Linear classifier with minimal classification error

- \( \mathcal{X} \) is a set of observations and \( \mathcal{Y} = \{+1, -1\} \) a set of hidden labels
- \( \phi: \mathcal{X} \to \mathbb{R}^n \) is fixed feature map embedding \( \mathcal{X} \) to \( \mathbb{R}^n \)
- **Task:** find linear classification strategy \( h: \mathcal{X} \to \mathcal{Y} \)

\[
h(x; \mathbf{w}, b) = \text{sign}(\langle \mathbf{w}, \phi(x) \rangle + b) = \begin{cases} +1 & \text{if } \langle \mathbf{w}, \phi(x) \rangle + b \geq 0 \\ -1 & \text{if } \langle \mathbf{w}, \phi(x) \rangle + b < 0 \end{cases}
\]

with minimal expected risk

\[
R_{0/1}^{0/1}(h) = \mathbb{E}_{(x,y) \sim p}(\ell_{0/1}^{0/1}(y, h(x))) \quad \text{where} \quad \ell_{0/1}^{0/1}(y, y') = [y \neq y']
\]

- We are given a set of training examples

\[
\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}
\]

drawn from i.i.d. with the distribution \( p(x, y) \).
The Empirical Risk Minimization principle leads to solving

\[(w^*, b^*) \in \operatorname{Argmin}_{(w, b) \in (\mathbb{R}^n \times \mathbb{R})} R_{T_m}^{0/1}(h(\cdot; w, b))\]  

where the empirical risk is

\[R_{T_m}^{0/1}(h(\cdot; w, b)) = \frac{1}{m} \sum_{i=1}^{m} \left[y^i \neq h(x^i; w, b)\right]\]

In this lecture we address the following issues:

1. The statistical consistency of the ERM for hypothesis space containing linear classifiers.

2. Algorithmic issues: in general, there is no known algorithm solving the task \((1)\) in time polynomial in \(m\).
Training linear classifier from separable examples

Definition 1. The examples $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ are linearly separable w.r.t. feature map $\phi: \mathcal{X} \to \mathbb{R}^n$ if there exists $(w, b) \in \mathbb{R}^{n+1}$ such that

$$y^i(\langle w, \phi(x^i) \rangle + b) > 0, \quad i \in \{1, \ldots, m\}$$

(2)

Perceptron algorithm:

**Input:** linearly separable examples $\mathcal{T}^m$

**Output:** linear classifier with $R^0_{\mathcal{T}^m}(h(\cdot; w, b)) = 0$

step 1: $w \leftarrow 0$, $b \leftarrow 0$

step 2: find $(x^i, y^i)$ such that $y^i(\langle w, \phi(x^i) \rangle + b) \leq 0$.

If not found exit, the current $(w, b)$ solves the problem.

step 3: $w \leftarrow w + y^i \phi(x^i)$, $b \leftarrow b + y^i$ and goto to step 2.
Training linear classifier from NON-separable examples

- The intractable ERM problem we wish to solve

\[
(w^*, b^*) \in \text{Argmin}_{(w, b) \in (\mathbb{R}^n \times \mathbb{R})} \frac{1}{m} \sum_{i=1}^{m} \left[ y^i \neq h(x^i; w, b) \right] \ell_0/1(y^i, h(x^i; w, b))
\]

where \( h(x; w, b) = \text{sign}(\langle w, \phi(x) \rangle + b) \).

- The ERM problem is approximated by a tractable convex problem

\[
(w^*, b^*) \in \text{Argmin}_{(w, b) \in (\mathbb{R}^n \times \mathbb{R})} \frac{1}{m} \sum_{i=1}^{m} \max \left\{ 0, 1 - y^i f(x^i; w, b) \right\} \psi(y^i, f(x^i; w, b))
\]

where \( f(x; w, b) = \langle w, \phi(x) \rangle + b \) and \( \psi(y, f(x)) \) is so called Hinge-loss.
The hinge-loss upper bounds the 0/1-loss

- The hinge-loss is an upper bound of the 0/1-loss evaluated for the predictor $h(x) = \text{sign}(f(x))$:

$$\ell_{0/1}(y,f(x)) = \left[ y f(x) \leq 0 \right] \leq \max\{0, 1 - y f(x)\}$$

\[
\begin{align*}
\left[ \text{sign}(f(x)) \neq y \right] &= \left[ y f(x) \leq 0 \right] \\
\psi(y, f(x)) &= \max\{0, 1 - y f(x)\}
\end{align*}
\]

\[
\begin{align*}
\max(0, 1 - t) &
\end{align*}
\]
Support Vector Machines

- Find linear classifier $h(x; \mathbf{w}, b) = \text{sign}(\langle \phi(x), \mathbf{w} \rangle + b)$ by solving

$$
(w^*, b^*) = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}} \left( \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \max\{0, 1 - y^i(\langle \mathbf{w}, \phi(x^i) \rangle + b)\} \right)
$$

- The regularization constant $C \geq 0$ helps to prevent overfitting (i.e. high estimation error) by constraining the parameter space.

  - $C_1 < C_2$ implies $\|w_1^*\| \leq \|w_2^*\|

- Small $\|\mathbf{w}\|$ implies score $f(x; \mathbf{w}, b) = \langle \mathbf{w}, \phi(x) \rangle + b$ varies slowly.

  - Cauchy inequality:
    $$
    (\langle \phi(x), \mathbf{w} \rangle - \langle \phi(x'), \mathbf{w} \rangle)^2 \leq \|\phi(x) - \phi(x')\|^2 \|\mathbf{w}\|^2
    $$
\[(w^*, b^*) = \arg\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \left( \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \max\{0, 1 - y^i(\langle w, \phi(x^i)\rangle + b)\} \right) \]

Example: Primal SVM problem

\[\frac{1}{|w|} = 0.34\]

\[C=1000.0, |w|=2.96, \text{trnErr}=0.10, \text{hingeLoss}=0.22\]
SVM as Quadratic Program

- Find linear classifier $h(x; \mathbf{w}, b) = \text{sign}(\langle \phi(x), \mathbf{w} \rangle + b)$ by solving

$$ (\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}} \left( \frac{1}{2} \| \mathbf{w} \|_2^2 + C \sum_{i=1}^m \max\{0, 1 - y^i(\langle \mathbf{w}, \phi(x^i) \rangle + b)\} \right) $$

where $C > 0$ is the regularization constant.

- It can be re-formulated as a convex quadratic program

$$ (\mathbf{w}^*, b^*, \xi^*) = \arg\min_{(\mathbf{w}, b) \in \mathbb{R}^{n+1}} \left( \frac{1}{2} \| \mathbf{w} \|_2^2 + C \sum_{i=1}^m \xi_i \right) $$

subject to

$$ y^i(\langle \mathbf{w}, \phi(x^i) \rangle + b) \geq 1 - \xi_i, \quad i \in \{1, \ldots, m\} $$

$$ \xi_i \geq 0, \quad i \in \{1, \ldots, m\} $$
From Primal SVM to Dual SVM problem

- Lagrangian of the primal SVM problem:

\[
L(w, b, \xi, \alpha, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i
\]

original objective

\[
- \sum_{i=1}^{m} \alpha_i (y^i (\langle w, \phi(x^i) \rangle + b) - 1 + \xi_i) - \sum_{i=1}^{m} \mu_i \xi_i
\]

constraint violation penalty

- Strong duality:

\[
\begin{align*}
\min_{w \in \mathbb{R}^n} \max_{\alpha \in \mathbb{R}^m_+} \min_{b \in \mathbb{R}} \min_{\xi \in \mathbb{R}^m} \max_{\mu \in \mathbb{R}^m_+} L(w, b, \xi, \alpha, \mu) &= \max_{\alpha \in \mathbb{R}^m_+} \min_{w \in \mathbb{R}^n} \min_{b \in \mathbb{R}} \min_{\xi \in \mathbb{R}^m} \max_{\mu \in \mathbb{R}^m_+} L(w, b, \xi, \alpha, \mu)
\end{align*}
\]

primal problem

dual problem
Dual SVM problem

- The dual SVM formulation is a convex quadratic program

\[
\alpha^* = \arg\max_{\alpha \in \mathbb{R}^m} \left( \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \phi(x^i), \phi(x^j) \rangle \right)
\]

s.t. \[ \sum_{i=1}^{m} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i \in \{1, \ldots, m\} \]

- The primal variables \((w, b)\) are obtained from the dual variables \(\alpha\) by

\[
w = \sum_{i=1}^{m} y^i \phi(x^i) \alpha_i = \sum_{i \in \mathcal{I}_{SV}} y^i \phi(x^i) \alpha_i
\]

\[
b = y^i - \langle w, \phi(x^i) \rangle, \quad \forall i \in \mathcal{I}_{SV}^b = \{ j \mid 0 < \alpha_j < C \}
\]

- \(\alpha\) is sparse; \(w\) is lin. combination of Support Vectors \(\mathcal{I}_{SV} = \{ j \mid \alpha_j > 0 \}\)
Example: SVM classifier

\[ f(x) = \langle w, \phi(x) \rangle + b = \langle \sum_{i=1}^{m} y^i \alpha^i \phi(x^i), \phi(x) \rangle + b \]

\[ y^i = +1 \]
\[ \alpha^i = C \quad \xi^i > 0 \]
\[ \alpha^i = 0 \quad \xi^i = 0 \]

\[ f(x) = +1 \]
\[ f(x) = 0 \]
\[ f(x) = -1 \]
\[ y^i = -1 \]
\[ \alpha^i \in (0, C) \quad \xi^i = 0 \]