Linear classifier with minimal classification error

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

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

with minimal expected risk

$$R^{0/1}(h) = \mathbb{E}_{(x,y) \sim p}\left(\ell^{0/1}(y, h(x))\right) \quad \text{where} \quad \ell^{0/1}(y, y') = [y \neq y']$$

- 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 $p(x, y)$. 
ERM learning for linear classifiers

- The Empirical Risk Minimization principle leads to solving

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

where the empirical risk is

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

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 \(m\).
Vapnik-Chervonenkis (VC) dimension

**Definition 1.** Let $\mathcal{H} \subseteq \{-1, +1\}^\mathcal{X}$ and $\{x^1, \ldots, x^m\} \in \mathcal{X}^m$ be a set of $m$ input observations. The set $\{x^1, \ldots, x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $y \in \{+1, -1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i) = y^i$, $i \in \{1, \ldots, m\}$.

**Definition 2.** Let $\mathcal{H} \subseteq \{-1, +1\}^\mathcal{X}$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest set of points from $\mathcal{X}$ which can be shattered by $\mathcal{H}$.
Vapnik-Chervonenkis (VC) dimension

Definition 1. Let $\mathcal{H} \subseteq \{-1, +1\}^\mathcal{X}$ and $\{x^1, \ldots, x^m\} \in \mathcal{X}^m$ be a set of $m$ input observations. The set $\{x^1, \ldots, x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $y \in \{+1, -1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i) = y^i$, $i \in \{1, \ldots, m\}$.

Definition 2. Let $\mathcal{H} \subseteq \{-1, +1\}^\mathcal{X}$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest set of points from $\mathcal{X}$ which can be shattered by $\mathcal{H}$.

Theorem 1. The VC-dimension of the hypothesis space of all linear classifiers operating in $n$-dimensional feature space
\[ \mathcal{H} = \{h(x; w, b) = \text{sign}(\langle w, \phi(x) \rangle + b) \mid (w, b) \in (\mathbb{R}^n \times \mathbb{R})\} \] is $n + 1$. 
Consistency of prediction with two classes and 0/1-loss

**Theorem 2.** Let \( \mathcal{H} \subseteq \{+1, -1\}^X \) be a hypothesis space with VC dimension \( d < \infty \) and \( \mathcal{T}_m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (X \times Y)^m \) a training set draw from i.i.d. rand vars with distribution \( p(x, y) \). Then, for any \( \varepsilon > 0 \) it holds

\[
\mathbb{P} \left( \sup_{h \in \mathcal{H}} \left| R^{0/1}_{0}(h) - R^{0/1}_{T_m}(h) \right| \geq \varepsilon \right) \leq 4 \left( \frac{2em}{d} \right)^d e^{-\frac{m \varepsilon^2}{8}}.
\]
Consistency of prediction with two classes and 0/1-loss

**Theorem 2.** Let $\mathcal{H} \subseteq \{+1, -1\}^X$ be a hypothesis space with VC dimension $d < \infty$ and $T_m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (X \times Y)^m$ a training set drawn from i.i.d. rand vars with distribution $p(x, y)$. Then, for any $\varepsilon > 0$ it holds

$$
\mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R_{0/1}(h) - R_{T_m}^{0/1}(h) \right| \geq \varepsilon \right) \leq 4 \left( \frac{2 e m}{d} \right)^d e^{- \frac{m \varepsilon^2}{8}}.
$$

**Corollary 1.** Let $\mathcal{H} \subseteq \{+1, -1\}^X$ be a hypothesis space with VC dimension $d < \infty$. Then ERM is statistically consistent in $\mathcal{H}$ w.r.t $\ell^{0/1}$ loss function.

**Corollary 2.** Let $\mathcal{H} \subseteq \{+1, -1\}^X$ be a hypothesis space with VC dimension $d < \infty$. Then, for any $0 < \delta < 1$ the inequality

$$
R_{0/1}(h) \leq R_{T_m}^{0/1}(h) + \sqrt{\frac{8 (d \log(2m) + 1) + \log \frac{4}{\delta}}{m}}
$$

holds for any $h \in \mathcal{H}$ with probability $1 - \delta$ at least.
Training linear classifier from separable examples

Definition 3. 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 $(\mathbf{w}, b) \in \mathbb{R}^{n+1}$ such that

$$y^i(\langle \mathbf{w}, \phi(x^i) \rangle + b) > 0, \quad i \in \{1, \ldots, m\} \quad (2)$$

Perceptron algorithm:

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

Output: linear classifier with $R_{\mathcal{T}^m}^{0/1}(h(\cdot; \mathbf{w}, b)) = 0$

step 1: $\mathbf{w} \leftarrow 0, \ b \leftarrow 0$

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

If not found exit, the current $(\mathbf{w}, b)$ solves the problem.

step 3: $\mathbf{w} \leftarrow \mathbf{w} + y^i \phi(x^i), \ b \leftarrow b + y^i$ 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 \( h(x; w, b) = \text{sign}(\langle w, \phi(x) \rangle + b) \).
Training linear classifier from NON-separable examples

■ The intractable ERM problem we wish to solve

\[
(w^*, b^*) \in \operatorname{Argmin}_{(w, b) \in (\mathbb{R}^n \times \mathbb{R})} \frac{1}{m} \sum_{i=1}^{m} [y^i \neq h(x^i; w, b)]_{\ell^0/1(y^i, h(x^i; w, b))}
\]

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

■ The ERM problem is approximated by a tractable convex problem

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

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 $h(x) = \text{sign}(f(x))$:

\[
\ell_{0/1}(y, f(x)) = \left\{ \begin{array}{ll}
[\text{sign}(f(x)) \neq y] &= [y f(x) \leq 0] \\
\psi(y, f(x)) &= \max\{0, 1 - y f(x)\}
\end{array} \right.
\]

\[
\ell_{0/1}(y, f(x)) = \left\{ \begin{array}{ll}
[t \leq 0] &= \max(0, 1 - t)
\end{array} \right.
\]

\[
\psi(y, f(x))
\]
Support Vector Machines

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

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

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

  - \( \lambda_1 > \lambda_2 \) implies \( \|w_1^*\| \leq \|w_2^*\| \)

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

  - Cauchy inequality:
    \[
    (\langle \phi(x), w \rangle - \langle \phi(x'), w \rangle)^2 \leq \|\phi(x) - \phi(x')\|^2 \|w\|^2
    \]
Summary

Topics covered in the lecture

- Linear classifier
- Vapnik-Chervonenkis dimension
- Consistency + generalization bound for two-class prediction and 0/1-loss
- ERM problem for linear classifiers
- Perceptron for separable examples
- SVM for non-separable examples
\[ \max(0, 1 - t) \]

\[ [t \leq 0] \]