Machine Learning

Discriminant Classification & the Linear SVM

Dariush Hosseini

dariush.hosseini@ucl.ac.uk
Department of Computer Science
University College London

Lecture Overview

Lecture Overview

By the end of this lecture you should:

1. Know the **Linear Support Vector Machine (SVM)** algorithm and its context as a **maximum margin** approach to **Discriminant Classification**
2. Know the **hard** and **soft** formulations of the SVM learning problem, and appreciate that even for the soft version the linear SVM has limitations
3. Be aware of the motivation of the SVM algorithm from **PAC learning**

Lecture Overview

Notation

- **Inputs**
  $\mathbf{x} = [x_1, ..., x_m]^T \in \mathbb{R}^m$

- **Binary Outputs**
  $y \in \{-1, 1\}$

- **Training Data**
  $\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$

- **Data-Generating Distribution**, $\mathcal{D}$
  $\mathcal{S} \sim \mathcal{D}$

Classification Problem

- **Representation**

$$f \in \mathcal{F}$$

- **Evaluation**
  - **Loss Measure:**

$$\mathcal{E}[f(\mathbf{x}), y] = \mathbb{I}[y \neq f(\mathbf{x})]$$

  - **Generalisation Loss:**

$$\mathsf{L}(\mathcal{E}, \mathcal{D}, f) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$$

    Where $\mathcal{D}$ is characterised by $p_{\mathcal{X},\mathcal{Y}}(\mathbf{x}, y) = p_{\mathcal{Y}}(y|\mathbf{x})p_{\mathcal{X}}(\mathbf{x})$ for some pmf, $p_{\mathcal{Y}|\mathcal{X}}$, and some pdf, $p_{\mathcal{X}}$

- **Optimisation**

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$$

Distribution-Free Classification

- Here we seek to learn the classification boundary (equivalently f^*) directly, without resorting to probabilistic inference
- In other words we seek to learn the **discriminant function** f^* directly
- In particular (initially) we are interested in **linear** discriminants:

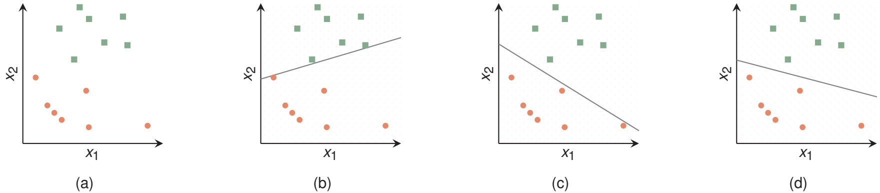
$$f = \text{sign}[\mathbf{w} \cdot \mathbf{x} + b] \qquad \text{where:} \qquad \mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R}$$

Distribution-Free Classification

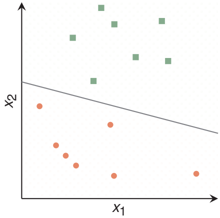
- An example is the PAC approach where we seek to approximate $\mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$ without reference to any explicit pdf and then to optimise this new quantity in order to learn f^*...
- ...But can we motivate **discriminant classification** more intuitively to begin with?

Margins

- Let us seek **linear** discriminants
- We want to learn a decision boundary that splits the input space so as to classify positive and negative instances
- Which boundary is the best?



Margins



- The ideal decision boundary is the line which runs halfway between the datapoints providing **maximum padding** for both classes
- The measure of this maximum padding is the perpendicular distance of the nearest point to the hyperplane - this is the **margin**
- So our goal is to find the decision boundary that has the **maximum margin** with respect to the training instances

Margins

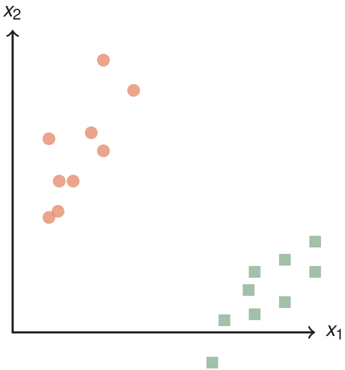
- Why?
- Intuition is that a large margin results in a **safer** boundary for which unseen test points are less likely (in some sense) to fall on the wrong side of the boundary
- Margin is somehow linked with **generalisation**

Lecture Overview

Separability

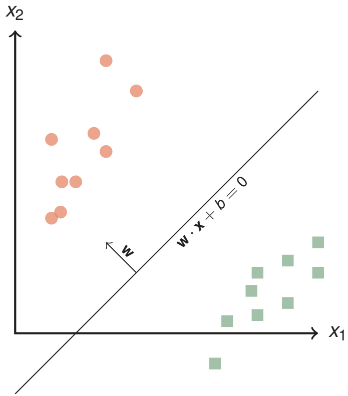
- Let us assume that the training data can be separated
- Let us seek the linear discriminant which maximises the margin
- We will proceed **geometrically**

Problem Motivation



- **Red circles** are classified $y = 1$, **Green squares** are classified $y = -1$

Problem Motivation: Separating Hyperplane



Problem Motivation: Separating Hyperplane

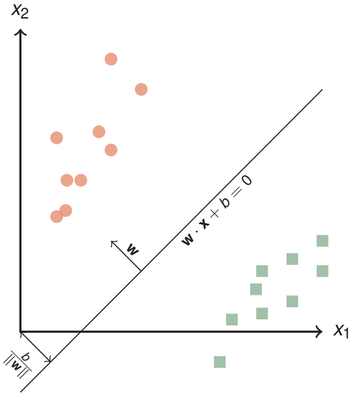
- The separating hyperplane is defined by:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

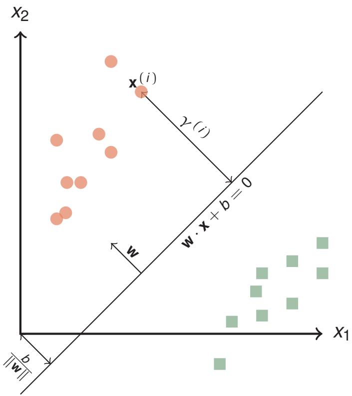
- For some point, $\widetilde{\mathbf{x}}$, the point on the hyperplane which is closest to the origin, the perpendicular distance to the origin is given by:

$$\begin{aligned} \mathbf{w} \cdot \widetilde{\mathbf{x}} + b &= 0 \\ \implies -\|\mathbf{w}\|\|\widetilde{\mathbf{x}}\| + b &= 0 \\ \implies \|\widetilde{\mathbf{x}}\| &= \frac{b}{\|\mathbf{w}\|} \end{aligned}$$

Problem Motivation: Separating Hyperplane



Problem Motivation: Margin of $\mathbf{x}^{(i)}$



Problem Motivation: Margin of $\mathbf{x}^{(i)}$

- The margin of some point, $\mathbf{x}^{(i)}$, is the perpendicular distance between the hyperplane and that point:

Problem Motivation: Margin of $\mathbf{x}^{(i)}$

- The margin of some point, $\mathbf{x}^{(i)}$, is the perpendicular distance between the hyperplane and that point:

For green squares :

$$\gamma^{(i)} = -\frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} - \frac{b}{\|\mathbf{w}\|}$$

For red circles :

$$\gamma^{(i)} = \frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}$$

Problem Motivation: Margin of $\mathbf{x}^{(i)}$

- The margin of some point, $\mathbf{x}^{(i)}$, is the perpendicular distance between the hyperplane and that point:

For green squares :

$$\gamma^{(i)} = -\frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} - \frac{b}{\|\mathbf{w}\|}$$

$$\gamma^{(i)} = -\frac{(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|}$$

For red circles :

$$\gamma^{(i)} = \frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}$$

$$\gamma^{(i)} = \frac{(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|}$$

Problem Motivation: Margin of $\mathbf{x}^{(i)}$

- The margin of some point, $\mathbf{x}^{(i)}$, is the perpendicular distance between the hyperplane and that point:

For green squares :

$$\gamma^{(i)} = -\frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} - \frac{b}{\|\mathbf{w}\|}$$

$$\gamma^{(i)} = -\frac{(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|}$$

For red circles :

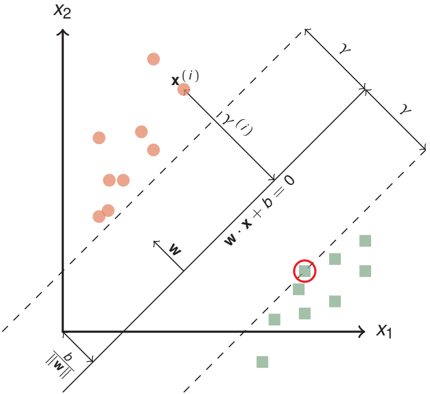
$$\gamma^{(i)} = \frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}$$

$$\gamma^{(i)} = \frac{(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|}$$

- Since, by the **hard margin** assumption, $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) > 0$ for all i, we may express the margin for both red and green points more compactly as:

$$\gamma^{(i)} = \frac{y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|}$$

Problem Motivation: Margin



UCL

Problem Motivation: Margin

- The **margin** of the system, γ, is defined as the smallest $\gamma^{(i)}$:

$$\gamma = \min_i \gamma^{(i)}$$
$$\gamma = \frac{1}{\|\mathbf{w}\|} \min_i \left[y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \right]$$

- Since $\gamma^{(i)}$ is invariant to multiplicative scaling of **w** and b, then w.l.o.g. we may write:

$$\min_i \left[y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \right] = 1$$
$$\implies y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geqslant 1 \qquad \forall i$$
$$\gamma = \frac{1}{\|\mathbf{w}\|}$$

Problem Formulation

- So our optimisation problem becomes:

$$\max_{\mathbf{w},b} \quad \frac{1}{\|\mathbf{w}\|}$$
$$\text{subject to:} \qquad y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geqslant 1 \qquad \forall i$$

- Or equivalently:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{1}$$
$$\text{subject to:} \qquad -y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) + 1 \leqslant 0 \qquad \forall i$$

Problem Solution

- We note that the objective here is **strictly convex** and that the constraints restrict **w**, *b* to be in a **convex set**
- So the **optimal solution** must be **unique** (Recall *Linear Regression Lecture*, Theorem (A.3))
- How should we solve this problem?
- We cannot apply **gradient descent** (without modification) because of constraints
- An alternative is to use **Lagrange Duality** to re-formulate the problem in a form which is more amenable to solution

Lagrange Duality

- First we write the Lagrangian for problem (1):

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha^{(i)}\left(1 - y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + b)\right)$$

where: $\qquad \alpha^{(i)} \geqslant 0$

- The dual objective can be written:

$$\mathcal{D}(\boldsymbol{\alpha}) = \min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$$

Lagrange Duality

- This is an unconstrained optimisation which we can solve by seeking stationary points:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w}^* - \sum_{i=1}^{n} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} = 0 \quad \Longrightarrow \quad \mathbf{w}^* = \sum_{i=1}^{n} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha^{(i)} y^{(i)} = 0$$

- Substituting these expressions back into $\mathcal{D}(\boldsymbol{\alpha})$ yields:

$$\mathcal{D}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$$

Dual Problem

- This leads to the following dual problem:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha^{(i)}\alpha^{(j)} y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$$

subject to:
$$\alpha_i \geqslant 0$$
$$\sum_{i=1}^{n} \alpha^{(i)} y^{(i)} = 0$$

- This is actually a simpler problem to solve than problem (1)
- There exists a bespoke numerical procedure for the solution of this problem, the **SMO** algorithm, which yields $\boldsymbol{\alpha}$

Some Observations

- The KKT **complementary slackness** condition, which must hold at optimality for this problem, tells us:

$$\alpha^{(i)} \left(1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)\right) = 0$$

- Therefore, either:

$$\alpha^{(i)} = 0 \qquad \text{and} \qquad y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) > 1$$

- Or:

$$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) = 1 \qquad \text{and} \qquad \alpha^{(i)} > 0$$

Some Observations

- Only points for which $\alpha_i > 0$ play an active role and contribute to the discriminant function - these points are called **support vectors**
- All other points are redundant - we could discard them and learn the same classifier!
- This feature leads to the **sparsity** property of SVM's
- Also, note that all support vectors sit on the margin hyperplanes defined by $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) = 1$

Primal Optimality

- Using equation (2) we can write:

$$\mathbf{w}^* = \sum_{i \in \mathcal{SV}} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)}$$

- Here $\mathcal{SV}$ is the set of support vectors

Primal Optimality

- We can also generate a value for b^* as follows:

$$
\begin{aligned}
y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) &= 1 \qquad \forall i \in \mathcal{SV} \\
\left(y^{(i)}\right)^2 (\mathbf{w} \cdot \mathbf{x}^{(i)} + b) &= y^{(i)} \\
\mathbf{w} \cdot \mathbf{x}^{(i)} + b &= y^{(i)} \\
b &= y^{(i)} - \mathbf{w} \cdot \mathbf{x}^{(i)} \\
\sum_{i \in \mathcal{SV}} b &= \sum_{i \in \mathcal{SV}} \left(y^{(i)} - \sum_{j \in \mathcal{SV}} \alpha^{(j)*} y^{(j)} \mathbf{x}^{(j)} \cdot \mathbf{x}^{(i)} \right) \\
b &= \frac{1}{|\mathcal{SV}|} \sum_{i \in \mathcal{SV}} \left(y^{(i)} - \sum_{j \in \mathcal{SV}} \alpha^{(j)*} y^{(j)} \mathbf{x}^{(j)} \cdot \mathbf{x}^{(i)} \right)
\end{aligned}
$$

Recap

- **Representation**

$$\mathcal{F} = \left\{ f_{\mathbf{w}}(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \,\middle|\, \mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R} \right\}$$

- **Evaluation**

$$\gamma$$

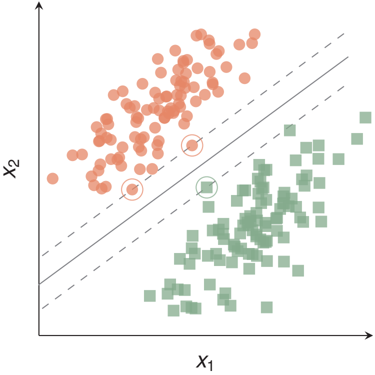
Where: $\gamma = \min_i \gamma^{(i)}$
And: $\gamma^{(i)} = y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geqslant 1$

- **Optimisation**

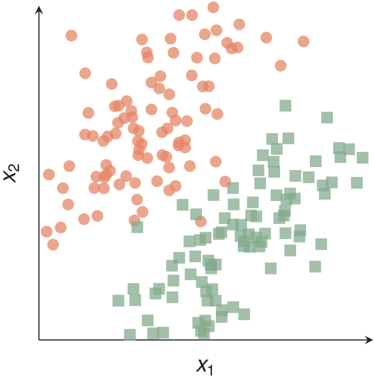
$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to:} \qquad -y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) + 1 \leqslant 0 \qquad \forall i$$

Perfectly Linearly Separable



Noisily Linearly Separable

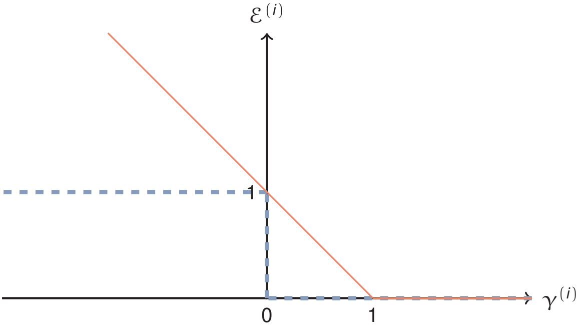


Noisily Linearly Separable

- When two classes are linearly separable, but there is some overlap between them, the hard margin SVM will not find a solution
- To overcome this problem we need to find a mechanism for tolerating errors and so obtain a **soft margin** classifier
- We introduce a new loss function, the **hinge loss**, characterised as:

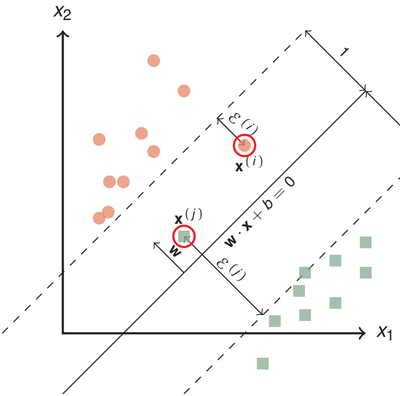
$$\max(0, 1 - \gamma^{(i)})$$

Slack Variables & Hinge Loss



- Note that the loss starts at the margin even for well-classified points
- The hinge loss is a **convex relaxation** of the **misclassification error**...
- ...Which will result in a tractable optimisation

Problem Motivation: Hinge Loss



Slack Variables & Hinge Loss

- We introduce **slack variables**, $\xi^{(i)}$, which are lower bounded by the **hinge loss** function and quantify a measure of error exhibited by a particular data point:

$$\xi^{(i)} \geqslant 0$$
$$\xi^{(i)} \geqslant 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$$

Problem Formulation

- We update problem (1) to include the hinge loss error:

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + b))$$

Or equivalently:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi^{(i)} && (3) \\ \text{subject to:} \quad & y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + b) \geqslant 1 - \xi^{(i)} \\ & \xi^{(i)} \geqslant 0 && \forall i \end{aligned}$$

- Where we have expressed the hinge loss via the two constraints

Tuning Parameter C

- C modulates the sum of $\xi^{(i)}$
- It determines the number and severity of the violations of the margin
- As C increases then we become less tolerant of errors and the margin will decrease

Lagrange Duality

- Once again we can make use of Lagrangian Duality:
- First we write the Lagrangian for problem (3):

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha^{(i)} \left(y^{(i)} (\mathbf{w} \cdot \mathbf{x}^{(i)} + b) - 1 + \xi^{(i)} \right) - \sum_{i=1}^{n} \beta^{(i)} \xi^{(i)} + C \sum_{i=1}^{n} \xi^{(i)}$$

where: $\alpha^{(i)}, \beta^{(i)} \geqslant 0$

- The dual objective can be written:

$$\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{w}, b, \boldsymbol{\xi}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Lagrange Duality

- This is an unconstrained optimisation which we can solve by seeking stationary points:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w}^* - \sum_{i=1}^{n} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} = 0 \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha^{(i)} y^{(i)} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi^{(i)}} = C - \alpha^{(i)} - \beta^{(i)} = 0$$

- Substituting these expressions back into $\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ yields:

$$\mathcal{D}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$$

Dual Problem

- This leads to the following dual problem:

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \\
\text{subject to:} \quad & \sum_{i=1}^{n} \alpha^{(i)} y^{(i)} = 0 \\
& 0 \leqslant \alpha^{(i)} \leqslant C
\end{aligned}
$$

- Again, we can solve this problem using the **SMO** algorithm

Some Observations

- The KKT **complementary slackness** conditions, which must hold at optimality for this problem, tell us:

$$\alpha^{(i)}\left(y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)}+b)-1+\xi^{(i)}\right)=0$$
$$\beta^{(i)}\xi^{(i)}=0$$

- From the first condition the support vectors (those points for which $\alpha^{(i)} > 0$) must satisfy: $y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)}+b)=1-\xi^{(i)}$

Some Observations

- Recall from the third stationary condition (equations (4)) that: $\alpha^{(i)} = C - \beta^{(i)}$

- So, for $\alpha^{(i)} = 0$:

$$\beta^{(i)} = C \qquad \Longrightarrow \qquad \xi^{(i)} = 0$$

- And, for $\alpha^{(i)} > 0$, either:

$$\beta^{(i)} > 0 \qquad \Longrightarrow \qquad 0 < \alpha^{(i)} < C \quad \text{and} \quad \xi^{(i)} = 0$$

- Or:

$$\beta^{(i)} = 0 \qquad \Longrightarrow \qquad \alpha^{(i)} = C \quad \text{and} \quad \xi^{(i)} > 0$$

Some Observations

- To sum up, each point lies in one of the following states:
  - **Beyond margin**:

$$\alpha^{(i)} = 0 \qquad \text{and} \qquad y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) > 1$$

  - **On margin**:

$$0 < \alpha^{(i)} < C \qquad \text{and} \qquad y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) = 1$$

  - **Within margin**:

$$\alpha^{(i)} = C \qquad \text{and} \qquad y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) < 1$$

Primal Optimality

- Using similar arguments to the Hard Margin case, we can write:

$$\mathbf{w}^* = \sum_{i \in \mathcal{SV}} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)}$$

And:

$$b = \frac{1}{|\widetilde{\mathcal{SV}}|} \sum_{i \in \widetilde{\mathcal{SV}}} \left(y^{(i)} - \sum_{j \in \widetilde{\mathcal{SV}}} \alpha^{(j)*} y^{(j)} \mathbf{x}^{(j)} \cdot \mathbf{x}^{(i)} \right)$$

Where $\mathcal{SV}$ is the set of support vectors, and $\widetilde{\mathcal{SV}}$ is the set of support vectors for which $0 < \alpha^{(i)} < C$

Recap

- **Representation**

$$\mathcal{F} = \left\{ f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \middle| \mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R} \right\}$$

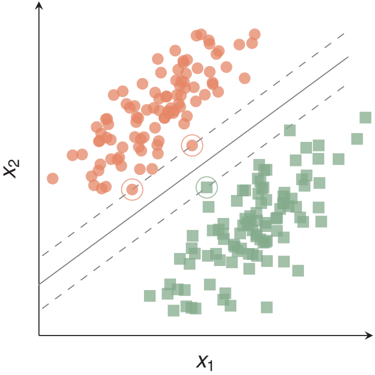
- **Evaluation**

$$\gamma \qquad \text{And:} \qquad \sum_{i=1}^{n} \max\left[0, 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)\right]$$

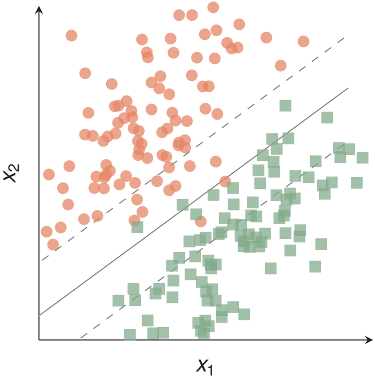
- **Optimisation**

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi^{(i)} \\ \text{subject to:} \quad & y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geqslant 1 - \xi^{(i)} \\ & \xi^{(i)} \geqslant 0 \qquad \forall i \end{aligned}$$

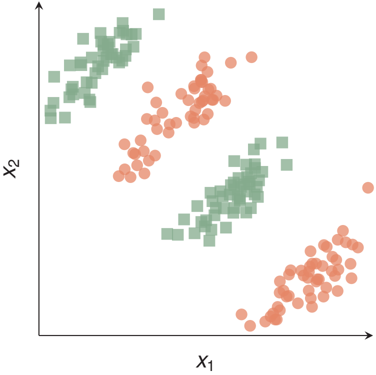
Linearly Separable with Hard Margin



Linearly Separable with Soft Margin



Non-Linearly Separable



Limits of the Linear SVM

- Here even a soft margin linear SVM would do badly
- It looks like we need something more flexible than a linear classifier
- This should remind us of the limitations of linear versus polynomial regression
- And again we'll need to enrich our function class to accommodate these cases

Limits of the Linear SVM

- Recall that we can do this by affecting a feature mapping of our input attributes, $\phi : \mathbf{x} \mapsto \phi(\mathbf{x})$
- We will see that we are able to handle very rich - even infinite dimensional - mappings of this type in a very efficient way...
- ...Because of the form of the dual problem which we developed earlier on

Motivation

- Note that thus far we have only motivated the SVM **intuitively**
- We **claimed** that maximising the margin was somehow linked to **generalisation**
- But how?

The PAC Approach

- One answer lies in the **PAC approach**
- Here we begin with the generalisation loss for misclassification: $\mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$
- Then we seek to express this as a **PAC bound**, in terms of:
  - The observable empirical training loss
    - Here we relax the misclassification (**Heaviside**) loss to the **hinge loss**
    - This is **conservative** and also assumes that the form of our bound is **convex**
  - Some complexity penalty, which takes into account the size of the **representation space**, $\mathcal{F}$
    - This term acts as a **regulariser** and penalises high weights

The PAC Approach

- We end up with a probabilistic ‘worst-case’ bound for the generalisation performance of our algorithm...
- ...And the problem of optimising this bound is identical to the SVM optimisation problem

Lecture Overview

Summary

1. The **SVM** is a classification algorithm which seeks **linear separating hyperplanes**, such that the **margin** of the system is maximised
   - **PAC Theory** shows us that the margin of a system and **generalisability** are related
2. The SVM can be formulated in a **hard margin** or **soft margin** version depending on whether our training data is linearly separable or not
3. When the decision boundary is non-linear we cannot use the linear SVM...unless we modify it...

Lecture Overview

Multiple Constraints: Problem

- $$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x})$$

  $$\text{subject to:} \quad \begin{cases} \{g^{(i)}(\mathbf{x}) \leqslant 0\}_{i=1}^{m} \\ \{h^{(j)}(\mathbf{x}) = 0\}_{j=1}^{p} \end{cases}$$

Multiple Constraints: Lagrangian

- We express the Lagrangian as:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu^{(i)} g^{(i)}(\mathbf{x}) + \sum_{j=1}^{p} \lambda^{(j)} h^{(j)}(\mathbf{x})$$

Where:
$\boldsymbol{\lambda} = [\lambda^{(1)}, ..., \lambda^{(p)}]^T, \{\lambda^{(j)} \in \mathbb{R}\}_{j=1}^{p};$
$\boldsymbol{\mu} = [\mu^{(1)}, ..., \mu^{(m)}]^T, \{\mu^{(i)} \in \mathbb{R}^{\geqslant 0}\}_{i=1}^{m};$
are Lagrange multipliers

Multiple Constraints: Problem Reformulation

- And we can solve our problem by seeking stationary solutions $(\mathbf{x}^*, \{\mu^{(i)*}\}, \{\lambda^{(j)*}\})$ which satisfy the following:

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0}$$

$$\text{subject to:} \quad \begin{cases} \{g^{(i)}(\mathbf{x}) \leqslant 0\}_{i=1}^{m}, \{h^{(j)}(\mathbf{x}) = 0\}_{j=1}^{p} \\ \{\mu^{(i)} \geqslant 0\}_{i=1}^{m} \\ \{\mu^{(i)} g^{(i)}(\mathbf{x}) = 0\}_{i=1}^{m} \end{cases}$$

Duality: Primal Problem

- The original problem is sometimes know as the **primal problem**, and its variables, $\mathbf{x}$, are known as the **primal variables**
- It is equivalent to the following formulation:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

- Here the bracketed term is known as the **primal objective** function

Duality: Barrier Function

- We can re-write the primal objective as follows:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geqslant 0} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\mathbf{x}) + \max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geqslant 0}\left[\sum_{i=1}^{m} \mu^{(i)} g^{(i)}(\mathbf{x}) + \sum_{j=1}^{p} \lambda^{(j)} h^{(j)}(\mathbf{x})\right]$$

- Here the second term gives rise to a **barrier function** which enforces the constraints as follows:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geqslant 0}\left[\sum_{i=1}^{m} \mu^{(i)} g^{(i)}(\mathbf{x}) + \sum_{j=1}^{p} \lambda^{(j)} h^{(j)}(\mathbf{x})\right] = \begin{cases} 0 & \text{if } \mathbf{x} \text{ is feasible} \\ \infty & \text{if } \mathbf{x} \text{ is infeasible} \end{cases}$$

Duality: Minimax Inequality

- In order to make use of this barrier function formulation, we will need the **minimax inequality**:

$$\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leqslant \min_{\mathbf{x}} \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$$

- **Proof:**

$$\min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leqslant \phi(\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y}$$

This is true for all **y**, therefore, in particular the following is true:

$$\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leqslant \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x}$$

This is true for all **x**, therefore, in particular the following is true:

$$\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leqslant \min_{\mathbf{x}} \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$$

Duality: Weak Duality

- We can now introduce the concept of **weak duality**:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right] \geqslant \max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} \left[\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

- Here the bracketed term on the right hand side is known as the **dual objective** function, $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu})$

- If we can solve the right hand side of the inequality then we have a lower bound on the solution of our optimisation problem

Duality: Weak Duality

- And often the RHS side of the inequality is an **easier** problem to solve, because:
  - $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is an **unconstrained** optimisation problem for a given value of $(\boldsymbol{\lambda}, \boldsymbol{\mu})$...
  - ...And if solving this problem is not hard then the overall problem is not hard to solve because:
  - $\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} [\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})]$ is a maximisation problem over a set of affine functions - thus it is a **concave maximisation** problem or equivalently a **convex minimisation** problem, and we know that such problems can be efficiently solved
  - Note that this is true regardless of whether $f, g^{(i)}, h^{(j)}$ are nonconvex

Duality: Strong Duality

- For certain classes of problems which satisfy **constraint qualifications** we can go further and **strong duality** holds:

$$\min_{\mathbf{x}} \left[\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right] = \max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} \left[\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

- There are several different constraint qualifications. One is **Slater's Condition** which holds for **convex optimisation** problems
- Recall, these are problems for which f is convex and $g^{(i)}, h^{(j)}$ are convex sets
- For problems of this type we may seek to solve the **dual optimisation** problem:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geqslant 0} \left[\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \right]$$

Duality: Strong Duality

- Another reason for adopting the dual optimisation approach to solving contrained optimisation problems is based on dimensionality:

- If the dimensionality of the dual variables, $(m + p)$, is less than the dimensionality of the primal variables, n, then dual optimisation often offers a more efficient route to solutions

- This is of particular importance if we are dealing with infinite dimensional primal variables