# 3. Support Vector Machines

COMP0078: Supervised Learning

Mark Herbster

17 October 2022

University College London
Department of Computer Science

svm22-v1

# Acknowledgments and References

**Thanks**

Thanks to Massi Pontil for many of the slides.

# Today

- Optimal Separating Hyperplane
- Soft Margin Separation
- Support Vector Machines
- Connection to Regularisation

# Optimal Separating Hyperplane

## Part I

Optimal Separating Hyperplane

# Separating hyperplane – 1

Let $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \in \mathbf{R}^n \times \{-1, 1\}$ be a training set

By a **hyperplane** we mean a set $H_{\boldsymbol{w},b} = \{\boldsymbol{x} \in \mathbf{R}^n : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}$ (affine linear space) parameterized by $\boldsymbol{w} \in \mathbf{R}^n$ and $b \in \mathbf{R}$

We assume that the data are linearly separable, that is, there exist $\boldsymbol{w} \in \mathbf{R}^n$ and $b \in \mathbf{R}$ such that

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0, \quad i = 1, \ldots, m \tag{1}$$

in which case we call $H_{\boldsymbol{w},b}$ a **separating hyperplane**

Note that we require the inequality in eq.(1) to be strict (we do not admit that the data lie on a hyperplane)

# Separating hyperplane – 2

The distance $\rho_{\boldsymbol{x}}(\boldsymbol{w}, b)$ of a point $\boldsymbol{x}$ from a hyperplane $H_{\boldsymbol{w},b}$ is

$$\rho_{\boldsymbol{x}}(\boldsymbol{w}, b) := \frac{|\boldsymbol{w}^\top \boldsymbol{x} + b|}{\|\boldsymbol{w}\|}$$

If $H_{\boldsymbol{w},b}$ separates the training set $S$ we define its **margin** as

$$\rho_S(\boldsymbol{w}, b) := \min_{i \in [m]} \rho_{\boldsymbol{x}_i}(\boldsymbol{w}, b)$$

If $H_{\boldsymbol{w},b}$ is a hyperplane (separating or not) we also define the *margin of a point* $\boldsymbol{x}$ as $\frac{\boldsymbol{w}^\top \boldsymbol{x} + b}{\|\boldsymbol{w}\|}$ (note that this can be positive or negative)

# Separating hyperplane – 3

Let's verify the observations on the previous slide.

The projection from a point $\boldsymbol{x}$ to $H_{\boldsymbol{w},b}$ is

$$\boldsymbol{p} = \boldsymbol{x} - \frac{\boldsymbol{w}(b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)}{\|\boldsymbol{w}\|^2}$$

To verify this we check that i) $\boldsymbol{p}$ is on the hyperplane and ii) $\boldsymbol{x} - \boldsymbol{p}$ is orthogonal to $\boldsymbol{p} - \boldsymbol{x}'$ where $\boldsymbol{x}'$ is some other point on the hyperplane $H_{\boldsymbol{w},b}$.

Verifying i: We observe that,

$$\langle \boldsymbol{w}, \boldsymbol{p} \rangle + b = \langle \boldsymbol{w}, \boldsymbol{x} \rangle - \frac{\langle \boldsymbol{w}, \boldsymbol{w} \rangle (b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)}{\|\boldsymbol{w}\|^2} + b = 0 .$$

Verifying ii: We observe that

$$\begin{aligned} \langle \boldsymbol{p} - \boldsymbol{x}, \boldsymbol{p} - \boldsymbol{x}' \rangle &= \left\langle -\frac{\boldsymbol{w}(b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)}{\|\boldsymbol{w}\|^2}, \boldsymbol{p} - \boldsymbol{x}' \right\rangle \\ &= \frac{(b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)^2}{\|\boldsymbol{w}\|^2} - \left\langle \boldsymbol{x} - \boldsymbol{x}', \frac{\boldsymbol{w}(b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)}{\|\boldsymbol{w}\|^2} \right\rangle \\ &= 0 \quad \%\% \quad \text{since } \langle \boldsymbol{w}, \boldsymbol{x}' \rangle + b = 0. \end{aligned}$$

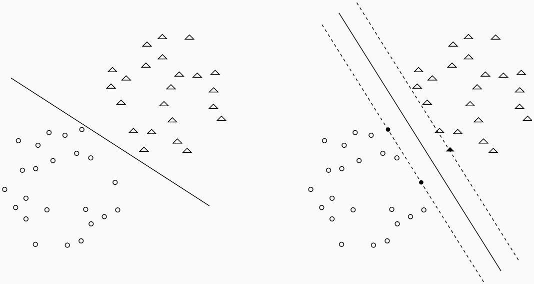
Computing the distance $\rho_{\boldsymbol{x}}(\boldsymbol{w}, b) = \|\boldsymbol{x} - \boldsymbol{p}\|$ of a point $\boldsymbol{x}$ from a hyperplane $H_{\boldsymbol{w},b}$ is

$$\sqrt{\langle \boldsymbol{p} - \boldsymbol{x}, \boldsymbol{p} - \boldsymbol{x} \rangle} = \sqrt{\left\langle \frac{\boldsymbol{w}(b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)}{\|\boldsymbol{w}\|^2}, \frac{\boldsymbol{w}(b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle)}{\|\boldsymbol{w}\|^2} \right\rangle} = \frac{|b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle|}{\|\boldsymbol{w}\|} .$$

# Optimal separating hyperplane (OSH)

This is the separating hyperplane with maximum margin. It solves the optimization problem

$$\rho(S) := \max_{\boldsymbol{w},b} \min_{1\leq i\leq m} \left\{ \frac{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} \;:\; y_j(\boldsymbol{w}^\top \boldsymbol{x}_j + b) \geq 0,\ j = 1, \ldots, m \right\} > 0$$



(a) (b)

# Choosing a parameterization

A separating hyperplane is parameterized by $(\boldsymbol{w}, b)$, but this choice is not unique (rescaling with a positive constant gives the same separating hyperplane). Two possible ways to fix the parameterization:

- *Normalized hyperplane:* set $\|\boldsymbol{w}\| = 1$, in which case $\rho_{\boldsymbol{x}}(\boldsymbol{w}, b) = |\boldsymbol{w}^\top \boldsymbol{x} + b|$ and $\rho_S(\boldsymbol{w}, b) = \min_{i \in [m]} y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)$
- *Canonical hyperplane:* choose $\|\boldsymbol{w}\|$ such that $\rho_S(\boldsymbol{w}, b) = \frac{1}{\|\boldsymbol{w}\|}$, i.e. we require that $\min_{i \in [m]} y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$ (a data-dependent parameterization)

We will mainly work with the second parameterization

# Optimal separating hyperplane

- If we work with normalized hyperplanes we have

$$\rho(S) = \max_{\boldsymbol{w},b} \min_{i} \{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \ : \ y_j(\boldsymbol{w}^\top \boldsymbol{x}_j + b) \geq 0, \|\boldsymbol{w}\| = 1, j \in [m]\}$$

- If we work with canonical hyperplanes, instead, we have

$$\begin{aligned}
\rho(S) &= \max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|} : \min_{i\in[m]} \{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\} = 1, y_j(\boldsymbol{w}^\top \boldsymbol{x}_j + b) \geq 0 \right\} \\
&= \max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|} : y_i(\boldsymbol{w}^\top \boldsymbol{x}_j + b) \geq 1, i \in [m] \right\} \\
&= \frac{1}{\min_{\boldsymbol{w},b} \{\|\boldsymbol{w}\| : y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, i \in [m]\}}
\end{aligned}$$

# Optimal separating hyperplane (cont.)

We choose to work with canonical hyperplanes and, so, look at the optimization problem

Problem **P1**

Minimize $\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} \qquad (\boldsymbol{w} \in \mathbf{R}^n)$

subject to $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ i = 1, \ldots, m,$

The quantity $1/\|\boldsymbol{w}\|$ is the **margin** of the OSH

# Solving a constrained optimisation with a Lagrangian

## Problem

Minimise a convex function subject to linear inequality constraints

$$f(\boldsymbol{z}) : A\boldsymbol{z} \leq \boldsymbol{c}$$

Where $f : \mathbf{R}^n \to \mathbf{R}$, is differentiable and the matrix $A$ is $m \times n$ and $\boldsymbol{c} \in \mathbf{R}^m$.

**If the optimisation is *feasible* that is $\{z : Az \leq c\}$ is non-empty then we may solve by forming the Lagrangian,**

$$L(\boldsymbol{z}, \boldsymbol{\alpha}) := f(\boldsymbol{z}) + \boldsymbol{\alpha}^\top (A\boldsymbol{z} - \boldsymbol{c}) \,,$$

**and we have that**

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \min_{\boldsymbol{z}} L(\boldsymbol{z}, \boldsymbol{\alpha}) = \min_{\boldsymbol{z}} f(\boldsymbol{z}) : A\boldsymbol{z} \leq \boldsymbol{c} \,.$$

## Necessary and sufficient conditions (KKT) for a solution $(\bar{\alpha}, \bar{z})$

1. $A\bar{\boldsymbol{z}} \leq \boldsymbol{c}$
2. $\bar{\boldsymbol{\alpha}} \geq \mathbf{0}$
3. $\nabla_z L|_{\bar{z}} = \mathbf{0}$
4. $(A\bar{\boldsymbol{z}} - \boldsymbol{c})_i \, \bar{\alpha}_i = 0 \quad i = 1, \ldots, m$ (“complementary slackness”)

# Saddle point

The solution of problem **P1** is equivalent to determine the **saddle point** of the Lagrangian function

$$L(\boldsymbol{w}, b; \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i \left[y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\right] \tag{2}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers

We minimize $L$ over $(\boldsymbol{w}, b)$ and maximise over $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} \geq \mathbf{0}$. Differentiating w.r.t $\boldsymbol{w}$ and $b$ we obtain:

$$\begin{aligned} \frac{\partial L}{\partial b} &= -\sum_{i=1}^{m} y_i \alpha_i = 0 && (3) \\ \frac{\partial L}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i = \mathbf{0} \;\Rightarrow \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i && (4) \end{aligned}$$

# Dual problem

Expanding (2) we have

$$\frac{1}{2}\overbrace{\boldsymbol{w}^\top \boldsymbol{w}}^{\boldsymbol{\alpha}^\top \boldsymbol{A}\boldsymbol{\alpha}} - \overbrace{\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w}^\top \boldsymbol{x}_i}^{\boldsymbol{\alpha}^\top \boldsymbol{A}\boldsymbol{\alpha}} - \overbrace{b\sum_{i=1}^{m} \alpha_i y_i}^{0} + \sum_{i=1}^{m} \alpha_i$$

Substituting (3) and (4) in eq.(2) and defining the $m \times m$ matrix $\boldsymbol{A} := (y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j : i, j = 1, \ldots, m)$ leads to the **dual problem**.

Problem **P2**

Maximize $\quad Q(\boldsymbol{\alpha}) := -\frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{A}\boldsymbol{\alpha} + \sum_i \alpha_i$

subject to $\quad \sum_i y_i \alpha_i = 0$

$\alpha_i \geq 0, \qquad i = 1, \ldots, m$

Note that the complexity of this problem depends on $m$, not on the number of input components $n$ (same as ridge regression)

# Kuhn-Tucker conditions and support vectors

If $\bar{\alpha}$ is a solution of the dual problem then the solution $(\bar{\boldsymbol{w}}, \bar{b})$ of the primal problem is given by

$$\bar{\boldsymbol{w}} = \sum_{i=1}^{m} \bar{\alpha}_i y_i \boldsymbol{x}_i$$

Note that $\bar{\boldsymbol{w}}$ is a linear combination of only the $\boldsymbol{x}_i$ for which $\bar{\alpha}_i > 0$. These $\boldsymbol{x}_i$ are termed **support vectors** (SVs)

Parameter $\bar{b}$ can be determined by looking at the Kuhn-Tucker conditions ("complementary slackness")

$$\bar{\alpha}_i \left( y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) - 1 \right) = 0$$

Specifically if $\boldsymbol{x}_j$ is a SV we have that

$$\bar{b} = y_j - \bar{\boldsymbol{w}}^\top \boldsymbol{x}_j$$

# Some remarks

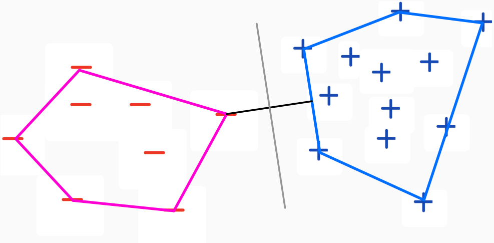
- The fact that the OSH is determined only by the SVs is most remarkable. Usually, the support vectors are a small subset of the training data
- All the information contained in the data set is summarized by the support vectors: The whole data set could be replaced by only these points and the **same** hyperplane would be found
- A new point $\boldsymbol{x}$ is classified as $\operatorname{sgn}\left(\sum_{i=1}^{m} y_i \bar{\alpha}_i \boldsymbol{x}_i^\top \boldsymbol{x} + \bar{b}\right)$

**Connection between number of support vectors and generalization**

Let $n_{\mathsf{SV}}$ equal the expected number of support vectors in an SVM trained on $m$ examples sampled IID then the expected generalization error of an SVM trained on $m-1$ examples is bounded above by $n_{\mathsf{SV}}/m$.

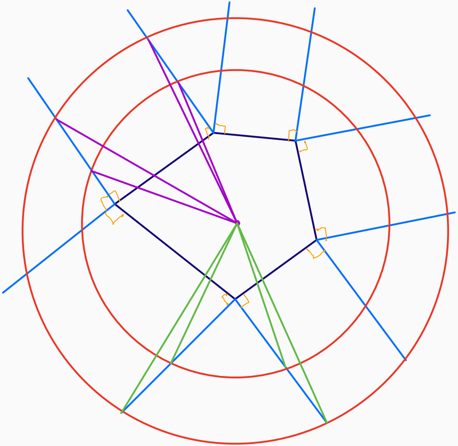
# Intuitions – 1

1. We've argued qualitatively that **sparsity** of support vectors implies good generalisation performance.
2. We've argued intuitively that large margin implies good generalisation performance. Note: *there are a number of arguments in literature which prove qualitative bounds on the generalisation error based on the margin but for our purposes these are beyond the scope of the class*.
3. We will give an intuitive argument that all else being equal a larger margin implies fewer support vectors.
4. The SVM optimisation problem can be shown to be equivalent to finding the (a) shortest line segment that connects the hull of the convex positive points to the negative points and then the perpendicular bisecting hyperplane of the line is the linear classifier.



# Intuitions – 2

1. The number of support vectors associated with a **face** is the number of vertices.
2. All else being equal a point nearer to a convex polytope tend to be “nearer” to a larger dimensional face than a “farther” point.



# Linearly nonseparable case – 1

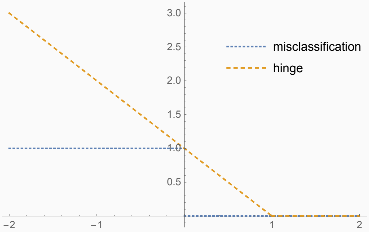
Ideally we would like to minimise

$$\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + C\sum_{i=1}^{m} V_{\mathrm{mc}}(y_i, \boldsymbol{w}^\top\boldsymbol{x}_i + b)$$

However this is known to be NP-hard! So instead we convexify by replacing the misclassification loss by the hinge loss

$$\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + C\sum_{i=1}^{m} V_{\mathrm{hinge}}(y_i, \boldsymbol{w}^\top\boldsymbol{x}_i + b) \tag{5}$$

$$V_{\mathrm{mc}}(y, \hat{y}) = \mathcal{I}[y = \mathrm{sign}(\hat{y})]$$
$$V_{\mathrm{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y})$$



$V_{\mathbf{mc}}(\mathbf{1}, \hat{y})$, **and** $V_{\mathbf{hinge}}(\mathbf{1}, \hat{y})$, .

We now have a convex (quadratic) optimisation problem.

# Linearly nonseparable case

Observe we can rewrite (5) as

Problem **P3**

Minimize $\frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{m} \xi_i$

subject to $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i,$

$\xi_i \geq 0, \qquad i = 1, \ldots, m$

**Note:** The slack variables $\xi_i$ relax the separation constraints ($\xi_i > 0 \Rightarrow y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) < 1$ i.e., the margin is less than 1).

# Saddle point

The solution of problem **P3** is equivalent to determining the **saddle point** of the Lagrangian function

$$L(\boldsymbol{w}, \boldsymbol{\xi}, b; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \left[y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) + \xi_i - 1\right] - \sum_{i=1}^{m} \beta_i \xi_i \tag{6}$$

where $\alpha_i, \beta_i \geq 0$ are the Lagrange multipliers

We minimize $L$ over $(\boldsymbol{w}, \boldsymbol{\xi}, b)$ and maximize over $\boldsymbol{\alpha}, \boldsymbol{\beta}$ with $\boldsymbol{\alpha}, \boldsymbol{\beta} \geq 0$. Differentiating w.r.t $\boldsymbol{w}$, $\boldsymbol{\xi}$, and $b$ we obtain:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= -\sum_{i=1}^{m} y_i \alpha_i = 0 \\
\frac{\partial L}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i = \boldsymbol{0} \ \Rightarrow \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \qquad (7) \\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0 \ \Rightarrow 0 \leq \alpha_i \leq C
\end{aligned}$$

# New dual problem

Analogous to (**P2**) substituting (7) into (6) leads to the dual problem.

Problem **P4**

Maximize $Q(\boldsymbol{\alpha}) := -\frac{1}{2}\boldsymbol{\alpha}^\top A\boldsymbol{\alpha} + \sum_i \alpha_i$

subject to $\sum_i y_i\alpha_i = 0$

$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m$

This is like problem **P2** except that now we have “box constraints" on $\alpha_i$. If the data is linearly separable, by choosing $C$ large enough we obtain the OSH

# Nonseparable case (cont)

Again we have

$$\bar{\boldsymbol{w}} = \sum_{i=1}^{m} \bar{\alpha}_i y_i \boldsymbol{x}_i,$$

while $\bar{b}$ can be determined from $\bar{\boldsymbol{\alpha}}$, solution of the problem **P4**, and from the new Kuhn-Tucker conditions ("complementary slackness")

$$\begin{array}{rcll} \bar{\alpha}_i \left( y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) - 1 + \bar{\xi}_i \right) & = & 0 & (*) \\ (C - \bar{\alpha}_i)\bar{\xi}_i & = & 0 & (**) \end{array}$$

Where (**) follows since $\beta_i = C - \alpha_i$. Again, points for which $\bar{\alpha}_i > 0$ are termed **support vectors**

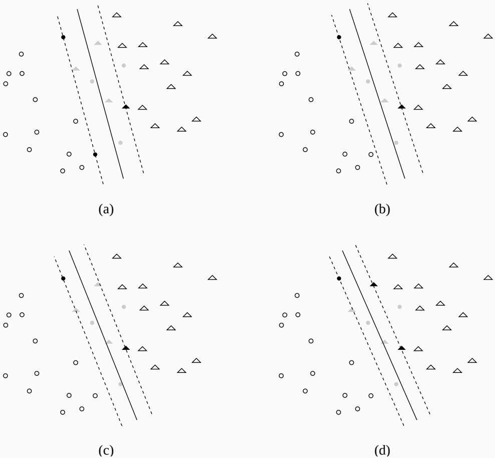
# A closer look at the KKT conditions

Equation (*) and (**) tell us that if

- $y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) > 1 \Rightarrow \bar{\alpha}_i = 0$  (not a SV)
- $y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) < 1 \Rightarrow \bar{\alpha}_i = C$  (a SV with positive slack $\bar{\xi}_i$)
- $y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) = 1 \Rightarrow \bar{\alpha}_i \in [0, C]$   (if $\bar{\alpha}_i > 0$ a SV "on the margin")

**Remark:** Conversely, from eqs.(*),(**) if $\bar{\alpha}_i = 0$ then $y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) \geq 1, \bar{\xi}_i = 0$; if $\bar{\alpha}_i \in (0, C)$ then $y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) = 1, \bar{\xi}_i = 0$; if $\bar{\alpha}_i = C$ then $y_i(\bar{\boldsymbol{w}}^\top \boldsymbol{x}_i + \bar{b}) \leq 1,\ \bar{\xi}_i \geq 0$

# The role of the parameter $C$



(a)

(b)

(c)

(d)

Optimal separating hyperplane for four increasing values of $C$. Both the margin and the training error are non-increasing functions of $C$

# The role of the parameter $C$ (cont.)

The parameter $C$ controls the trade-off between $\|\boldsymbol{w}\|^2$ and the training error $\sum_{i=1}^{m} \xi_i$

It can be shown that the optimal value of the Lagrange multipliers $\bar{\alpha}_i$ (and, so, $\bar{\boldsymbol{w}}, \bar{b}$) are piecewise quadratic functions of $C$. This helps re-computing the solution when varying $C$.

- $C$ is often selected by minimizing the leave-one-out (LOO) cross validation error.

### Observations on computing LOO

1. We need "retrain" the SVM no more than #SVs times (Why?) – retraining is "fast."
2. Alternatively observe that rather than compute LOO one could use $\frac{\#\text{SVs}}{m}$ as an upper bound on the LOO CV error.

# Support Vector Machines (SVMs)

The above analysis holds true if we work with a feature map $\phi : \mathcal{X} \to \mathcal{W}$. We simply replace $\boldsymbol{x}$ by $\phi(\boldsymbol{x})$ and $\boldsymbol{x}^\top \mathbf{t}$ by $\langle \phi(\boldsymbol{x}), \phi(\mathbf{t}) \rangle = K(\boldsymbol{x}, \mathbf{t})$

An SVM with kernel $K$ is the function

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b, \quad \boldsymbol{x} \in \mathcal{X}$$

where the parameters $\alpha_i$ solve problem **P4** with $\boldsymbol{A} = (y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) : i, j = 1, \dots, m)$ and $b$ is obtained as discussed above

A new point $\boldsymbol{x} \in \mathcal{X}$ is classified as $\operatorname{sgn}(f(\boldsymbol{x}))$

# Connection to regularization

The SVM formulation above is equivalent to the problem

$$E_\lambda(\boldsymbol{w}, b) = \sum_{i=1}^{m} \max(1 - y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b), 0) + \lambda \|\boldsymbol{w}\|^2$$

with $\lambda = \frac{1}{2C}$

In fact, we have

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \left\{ C \sum_{i=1}^{m} \xi_i + \frac{1}{2}\|\boldsymbol{w}\|^2 : y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b) \geq 1 - \xi_i, \xi_i \geq 0 \right\} =$$

$$\min_{\boldsymbol{w}, b} \left\{ \min_{\boldsymbol{\xi}} \left\{ C \sum_{i=1}^{m} \xi_i + \frac{1}{2}\|\boldsymbol{w}\|^2 : \xi_i \geq 1 - y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b), \xi_i \geq 0 \right\} \right\} =$$

$$\min_{\boldsymbol{w}, b} \left\{ C \sum_{i=1}^{m} \max\left(1 - y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle + b), 0\right) + \frac{1}{2}\|\boldsymbol{w}\|^2 \right\} = C E_{\frac{1}{2C}}(\boldsymbol{w}, b)$$

# SVM regression

SVM's can be developed for regression as well. Here we choose the loss $|y - f(\boldsymbol{x})|_\epsilon = \max(|y - f(\boldsymbol{x})| - \epsilon, 0)$

$$
\begin{array}{ll}
\text{Minimize} & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\
 & \\
\text{subject to} & \boldsymbol{w}^\top \boldsymbol{x}_i + b - y_i \leq \epsilon + \xi_i, \\
 & y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b \leq \epsilon + \xi_i^*, \\
 & \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, m
\end{array}
$$

SVMs loss functions (both for classification and regression) are **scale sensitive**: errors below a certain resolution do not count. This leads to sparse solutions!

# Solution methods

The above optimization problems are Quadratic Programming (QP) problems. Several methods (eg, interior point methods) from convex optimization exist for solving QP problems

If we work with a non-linear kernel, the number of underlying features, $N$, is typically much larger (or infinite) than the number of examples. Thus, we need to solve the dual problem

However, if $m \gg N$ it is more efficient to solve the primal problem

# Decomposition of the dual problem

For large datasets (say $m > 10^5$) it is practically impossible to solve the dual problem with standard optimization techniques (matrix $\boldsymbol{A}$ is dense!)

A typical approach is to iteratively optimize wrt. an “active set" $\mathcal{A}$ of variables. Set $\boldsymbol{\alpha} = 0$, choose $q \leq m$ and a subset $\mathcal{A}$ of $q$ variables, $\mathcal{A} = \{\alpha_{i_1}, \ldots, \alpha_{i_q}\}$. We repeat until convergence:

- Optimize $Q(\alpha)$ wrt. the variables in $\mathcal{A}$
- Remove one variable from $\mathcal{A}$ which satisfies the KKT conditions and add one variable, if any, which violates the KKT conditions. If no such variable exists stop

One can show that after each iteration $Q$ increases

# Suggested Readings

Recommended: *The Elements of Statistical Learning ...*, Chapters 12.1-12.3.

Background on Lagrangian and KKT conditions see Chapter 5 *Convex Optimization* by Boyd and Vandenberghe.

# Going Deeper ...

- *Making Large-Scale SVM Learning Practical.* This is an early work on efficient optimization for SVMs. The state of the art has considerably advanced since this paper but it covers the elementary concepts.
- *When Do Neural Networks Outperform Kernel Methods?.* For what tasks are NNs good? For what tasks do Kernel Methods do well? This paper consider this question in terms of potential latent low-dimensional representations of the data.

# Problems – 1

1. Describe the criterion used by hard margin Support Vector Machines to choose a separating hyperplane for a linearly separable dataset, illustrating with a diagram indicating the margin and the support vectors.
2. For a hard-margin SVM. If we remove one of
   2.1 the examples which is a support vector from the training set,
   2.2 the examples which is not a support vector from the training set

   and retrain with out that example. How does the maximum margin change?
3. Consider the optimisation given by

$$\begin{array}{ll} \min_{\mathbf{w},b,\gamma,\xi} & \|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \\ \text{subject to} & y_i\langle \mathbf{w},\mathbf{x}_i\rangle \geq 1-\xi_i,\ \xi_i \geq 0, \\ & i=1,\ldots,m. \end{array}$$

   Which part of the optimisation corresponds to the regulariser? What is the loss function incorporated into the optimisation?

# Problems – 2

1. Assume that the set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \subset \mathbf{R}^2 \times \{-1, 1\}$ of binary examples is strictly linearly separable by a line going through the origin, that is, there exists a vector $\boldsymbol{w} \in \mathbf{R}^2$ such that the linear function $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, $\boldsymbol{x} \in \mathbf{R}^2$ has the property that $y_i f(\boldsymbol{x}_i) > 0$ for every $i = 1, \dots, m$. Consider the optimisation problem (linearly separable SVM):

$$\mathbf{P1}: \quad \text{minimise} \left\{ \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{w} \ : \ y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1, i = 1, \dots, m \right\}.$$

   Argue that the above problem has a unique solution. Describe the geometric meaning of this solution.

2. Show that the vector $\boldsymbol{w}$ solving problem **P1** has the form $\boldsymbol{w} = \sum_{i=1}^m c_i y_i \boldsymbol{x}_i$ where $c_1, \dots, c_m$ are some nonnegative coefficients. [HINT: use the method of Lagrange multipliers]

3. **Show that the coefficients $c_1, \dots, c_m$ in the above formula solve the optimization problem**

$$\mathbf{P2}: \quad \max \left\{ -\frac{1}{2} \sum_{i,j=1}^m c_i c_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j + \sum_{i=1}^m c_i \ : \ c_j \geq 0, \ j = 1, \dots, m \right\}.$$

   **Finally, if $(\hat{c}_1, \dots, \hat{c}_m)$ is the solution to this problem and $\hat{\boldsymbol{w}}$ is the solution to problem P1, argue that $\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{w}} = \sum_{i=1}^m \hat{c}_i$.**