Maximum Margin Classifier

- Maximizing the margin:
  \[ \arg\min_{w,b} \frac{1}{2} \|w\|^2 \text{ subject to } N \text{ constraints } t_n(w^T x_n + b) \geq 1 \]

- We decided to "calibrate" \( w \) s.t. for the nearest point \( t_n(w^T x_n + b) = 1 \)

- Then the size of the margin is given by \( \frac{1}{\|w\|} \)

- And for all data points we have \( t_n(w^T x_n + b) \geq 1 \)
Maximum Margin Classifier

- Maximizing the margin:
  \[ \arg \min_{w,b} \frac{1}{2} \|w\|^2 \text{ subject to } N \text{ constraints } t_n(w^T x_n + b) \geq 1 \]

- Primal Lagrangian function:
  \[ L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{n=1}^{N} a_n \{ t_n(w^T x_n + b) - 1 \} \]

- With KKT conditions:
  (primal feasibility) \[ t_n(w^T x_n + b) - 1 \geq 0 \] for \( n = 1, \ldots, N \)
  (dual feasibility) \[ a_n \geq 0 \] for \( n = 1, \ldots, N \)
  (complimentary slackness) \[ a_n(t_n(w^T x_n + b) - 1) = 0 \] for \( n = 1, \ldots, N \)

- Dual Lagrangian obtained via (stationarity conditions)
  \[ \frac{\partial L}{\partial w} = 0, \quad \frac{\partial L}{\partial b} = 0 \]
  \[ \tilde{L}(a) = \min_{x,b} L(x, b, a) \]

- Solution: \[ a^* = \arg \max_a \tilde{L}(a) \quad \Rightarrow \quad w^*, b^* = \arg \min_{w,b} L(w, b, a^*) \]
Maximum Margin Classifier

- Primal Lagrangian function:
  
  \[
  L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{n=1}^{N} a_n \{ t_n (w^T x_n + b) - 1 \}
  \]

  with Langrange multipliers: \( a_n \geq 0 \) for \( n = 1, \ldots, N \)

- First step towards dual Langrangian: obtain stationarity conditions

  \[
  \frac{\partial L}{\partial w} = w^T - \sum_{n=1}^{N} a_n t_n x_n^T = 0 \quad \rightarrow \quad w = \sum_{n=1}^{N} a_n t_n x_n
  \]

  \[
  \frac{\partial L}{\partial b} = - \sum_{n=1}^{N} a_n t_n = 0 \quad \rightarrow \quad \sum_{n=1}^{N} a_n t_n = 0
  \]

- Eliminate \( w \) and \( b \) from \( L \) then gives the dual representation!
Maximum Margin Classifier

Stationarity conditions:

\[
\frac{\partial L}{\partial w} = w^T - \sum_{n=1}^{N} a_n t_n x_n^T = 0 \quad \rightarrow \quad w = \sum_{n=1}^{N} a_n t_n x_n
\]

\[
\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \quad \rightarrow \quad \sum_{n=1}^{N} a_n t_n = 0
\]

Eliminate \( w \) and \( b \) from \( L \) then gives the dual representation!

- Primal: \( L(w, b, a) = \frac{1}{2} w^T w - \sum_{n=1}^{N} a_n \{ t_n(w^T x_n + b) - 1 \} \), with \( a_n \geq 0 \) for \( n = 1, \ldots, N \)

- Dual: \( \bar{L}(a) = w^T (\frac{1}{2} w - \sum_{n=1}^{N} a_n t_n x_n^T) - \sum_{n=1}^{N} a_n b_n + \sum_{n=1}^{N} a_n \)

\[
= -\frac{1}{2} w^T w - b \sum_{n=1}^{N} a_n b_n + \sum_{n=1}^{N} a_n
\]

\[
= \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m x_n^T x_m
\]

with \( a_n \geq 0 \) for \( n = 1, \ldots, N \)

and with \( \sum_{n=1}^{N} a_n t_n = 0 \)
The dual representation of the maximum margin, where we maximize w.r.t. $\mathbf{a}$:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

with constraints:

- $a_n \geq 0$ for $n = 1, \ldots, N$
- $\sum_{n=1}^{N} a_n t_n = 0$

Apply the **KERNEL TRICK**: replace $\mathbf{x}_n^T \mathbf{x}_m$ with $k(\mathbf{x}_n, \mathbf{x}_m)$

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

Advantage: can now learn complex nonlinear decision boundaries!
Maximum Margin Classifier

- Prediction of class for datapoint $x_n$:
  \[ y(x_n) = w^T x_n + b \]

- Use $w = \sum_{n=1}^{N} a_n t_n x_n$ so that
  \[ y(x) = \sum_{n=1}^{N} a_n t_n x_n^T x + b \]

- Remember the KKT conditions:
  (primal feasibility) \[ t_n(w^T x_n + b) - 1 \geq 0 \] for $n = 1, \ldots, N$
  (dual feasibility) \[ a_n \geq 0 \] for $n = 1, \ldots, N$
  (complimentary slackness) \[ a_n(t_n(w^T x_n + b) - 1) = 0 \] for $n = 1, \ldots, N$

- Support vectors lie on maximum margin hyperplanes
  \[ a_n > 0 \quad \rightarrow \quad t_n y(x_n) = 1 \] (support vectors)
  \[ a_n = 0 \quad \leftrightarrow \quad t_n y(x_n) > 1 \] (all other points)
Maximum Margin Classifier

- Prediction of class for datapoint $\mathbf{x}$:
  
  $$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T \mathbf{x} + b \quad \rightarrow \quad y(\mathbf{x}) = \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}) + b$$

- Find $b$ by using that $t_n y_n(\mathbf{x}) = 1$ if $\mathbf{x}_n$ lies on the margin boundary! ($\mathbf{x}_n$ is a support vector)

  $$t_n \left( \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n) + b \right) = 1$$
  
  $$\sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n) + b = t_n$$

  $$b = t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n)$$

- More stable to average over all support vectors (depending on optimizer, $a_n$ may not be perfect)

  $$b = \frac{1}{N_S} \sum_{n \in S} \left( t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_m, \mathbf{x}_n) \right)$$
Maximum Margin Classifier

- Maximum Margin Classifier with Gaussian Kernel

\[ y(x) = \sum_{m \in S} a_m t_m k(x_m, x) + b, \quad \text{with} \quad k(x_n, x_m) = \exp \left( -\frac{1}{2\sigma^2} \|x_n - x_m\|^2 \right) \]

- Dataset is not linearly separable

- Nonlinear kernel can still separate the data perfectly!