Statistical Machine Learning

Lecture 8: Support Vector Machines

2022-23

Linear Classification

1. Suppose there exists a function $f : \mathscr{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathscr{X}$ with $f(\boldsymbol{x}) > 0$ is assigned to class 1 and otherwise to class -1. We consider the case where f is a linear function of $\boldsymbol{x} \in \mathscr{X}$, so it can be expressed as:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0,$$

   where $\boldsymbol{W} = (w_0, \boldsymbol{w}) \in \mathbb{R} \times \mathbb{R}^d$ are the parameters that control the function and the decision rule $\operatorname{sgn}(f(\boldsymbol{x}))$.
2. We usually assume that $\|\boldsymbol{w}\|^2 = 1$.
3. The parameters $\boldsymbol{W}$ must be learned from the data.
4. The data set is expected to be $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1}^N$, where $(\boldsymbol{x}_n, t_n) \in \mathscr{X} \times \{-1, +1\}$ for all $n = 1, \ldots, N$.

Perceptron Algorithm and Dual Representation

1. Let $\boldsymbol{y} \in \{1\} \times \mathscr{X}$ such that $\boldsymbol{y}_n = (1, \boldsymbol{x}_n^T)^T$ for all $(\boldsymbol{x}_n, t_n) \in \mathcal{D}$. The perceptron algorithm is:

$$\boldsymbol{W}_{l+1} = \boldsymbol{W}_l + \Theta(-t_{l+1} \boldsymbol{W}_l^T \boldsymbol{y}_{l+1}) t_{l+1} \boldsymbol{y}_{l+1}.$$

2. Assume that $\boldsymbol{W}_0 = \boldsymbol{0}$. Thus

$$\boldsymbol{W}_l = \sum_{j=1}^{l} \alpha_j t_j \boldsymbol{y}_j,$$

where α_j is the number of times $\boldsymbol{y}_j$ has been misclasified.

Perceptron Algorithm and Dual Representation

1. The decision function becomes:

$$\operatorname{sgn}(f(\boldsymbol{x})) = \operatorname{sgn}\left(\sum_{j=1}^{I} \alpha_j t_j \boldsymbol{y}_j^T \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}\right),$$

   where, instead of depending on $\boldsymbol{W}$ it depends on $\{\alpha_j\}$. This is the *dual form* of the function f.

2. Observe that the points close to the boundary, and therefore harder to learn, are expected to have larger α.

Regression Problem

1. Let $\boldsymbol{y} \in \{1\} \times \mathscr{X}$ such that $\boldsymbol{y}_n = (1, \boldsymbol{x}_n^T)^T$ for all $(\boldsymbol{x}_n, t_n) \in \mathcal{D} \subset (\mathscr{X} \times \mathbb{R})^N$. The model and the least-square loss function are:

$$f(\boldsymbol{x}) = \boldsymbol{W}^T \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

$$\mathcal{E}(\boldsymbol{W}) = \sum_{n=1}^{N} \left[t_n - \boldsymbol{W}^T \boldsymbol{y}_n \right]^2,$$

where $\boldsymbol{W} = (w_0, \boldsymbol{w}^T)^T \in \mathbb{R} \times \mathbb{R}^d$.

Regression Problem: Ridge Regression

1. Let $\boldsymbol{y} \in \{1\} \times \mathscr{X}$ such that $\boldsymbol{y}_n = (1, \boldsymbol{x}_n^T)^T$ for all $(\boldsymbol{x}_n, t_n) \in \mathcal{D} \subset (\mathscr{X} \times \mathbb{R})^N$. The model and the least-square loss function are:

$$f(\boldsymbol{x}) = \boldsymbol{W}^T \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

$$\mathcal{E}_{RR}(\boldsymbol{W}, \lambda) = \lambda \boldsymbol{w}^T \boldsymbol{w} + \sum_{n=1}^{N} \left[t_n - \boldsymbol{W}^T \boldsymbol{y}_n \right]^2,$$

   where $\boldsymbol{W} = (w_0, \boldsymbol{w}^T)^T \in \mathbb{R} \times \mathbb{R}^d$.

2. λ is a parameter that controls a trade-off between the low norm of $\boldsymbol{w}$ and the low quadratic error.

Regression Problem: Ridge Regression

1. The equation to be solved is

$$\nabla_{\boldsymbol{w}} \mathcal{E}_{RR} = \begin{pmatrix} \partial_{w_0} \\ \nabla_{\boldsymbol{w}} \end{pmatrix} \mathcal{E}_{RR} = \mathbf{0},$$

2. That admits the solution:

$$w_0^\star = \frac{1}{N} \sum_{n=1}^{N} \left(t_n - \boldsymbol{x}_n^T \boldsymbol{w}^\star \right) = \bar{t} - \bar{\boldsymbol{x}}^T \boldsymbol{w}^\star$$

$$\boldsymbol{w}^\star = \sum_{n=1}^{N} \frac{t_n - \boldsymbol{x}_n^T \boldsymbol{w}^\star - w_0^\star}{\lambda} \boldsymbol{x}_n = \sum_{n=1}^{N} \frac{(t_n - \bar{t}) - (\boldsymbol{x}_n - \bar{\boldsymbol{x}})\boldsymbol{w}^\star}{\lambda} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})$$

$$\alpha_n \equiv \frac{(t_n - \bar{t}) - (\boldsymbol{x}_n - \bar{\boldsymbol{x}})\boldsymbol{w}^\star}{\lambda}$$

$$\boldsymbol{w}^\star = \sum_{n=1}^{N} \alpha_n (\boldsymbol{x}_n - \bar{\boldsymbol{x}}).$$

Ridge Regression: Dual problem

1. Let us define the matrix $\boldsymbol{X} \in \mathscr{X}^N$ such that

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} \qquad \overline{\boldsymbol{X}} = \left.\begin{pmatrix} \overline{\boldsymbol{x}}^T \\ \overline{\boldsymbol{x}}^T \\ \vdots \\ \overline{\boldsymbol{x}}^T \end{pmatrix}\right\} N \text{ times.}$$

2. Observe that $\boldsymbol{w}^\star = \left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T \boldsymbol{\alpha}$.

Ridge Regression: Dual problem

1. Then $\boldsymbol{w}^\star = \left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_N)^T \in \mathbb{R}^N$.
   Let $\boldsymbol{\tau} \in \mathbb{R}^N$ such that $[\boldsymbol{\tau}]_n = t_n - \bar{t}$. Therefore:

$$w_0^\star = \bar{t} - \bar{\boldsymbol{x}}^T \left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T \boldsymbol{\alpha}$$

$$\mathcal{E}_{RR}(\boldsymbol{\alpha}, \lambda) = \lambda \boldsymbol{\alpha}^T \left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T \boldsymbol{\alpha} + \sum_{n=1}^{N} \left[t_n - \bar{t} - \boldsymbol{\alpha}^T \left(\boldsymbol{X} - \right.\right.$$

$$= \lambda \boldsymbol{\alpha}^T \left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T \boldsymbol{\alpha} + \sum_{n=1}^{N} \left[(t_n - \bar{t}) - \boldsymbol{\alpha}^T \left(\boldsymbol{X} - \right.\right.$$

$$= \lambda \boldsymbol{\alpha}^T \boldsymbol{G} \boldsymbol{\alpha} + \boldsymbol{\tau}^T \boldsymbol{\tau} - 2\boldsymbol{\alpha}^T \boldsymbol{G} \boldsymbol{\tau} + \boldsymbol{\alpha}^T \boldsymbol{G} \boldsymbol{G} \boldsymbol{\alpha},$$

   where $\boldsymbol{G}^T = \boldsymbol{G} = \left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T$.

Ridge Regression: Dual problem

1. The derivatives with respect to $\boldsymbol{\alpha}$ are

$$\nabla_{\boldsymbol{\alpha}} \mathcal{E}_{RR} = 2\boldsymbol{G}\left(\lambda\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{\alpha} - \boldsymbol{\tau}\right) = \mathbf{0}.$$

2. The solution is

$$\boldsymbol{\alpha} = \left(\lambda\boldsymbol{I} + \boldsymbol{G}\right)^{-1}\boldsymbol{\tau}.$$

3. The predictive function becomes:

$$\begin{aligned} f(\boldsymbol{x}) &= \left(\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^T \left(\lambda\boldsymbol{I} + \boldsymbol{G}\right)^{-1}\boldsymbol{\tau}\right)^T \left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right) + \overline{t} \\ &= \boldsymbol{\tau}^T \left(\lambda\boldsymbol{I} + \boldsymbol{G}\right)^{-1}\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right) + \overline{t} \end{aligned}$$

Ridge Regression: Dual problem

1. The predictive function becomes:

$$f(\boldsymbol{x}) = \boldsymbol{\tau}^T (\lambda \boldsymbol{I} + \boldsymbol{G})^{-1} \begin{pmatrix} (\boldsymbol{x}_1 - \overline{\boldsymbol{x}})^T (\boldsymbol{x} - \overline{\boldsymbol{x}}) \\ (\boldsymbol{x}_2 - \overline{\boldsymbol{x}})^T (\boldsymbol{x} - \overline{\boldsymbol{x}}) \\ \vdots \\ (\boldsymbol{x}_N - \overline{\boldsymbol{x}})^T (\boldsymbol{x} - \overline{\boldsymbol{x}}) \end{pmatrix} + \overline{t}$$
$$= \boldsymbol{\tau}^T (\lambda \boldsymbol{I} + \boldsymbol{G})^{-1} z + \overline{t}.$$

2. Observe that in the dual representation the data only appears through the Gram matrix $\boldsymbol{G}$, and that in the decision function it is only the inner product of the data points with the test point ($\boldsymbol{x}$) that are needed.

Optimisation Theory

1. Primal Optimisation Problem: Given the functions $f, g_i, h_j : \Omega \to \mathbb{R}$, with $i = 1, \ldots, K$, and $j = 1, \ldots, P$,

$$\begin{array}{lll} \text{minimise} & f(\boldsymbol{w}), & \boldsymbol{w} \in \Omega, \\ \text{subject to} & g_i(\boldsymbol{w}) \leq 0, & i = 1, \ldots, K, \\ & h_j(\boldsymbol{w}) = 0, & j = 1, \ldots, P, \end{array}$$

   where $f(\boldsymbol{w})$ is called the *objective function*, $g_i(\boldsymbol{w})$ are the *inequality constraints* and $h_j(\boldsymbol{w})$ are the *equality constraints*.

2. The region defined as

$$\mathcal{R} \equiv \{\boldsymbol{w} \in \Omega : \boldsymbol{g}(\boldsymbol{w}) \leq \mathbf{0}_K, \, \boldsymbol{h}(\boldsymbol{w}) = \mathbf{0}_P\},$$

   is known as the *feasible region*.

3. The solution $\boldsymbol{w}^\star \in \mathcal{R}$ such that $f(\boldsymbol{w}^\star) \leq f(\boldsymbol{w})$ for all $\boldsymbol{w} \in \Omega$.

Lagrangian Theory

1. Given an optimisation problem with domain $\Omega \subseteq \mathbb{R}^d$,

$$\begin{array}{lll} \text{minimise} & f(\boldsymbol{w}), & \boldsymbol{w} \in \Omega, \\ \text{subject to} & g_i(\boldsymbol{w}) \leq 0, & i = 1, \ldots, K, \\ & h_j(\boldsymbol{w}) = 0, & j = 1, \ldots, P, \end{array}$$

we define the Lagrangian function as

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{w}) + \boldsymbol{\alpha}^T \boldsymbol{g}(\boldsymbol{w}) + \boldsymbol{\beta}^T \boldsymbol{h}(\boldsymbol{w}),$$

with $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{\beta} \in \mathbb{R}^P$ are known as the Lagrange Multipliers.

Lagrangian Theory

1. The dual problem of the primal defined above is

$$\begin{aligned} &\text{maximise} && \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}), \\ &\text{subject to} && \boldsymbol{\alpha} \geq 0, \end{aligned}$$

where

$$\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\boldsymbol{w} \in \Omega} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Weak and Strong Duality

1. WD: Let $\boldsymbol{w} \in \Omega$ be a feasible solution of the primal problem and $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^K \times \mathbb{R}^P$ a feasible solution of the dual problem. Then $f(\boldsymbol{w}) \geq \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The $\Delta(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv f(\boldsymbol{w}) - \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is known as the *Duality Gap.*
2. A set Υ is *convex* if for all $\alpha, \beta \in \Upsilon$ and $t \in [0, 1]$ the convex combination $\gamma = t\alpha + (1 - t)\beta$ is also an element of $\Upsilon \ni \gamma$.
3. SD: Given an optimisation problem in a convex domain $\Omega \subseteq \mathbb{R}^d$,

$$\begin{array}{lll} \text{minimise} & f(\boldsymbol{w}), & \boldsymbol{w} \in \Omega, \\ \text{subject to} & g_i(\boldsymbol{w}) \leq 0, & i = 1, \ldots, K, \\ & h_j(\boldsymbol{w}) = 0, & j = 1, \ldots, P, \end{array}$$

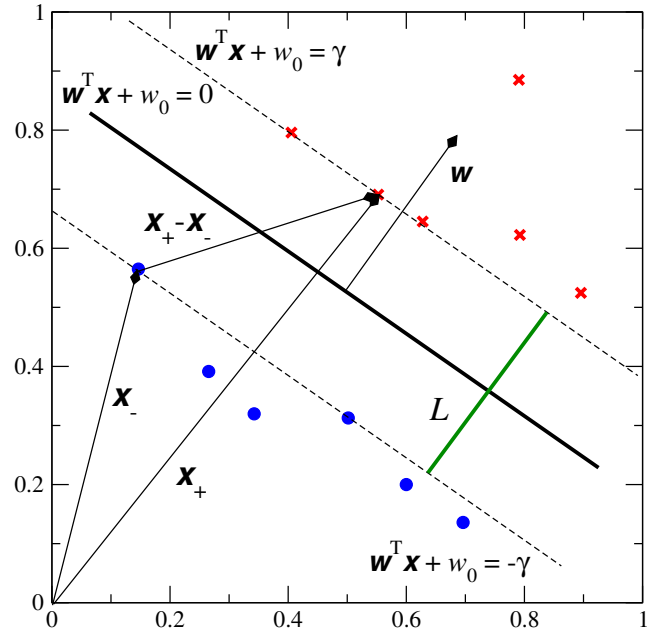
where there exist $\boldsymbol{A} \in \mathbb{R}^{K \times d}$, $\boldsymbol{B} \in \mathbb{R}^{P \times d}$, $\boldsymbol{a} \in \mathbb{R}^K$, and $\boldsymbol{b} \in \mathbb{R}^P$ such that $\boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{a}$ and $\boldsymbol{h}(\boldsymbol{w}) = \boldsymbol{B}\boldsymbol{w} - \boldsymbol{b}$, then the duality gap is zero.

Karush-Kuhn-Tucker Conditions

1. Given an optimisation problem in a convex domain $\Omega \subseteq \mathbb{R}^d$, minimise $f(\boldsymbol{w})$, $\boldsymbol{w} \in \Omega$, subject to $\boldsymbol{g}(\boldsymbol{w}) \leq \mathbf{0}_K$ and $\boldsymbol{h}(\boldsymbol{w}) = \mathbf{0}_P$, with $f \in C^1$ (the derivative exists and it is continuous) and convex (i.e. $f(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2) \leq tf(\boldsymbol{w}_1) + (1-t)f(\boldsymbol{w}_2)$) the necessary and sufficient conditions for a point $\boldsymbol{w}^\star$ to be the optimum is the existence of $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\beta}^\star$ such that:

$$
\begin{aligned}
\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) &= \mathbf{0}_d, \\
\nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{w}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) &= \mathbf{0}_P, \\
\alpha_i^\star g_i(\boldsymbol{w}^\star) &= 0, \qquad i = 1, \ldots, K \\
g_i(\boldsymbol{w}^\star) &\leq 0, \qquad i = 1, \ldots, K \\
\alpha_i^\star &\geq 0, \qquad i = 1, \ldots, K.
\end{aligned}
$$

Maximising the Classification Margin



2. The red crosses satisfy $\mathbf{w}^T\mathbf{x}_+ + w_0 > \gamma$ (classified with $t = 1$) whereas the blue dots satisfy $\mathbf{w}^T\mathbf{x}_- + w_0 < -\gamma$ (classified with $t = -1$).
3. Observe that for all $(\mathbf{x}_n, t_n) \in \mathcal{D}$ we have that $t_n(\mathbf{w}^T\mathbf{x}_n + w_0) - \gamma > 0$.

Maximising the Classification Margin

1. Observe that

$$\frac{\boldsymbol{w}^T}{\|\boldsymbol{w}\|_2}(\boldsymbol{x}_+ - \boldsymbol{x}_-) > \frac{2\gamma}{\|\boldsymbol{w}\|_2} = L,$$

   where the left-hand side is the perpendicular distance between the margins.

2. To maximise the margin L we have to reduce $\|\boldsymbol{w}\|_2 = \sqrt{\sum_{\ell=1}^{d} w_\ell^2}$.

Maximising the Classification Margin

1. The loss function is then $f(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$ and the primal Lagrangian becomes:

$$\mathcal{L}_P(\boldsymbol{w}, w_0, \boldsymbol{\alpha}') = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N} \alpha'_n[\gamma - t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0)]$$

   where $\{\alpha'_n\}$ are the Lagrange multiplier for the N constraints $t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - \gamma > 0$.

2. By defining $\alpha'_n = \frac{\alpha_n}{\gamma}$ and re-scaling in γ w_0 and $\boldsymbol{x}_n$, we can write

$$\mathcal{L}_P(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N} \alpha_n \left[1 - t_n\left(\boldsymbol{w}^T\frac{\boldsymbol{x}_n}{\gamma} + \frac{w_0}{\gamma}\right)\right].$$

KKT Conditions

1. The loss function $f(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$ is convex. Observe that for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \Omega$ and $s \in [0,1]$

$$
\begin{aligned}
f(s\boldsymbol{w}_1 + (1-s)\boldsymbol{w}_2) &= \frac{1}{2}[s\boldsymbol{w}_1 + (1-s)\boldsymbol{w}_2]^T[s\boldsymbol{w}_1 + (1-s)\boldsymbol{w}_2] \\
&= \frac{1}{2}\left\{s^2\boldsymbol{w}_1^T\boldsymbol{w}_1 + 2s(1-s)\boldsymbol{w}_1^T\boldsymbol{w}_2 + (1-s)^2\boldsymbol{w}_2^T\boldsymbol{w}_2\right\} \\
&= sf(\boldsymbol{w}_1) + (1-s)f(\boldsymbol{w}_2) - \frac{1}{2}s(1-s)\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2^2 \\
&\leq sf(\boldsymbol{w}_1) + (1-s)f(\boldsymbol{w}_2).
\end{aligned}
$$

KKT Conditions

1. The first condition produces the following equations:

$$\nabla_{\boldsymbol{w}} \mathcal{L}_P = \mathbf{0}_d = \boldsymbol{w}^\star - \sum_{n=1}^{N} \alpha_n^\star t_n \boldsymbol{x}_n$$

$$\frac{\partial \mathcal{L}_P}{\partial w_0} = 0 = \sum_{n=1}^{N} \alpha_n^\star t_n$$

2. The second condition does not apply because we do not have equation constraints.
3. The last three conditions are:

$$\alpha_n^\star [1 - t_n \left(\boldsymbol{x}_n^T \boldsymbol{w}^\star + w_0^\star \right)] = 0, \qquad n = 1, \ldots, N$$

$$t_n \left(\boldsymbol{x}_n^T \boldsymbol{w}^\star + w_0^\star \right) \geq 1, \qquad n = 1, \ldots, N$$

$$\alpha_n^\star \geq 0, \qquad n = 1, \ldots, N.$$

Dual Lagrangian

1. The dual Lagrangian of the problem is:

$$\begin{aligned}\mathcal{L}_D(\boldsymbol{\alpha}) &= \inf_{\boldsymbol{w}\in\Omega} \mathcal{L}_P(\boldsymbol{w},\boldsymbol{\alpha}) \\ &= -\frac{1}{2}\left(\sum_{n=1}^{N}\alpha_n t_n \boldsymbol{x}_n\right)^T\left(\sum_{n=1}^{N}\alpha_n t_n \boldsymbol{x}_n\right)+\sum_{n=1}^{N}\alpha_n \\ &= -\frac{1}{2}\sum_{n,m=1}^{N}\alpha_n\alpha_m t_n t_m \boldsymbol{x}_n^T\boldsymbol{x}_m+\sum_{n=1}^{N}\alpha_n,\end{aligned}$$

$$\text{subject to } 0=\sum_{n=1}^{N}\alpha_n t_n.$$

2. Through the KKT conditions the *active constraints*, i.e. $t_\ell\left(\boldsymbol{x}_\ell^T\boldsymbol{w}^\star+w_0^\star\right)=1$ are the only ones that have $\alpha_\ell>0$. The $\boldsymbol{x}_\ell$ are the *support vectors.*

3. We can define
$\mathrm{sv} = \{k : 1 \leq k \leq N \text{ and } t_k \left(\boldsymbol{x}_k^T \boldsymbol{w}^\star + w_0^\star\right) = 1\}.$

Example

1. Consider the data set
   $\mathcal{D} = \{([1,1]^T, 1); ([-1,-1]^T, -1); ([1,2]^T, 1)\}$.
2. The Dual Lagrangian is:

$$\mathcal{L}_D = -\frac{1}{2}\left\{2\alpha_1^2 + 2\alpha_2^2 + 5\alpha_3^2 + 4\alpha_1\alpha_2 + 6\alpha_1\alpha_3 + 6\alpha_2\alpha_3\right\} +$$
$$+ \alpha_1 + \alpha_2 + \alpha_3$$
$$\text{subject to } 0 = \alpha_1 - \alpha_2 + \alpha_3.$$

Example

1. The system of linear equations becomes:

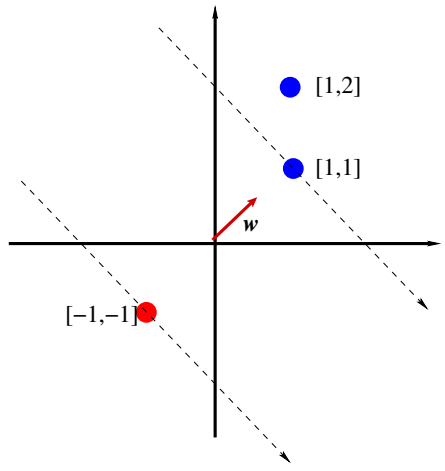
$$2\alpha_1 + 2\alpha_2 + 3\alpha_3 = 1$$
$$2\alpha_1 + 2\alpha_2 + 3\alpha_3 = 1$$
$$3\alpha_1 + 3\alpha_2 + 5\alpha_3 = 1$$

   which, to be satisfied, requires that $\alpha_3 = -1$, which violates one of the KKT conditions. This implies that $[1, 2]^T$ is not a support vector (does not produce an active constraint), therefore we made $\alpha_3 = 0$.

2. This implies that the new system of linear equations is $2\alpha_1 + 2\alpha_2 = 1$ subject to $\alpha_1 = \alpha_2 \left(= \frac{1}{4}\right)$.

Results

1. The weight $\boldsymbol{w}^\star = \frac{1}{2}[1,1]^T$ and bias $w_0 = 1 - [1,1]^T \boldsymbol{w}^\star = 0$ can be immediately computed.

2. 

Remarks

1. The bias cannot be computed from the equations derived from the dual problem. It has to be computed from the primal constraints:

$$w_0^\star = -\frac{\min_{t_\ell=1}\{\boldsymbol{x}_\ell^T \boldsymbol{w}^\star\} + \max_{t_\ell=-1}\{\boldsymbol{x}_\ell^T \boldsymbol{w}^\star\}}{2}.$$

2. The optimal hyperplane can be expressed as:

$$h(\boldsymbol{x}, \boldsymbol{\alpha}^\star, w_0^\star) = \sum_{\ell \in \mathrm{sv}} t_\ell \alpha_\ell^\star \boldsymbol{x}_\ell^T \boldsymbol{x} + w_0^\star.$$

Remarks

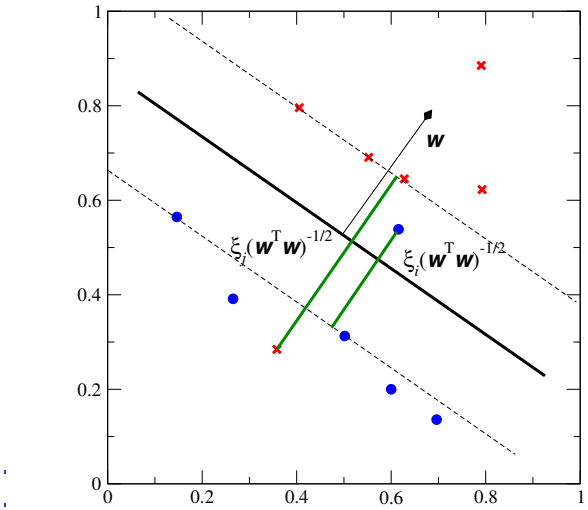
1. Observe that for all $k \in \mathrm{sv}$

$$t_k h(\boldsymbol{x}_k, \boldsymbol{\alpha}^\star, w_0^\star) = t_k \left(\sum_{\ell \in \mathrm{sv}} t_\ell \alpha_\ell^\star \boldsymbol{x}_\ell^T \boldsymbol{x}_k + w_0^\star \right) = 1$$

$$\begin{aligned}
(\boldsymbol{w}^\star)^T \boldsymbol{w}^\star &= \left(\sum_{\ell \in \mathrm{sv}} t_\ell \alpha_\ell^\star \boldsymbol{x}_\ell \right)^T \sum_{k \in \mathrm{sv}} t_k \alpha_k^\star \boldsymbol{x}_k \\
&= \sum_{k \in \mathrm{sv}} \alpha_k^\star t_k \sum_{\ell \in \mathrm{sv}} t_\ell \alpha_\ell^\star \boldsymbol{x}_\ell^T \boldsymbol{x}_k \\
&= \sum_{k \in \mathrm{sv}} \alpha_k^\star (1 - t_k w_0^\star) = \sum_{k \in \mathrm{sv}} \alpha_k^\star - w_0^\star \sum_{k \in \mathrm{sv}} \alpha_k^\star t_k
\end{aligned}$$

$$\|\boldsymbol{w}^\star\|_2 = \sqrt{\sum_{k \in \mathrm{sv}} \alpha_k^\star}$$

Soft Margins



Consider the situation presented in the figure, where the problem is not linearly separable (due to noise in the data set).

2. We can soften the margins by considering the following constraints:

$$t_k(\mathbf{w}^T\mathbf{x}_k + w_0) \geq 1 - \xi_k$$
$$\xi_k \geq 0,$$

where $\{\xi_k\}$ is a set of suitable slack variables.

Soft Margins- Optimisation Problem

1. The solution to the constraint optimisation problem requires now to reduce the size of $\boldsymbol{w}$ and $\boldsymbol{\xi}$ while satisfying the constraints $t_k(\boldsymbol{w}^T\boldsymbol{x}_k + w_0) \geq 1 - \xi_k$ ($\xi_k \geq 0$ is not needed, allow $\xi_k < 0$ will not change the optimal solution).
2. The problem becomes:

$$\begin{aligned} &\text{minimise} && \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}, && \boldsymbol{w} \in \Omega_{\boldsymbol{w}},\ \boldsymbol{\xi} \in \Omega_{\boldsymbol{\xi}}, \\ &\text{subject to} && t_k(\boldsymbol{w}^T\boldsymbol{x}_k + w_0) \geq 1 - \xi_k, && k = 1, \ldots, N. \end{aligned}$$

3. The primal Lagrangian is

$$\mathcal{L}_P(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} - \sum_{n=1}^{N} \alpha_n[t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - 1 + \xi_n].$$

Soft Margins- Optimisation Problem

1. The equations to be solved are:

$$\nabla_{\boldsymbol{w}}\mathcal{L} = \boldsymbol{w}^\star - \sum_{n=1}^{N} \alpha_n^\star t_n \boldsymbol{x}_n = \mathbf{0}_d$$
$$\nabla_{\boldsymbol{\xi}}\mathcal{L} = C\boldsymbol{\xi}^\star - \boldsymbol{\alpha}^\star = \mathbf{0}_N$$
$$\partial_{w_0}\mathcal{L} = \boldsymbol{t}^T\boldsymbol{\alpha}^\star = 0.$$

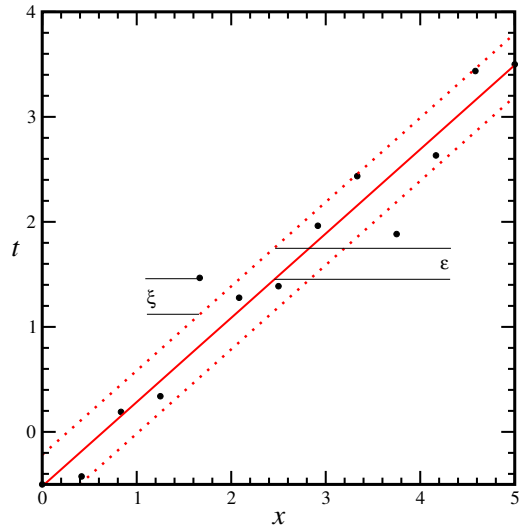
2. The dual Lagrangian becomes:

$$\text{maximise} \qquad \mathcal{L}_D(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{m,n=1}^{N} t_m t_n \alpha_m \alpha_n \left(\boldsymbol{x}_m^T\boldsymbol{x}_n + \frac{\delta_{m,n}}{C}\right)$$
$$\text{subject to} \qquad \boldsymbol{t}^T\boldsymbol{\alpha} = 0, \qquad \alpha_n \geq 0,\ 1 \leq n \leq N.$$

3. The corresponding KKT complementary conditions are: $\alpha_i[t_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 + \xi_i] = 0.$

SVM for Regression

1. Relies on a loss function that ignores errors within a margin ε of the expected value. The variable ξ measures the cost of errors.



2. We define the slack variables
$\xi_n = \max\{0, |t_n - \boldsymbol{w}^T \boldsymbol{x}_n - w_0| - \varepsilon\}$.

Quadratic ε-Insensitive Loss

1. We define the quadratic ε-insensitive loss as:

$$L_2^{\varepsilon}(\boldsymbol{x}, t, f) = \left(\max\left\{0, |t - f(\boldsymbol{x})| - \varepsilon\right\}\right)^2.$$

2. The objective function for the problem is:

$$R\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N} L_2^{\varepsilon}(\boldsymbol{x}_n, t_n, f),$$

   for a suitable $R \in \mathbb{R}^+$.

3. The primal problem is defined as:

$$\begin{aligned}
\text{minimise} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\left(\boldsymbol{\xi}^T\boldsymbol{\xi} + \hat{\boldsymbol{\xi}}^T\hat{\boldsymbol{\xi}}\right), && \boldsymbol{w} \in \Omega_{\boldsymbol{w}},\ \boldsymbol{\xi}, \hat{\boldsymbol{\xi}} \in \Omega_{\boldsymbol{\xi}}, \\
\text{subject to} \quad & \boldsymbol{w}^T\boldsymbol{x}_k + w_0 - t_k \leq \varepsilon + \xi_k, && k = 1, \ldots, N, \\
& t_k - \boldsymbol{w}^T\boldsymbol{x}_k + w_0 \leq \varepsilon + \hat{\xi}_k, && k = 1, \ldots, N, \\
& \xi_k, \hat{\xi}_k \geq 0, && k = 1, \ldots, N.
\end{aligned}$$

Quadratic ε-Insensitive Loss

1. The primal Lagrangian is

$$\mathcal{L}_P(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\left(\boldsymbol{\xi}^T\boldsymbol{\xi} + \hat{\boldsymbol{\xi}}^T\hat{\boldsymbol{\xi}}\right) + \\ + \sum_{n=1}^{N} \alpha_n \left(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - t_n - \varepsilon - \xi_n\right) + \\ + \sum_{n=1}^{N} \hat{\alpha}_n \left(t_n - \boldsymbol{w}^T\boldsymbol{x}_n - w_0 - \varepsilon - \hat{\xi}_n\right)$$

Quadratic ε-Insensitive Loss

1. The stability conditions are:

$$\nabla_{\boldsymbol{w}} \mathcal{L}_P = \boldsymbol{w} - \sum_{n=1}^{N} (\hat{\alpha}_n - \alpha_n) \boldsymbol{x}_n = \mathbf{0}_d,$$

$$\nabla_{\boldsymbol{\xi}} \mathcal{L}_P = C\boldsymbol{\xi} - \boldsymbol{\alpha} = \mathbf{0}_N,$$

$$\nabla_{\hat{\boldsymbol{\xi}}} \mathcal{L}_P = C\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{\alpha}} = \mathbf{0}_N,$$

$$\partial_{w_0} \mathcal{L}_P = \sum_{n=1}^{N} (\hat{\alpha}_n - \alpha_n) = 0.$$

Quadratic ε-Insensitive Loss

1. The dual Lagrangian is:

$$\mathcal{L}_D(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = \sum_{n=1}^{N} t_n(\hat{\alpha}_n - \alpha_n) - \varepsilon \sum_{n=1}^{N}(\alpha_n + \hat{\alpha}_n) - \\ - \frac{1}{2} \sum_{m,n=1}^{N} (\hat{\alpha}_m - \alpha_m)(\hat{\alpha}_n - \alpha_n)\boldsymbol{x}_m^T \boldsymbol{x}_n - \\ - \frac{1}{2C}\left(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \hat{\boldsymbol{\alpha}}^T\hat{\boldsymbol{\alpha}}\right)$$

$$\text{subject to } \sum_{n=1}^{N}(\hat{\alpha}_n - \alpha_n) = 0.$$

Quadratic ε-Insensitive Loss

1. The KKT complementary conditions are:

$$\begin{aligned}
\alpha_n \left(\boldsymbol{w}^T \boldsymbol{x}_n + w_0 - t_n - \varepsilon - \xi_n \right) &= 0, \qquad && n = 1, \ldots, N, \\
\hat{\alpha}_n \left(t_n - \boldsymbol{w}^T \boldsymbol{x}_n - w_0 - \varepsilon - \hat{\xi}_n \right) &= 0, \qquad && n = 1, \ldots, N, \\
\alpha_n \hat{\alpha}_n &= 0, \qquad && i = 1, \ldots, N, \\
\xi_n \hat{\xi}_n &= 0, \qquad && i = 1, \ldots, N,
\end{aligned}$$

the last two conditions indicate that it is not possible a data point, at the same time, to be wrong for excess and defect.

Quadratic ε-Insensitive Loss

1. The dual becomes:

$$\begin{aligned}
\mathcal{L}_D(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) &= \sum_{n=1}^{N} t_n(\hat{\alpha}_n - \alpha_n) - \varepsilon \sum_{n=1}^{N}(\alpha_n + \hat{\alpha}_n) - \\
&\quad - \frac{1}{2} \sum_{m,n=1}^{N} (\hat{\alpha}_m - \alpha_m)(\hat{\alpha}_n - \alpha_n) \boldsymbol{x}_m^T \boldsymbol{x}_n - \\
&\quad - \frac{1}{2C} \left(\boldsymbol{\alpha}^T \boldsymbol{\alpha} + \hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{\alpha}} \right) + \frac{1}{C} \sum_{n=1}^{N} \alpha_n \hat{\alpha}_n \\
&= \sum_{n=1}^{N} t_n(\hat{\alpha}_n - \alpha_n) - \varepsilon \sum_{n=1}^{N}(\alpha_n + \hat{\alpha}_n) - \\
&\quad - \frac{1}{2} \sum_{m,n=1}^{N} (\hat{\alpha}_m - \alpha_m)(\hat{\alpha}_n - \alpha_n) \left(\boldsymbol{x}_m^T \boldsymbol{x}_n + \frac{\delta_{m,n}}{C} \right)
\end{aligned}$$

Quadratic ε-Insensitive Loss

1. By making the substitution $\boldsymbol{\beta} = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$ and by using that $\hat{\alpha}_n \alpha_n = 0$ for all $1 \leq n \leq N$,

$$\mathcal{L}_D(\boldsymbol{\beta}) = \sum_{n=1}^{N} t_n \beta_n - \varepsilon \sum_{n=1}^{N} |\beta_n| - \frac{1}{2} \beta_m \beta_n \left(\boldsymbol{x}_m^T \boldsymbol{x}_n + \frac{\delta_{m,n}}{C} \right)$$

$$\text{subject to } \sum_{n=1}^{N} \beta_n = 0.$$