambda_j \geq 0$ imply that we must ensure $C - \alpha_j \geq 0$, hence $\alpha_j \leq C$ for all $j = 1, \ldots, l$.
- Finally, we obtain the same objective function

$$\mathcal{L}(\alpha_1, \ldots, \alpha_l) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j \, y_i \, y_j \, \mathbf{x}_i \cdot \mathbf{x}_j .$$

- Solution: maximize $\mathcal{L}$ with respect to α_i subject to $\alpha_i \geq 0$ (for all $i = 1, \ldots, l$), $\sum_{i=1}^{l} \alpha_i y_i = 0$, and additional constraints $\alpha_i \leq C$ (for all $i = 1, \ldots, l$).
- Dual problem:

$$\begin{array}{rl} \text{Minimize} & \dfrac{1}{2} \boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \\ \text{wrt.} & \boldsymbol{\alpha} \\ \text{subject to} & \boldsymbol{\alpha}^T \mathbf{y} = 0 \text{ and } \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}. \end{array}$$

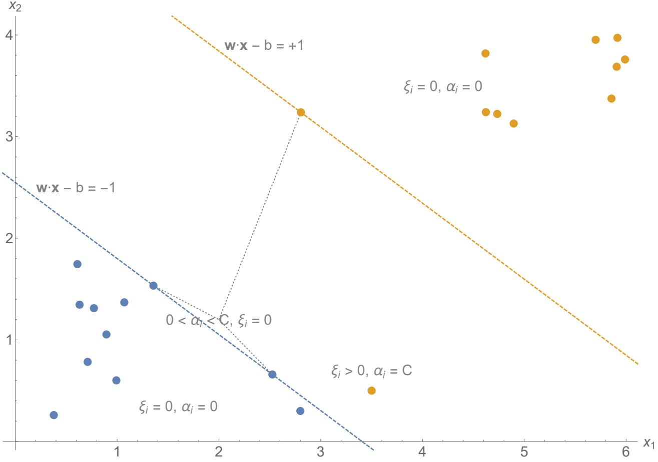
C-SVMs: Mathematical details: Part 3

- Analogously:

$$g(\mathbf{x}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x} - b) = \operatorname{sign}\Big(\sum_{i=1}^{l} \alpha_i \, y_i \, \mathbf{x}_i \cdot \mathbf{x} - b \Big).$$

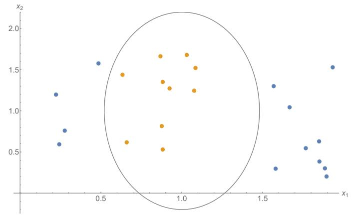
- Computation of b, however, requires a bit more caution.
- In non-separable case, the KKT-conditions tell us that $\alpha_i \big(y_i(\mathbf{w} \cdot \mathbf{x}_i - b) - 1 + \xi_i \big) = 0$ holds for all $i = 1, \ldots, l$. If we choose an i such that $\alpha_i > 0$, we would need ξ_i to determine b.
- However, note that KKT conditions also imply (for the other set of constraints $\xi_i \geq 0$) that $\lambda_i \xi_i = (C - \alpha_i)\xi_i = 0$ holds for all $i = 1, \ldots, l$.
- If we find j with $0 < \alpha_j < C$, we can infer $\xi_j = 0$ and thus $y_j(\mathbf{w} \cdot \mathbf{x}_j - b) - 1 = 0$, i.e. can use same method as before.
- Every $\alpha_j > 0$ corresponds to a support vector $\mathbf{x}_j$.

C-SVMs: Illustration

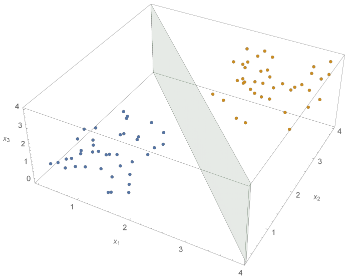


Nonlinear SVM: Part 1

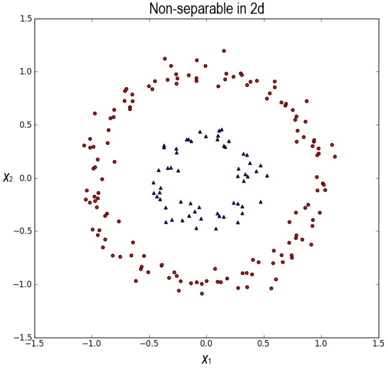
- Linear separability is very restrictive.
- The higher the dimensionality, however, the easier linear separability can be achieved.

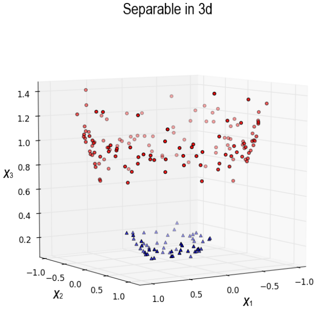


$$\underset{\Phi^{-1}}{\overset{\Phi}{\rightleftharpoons}}$$



Nonlinear SVM: Part 2





- Left: (x_1, x_2), data not linearly separable
- Right: $(x_1, x_2, x_3 = x_1^2 + x_2^2)$, data linearly separable

Nonlinear SVM: Part 3

- Basic idea of nonlinear SVMs: transform data into a higher-dimensional space such that problem hopefully becomes linearly separable there.
- More formal: choose a Hilbert space $\mathcal{H}$ and a (nonlinear) mapping $\Phi : X \to \mathcal{H}$.
- Then try to apply linear method (presented in earlier slides) in the space $\mathcal{H}$.
- Problem: how to specify $\mathcal{H}$ and Φ?

Nonlinear SVM: Part 4

- Recall: In solving the dual problem and computing the final classification function: need only scalar products of pairs of samples. Therefore: not necessary to explicitly know $\mathcal{H}$ and Φ.
- Only need $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ for all $\mathbf{x}_i, \mathbf{x}_j$ ($i, j = 1, \ldots, l$).
- Required for computing the classification of a new sample $\mathbf{x}$: $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}}$ for all $i = 1, \ldots, l$.
- Suppose we are given a mapping $k : X \times X \to \mathbb{R}$ (the kernel) for which we know that there exists Hilbert space $\mathcal{H}$ and mapping $\Phi : X \to \mathcal{H}$ such that $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}}$ for all $\mathbf{x}, \mathbf{y} \in X$.

Nonlinear SVM: Part 5

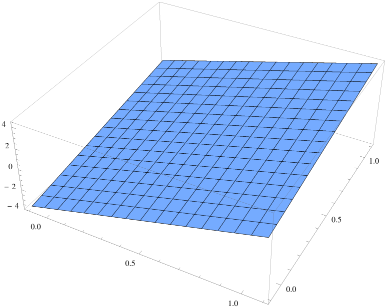
- This is the case $\Leftrightarrow$(Aronszajn) k is positive semi-definite and symmetric, i.e.
  1. $\sum_{i,j} c_i k(\mathbf{x}_i, \mathbf{x}_j) c_j \geq 0$
  2. $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$

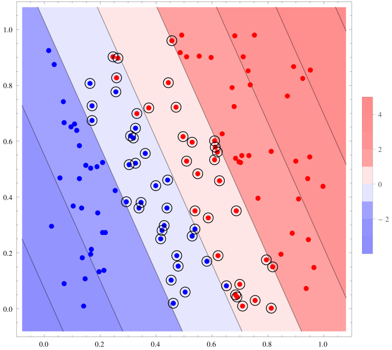
  for $i, j = 1, ..., l$, $c_i, c_j \in \mathbb{R}$, $\mathbf{x}_i, \mathbf{x}_j \in X$.
- Equivalent formulation: Gram matrix $\mathbf{K} = (k_{ij})_{i=1,\dots,l}^{j=1\dots,l} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i=1,\dots,l}^{j=1\dots,l}$ is positive semi-definite and symmetric.
- In practice: make an a priori choice of k using common sense and, if available, prior knowledge about problem: $\rightarrow$ "kernel trick".

Nonlinear SVM: Part 6

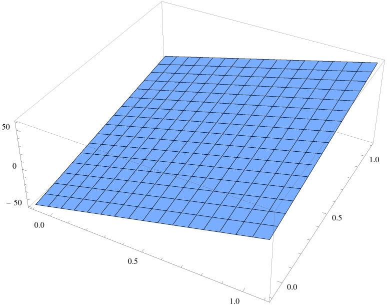
- Which kernels? $\to$ Assume $X = \mathbb{R}^n$ and $k : X^2 \to \mathbb{R}$ continuous. The following statements are equivalent:
  - k is a kernel
  - For $(Af)(\mathbf{x}) = \int_X k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \mathrm{d}\mathbf{y}$ the inequality
  $$\langle Af, f \rangle_{L^2(X)} = \int_{X^2} k(\mathbf{x}, \mathbf{y})\, f(\mathbf{x})\, f(\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \geq 0.$$
  holds for all square-integrable functions $f \in L^2(X)$.
- Mercer's Theorem: If in addition the diagonal $k(x, x)$ is integrable:
  - There are sequences $(\varphi_m)_{m \in \mathbb{N}}$ of continuous eigenfunctions and positive eigenvalues $(\sigma_m)_{m \in \mathbb{N}}$ of A.
  - $k(\mathbf{x}, \mathbf{y}) = \sum_{m \geq 1} \sigma_m \varphi_m(\mathbf{x}) \varphi_m(\mathbf{y})$ and sum converges uniformly on compact sets of X^2.
- More details with proofs: e.g. these notes, chapter 3.5.
- Standard kernels ("$\cdot$" here denotes Euclidean inner product):
  1. Linear: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$
  2. Polynomial: $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + \beta)^\alpha$
  3. Gaussian / RBF: $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{y}\|^2\right)$
  4. Sigmoid: $k(\mathbf{x}, \mathbf{y}) = \tanh(\alpha\, \mathbf{x} \cdot \mathbf{y} + \beta)$

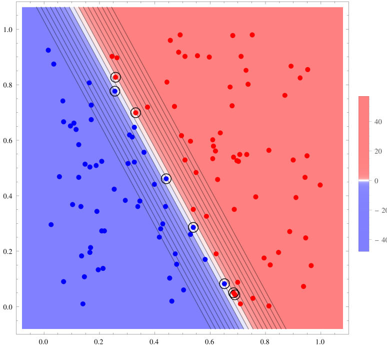
C-SVM Illustration Part 1: $C = 1$, Kernel = linear



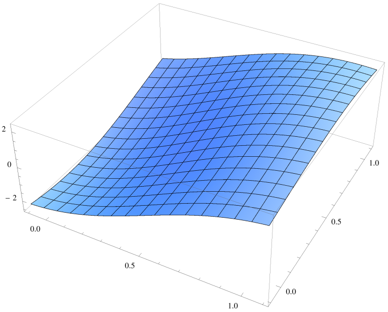


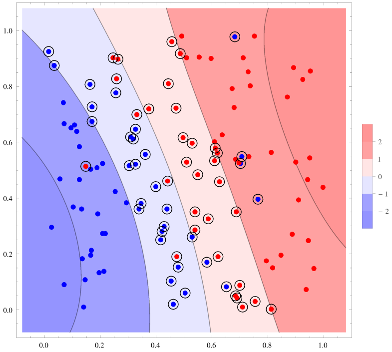
C-SVM Illustration Part 2: $C = 1000$, Kernel = linear





C-SVM Illustration Part 3: $C = 1$, Kernel = RBF, $\frac{1}{2\sigma^2} = 1$





C-SVM Illustration Part 4: $C = 1000$, Kernel = RBF, $\frac{1}{2\sigma^2} = 100$