Statistical Machine Learning (BE4M33SSU)
Lecture 4: Support Vector Machines II

Czech Technical University in Prague

- Lagrange duality
- Dual formulation of Support Vector Machines
- Positive definite kernel
- Examples of kernels

Constrained optimization problem and its Lagrange function

- For convex $f\colon \mathbb{R}^n \to \mathbb{R}$, $g_i \in \mathbb{R}^n \to \mathbb{R}$, $i \in \{1, \ldots, m\}$, we want to solve:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\operatorname{argmin}} f(\boldsymbol{\theta}) \quad \text{s.t.} \quad g_i(\boldsymbol{\theta}) \leq 0\,, i \in \{1, \ldots, m\}$$

- Let us define a Lagrange function $L\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ such that

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = f(\boldsymbol{\theta}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{\theta})$$

- The original problem is equivalent to and unconstrained problem

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\operatorname{argmin}} P(\boldsymbol{\theta})$$

where

$$P(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \begin{cases} f(\boldsymbol{\theta}) & \text{if} \quad \sum_{i=1}^{m} g_i(\boldsymbol{\theta}) \leq 0\,, i \in \{1, \ldots, m\} \\ \infty & \text{otherwise} \end{cases}$$

Lagrange dual problem

- The dual problem reads

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \geq \mathbf{0}}{\operatorname{argmax}}\, D(\boldsymbol{\alpha}) \quad \text{where} \quad D(\boldsymbol{\alpha}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^n} L(\boldsymbol{\theta}, \boldsymbol{\alpha})$$

- **Strong duality**: if the problem is convex and there exists $\boldsymbol{\theta} \in \mathbb{R}^n$ such that $g_i(\boldsymbol{\theta}) < 0, i \in \{1, \ldots, m\}$, then the duality gap is zero:

$$P(\boldsymbol{\theta}^*) = D(\boldsymbol{\alpha}^*)$$

- **Karush-Kunh-Tucker conditions**: $(\boldsymbol{\theta}^*, \boldsymbol{\alpha}^*)$ is the solution of the primal and the dual problem with zero duality gap if and only if:

$$\begin{aligned}
&1) \quad \nabla f(\boldsymbol{\theta}^*) + \sum_{i=1}^{m} \alpha_i^* \nabla g_i(\boldsymbol{\theta}^*) = \mathbf{0} \\
&2) \quad g_i(\boldsymbol{\theta}^*) \leq 0\,, \qquad i \in \{1, \ldots, m\} \\
&3) \quad \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{\theta}^*) = 0
\end{aligned}$$

Primal SVM problem

- The formulation of the linear SVM algorithm reads

$$(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*) = \underset{\substack{(\boldsymbol{w},b)\in\mathbb{R}^{n+1}\\ \boldsymbol{\xi}\in\mathbb{R}^m}}{\operatorname{argmin}} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i\right)$$

$$\begin{aligned} \text{s.t.} \quad y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) &\geq 1-\xi_i\,, \quad i\in\{1,\dots,m\} \\ \xi_i &\geq 0\,, \quad i\in\{1,\dots,m\} \end{aligned}$$

- The Lagrange function

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\big(y^i(\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i)\rangle + b) - 1 + \xi_i\big) - \sum_{i=1}^{m}\mu_i\xi_i$$

- The SVM learning expressed as an unconstrained problem

$$(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*) \in \underset{\substack{(\boldsymbol{w},b)\in\mathbb{R}^{n+1}\\ \boldsymbol{\xi}\in\mathbb{R}^m}}{\operatorname{argmin}} P(\boldsymbol{w}, b, \boldsymbol{\xi}) \quad \text{where} \quad P(\boldsymbol{w}, b, \boldsymbol{\xi}) = \max_{\substack{\boldsymbol{\alpha}\geq\mathbf{0}\\ \boldsymbol{\mu}\geq\mathbf{0}}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})$$

Dual SVM problem

- The dual SVM problem reads

$$(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*) = \operatorname*{argmax}_{\substack{\boldsymbol{\alpha}\geq\mathbf{0}\\ \boldsymbol{\mu}\geq\mathbf{0}}} D(\boldsymbol{\alpha}, \boldsymbol{\mu}) \quad \text{where} \quad D(\boldsymbol{\alpha}, \boldsymbol{\mu}) = \min_{\substack{\boldsymbol{w}\in\mathbb{R}^n\\ b\in\mathbb{R}\\ \boldsymbol{\xi}\in\mathbb{R}^m}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})$$

- Given $(\boldsymbol{\alpha}, \boldsymbol{\mu})$, the function $L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ w.r.t. $(\boldsymbol{w}, b, \boldsymbol{\xi})$ is convex and differentiable hence we find the optimum by solving:

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}{\partial \boldsymbol{w}} = \mathbf{0}, \quad \frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}{\partial b} = \mathbf{0}, \quad \frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}{\partial \boldsymbol{\xi}} = \mathbf{0}$$

by which we get

$$D(\boldsymbol{\alpha}, \boldsymbol{\mu}^*(\boldsymbol{\alpha})) = \begin{cases} \sum_{i=1}^m \alpha_i - \frac{1}{2\lambda} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^i y^j \langle \boldsymbol{\phi}(x^i), \boldsymbol{\phi}(x^j) \rangle & \text{if} \quad \sum_{i=1}^m \alpha_i y^i = 0 \\ & \boldsymbol{\alpha} \in [0, \frac{1}{m}]^m \\ -\infty & \text{otherwise} \end{cases}$$

Dual SVM problem

- The dual SVM formulation is a **convex quadratic program**

$$\boldsymbol{\alpha}^* = \operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left(\sum_{i=1}^m \alpha_i - \tfrac{1}{2\lambda} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^i y^j \langle \boldsymbol{\phi}(x^i), \boldsymbol{\phi}(x^j) \rangle \right)$$
$$\text{s.t.} \qquad \sum_{i=1}^m \alpha_i \, y^i = 0 \;, \quad 0 \leq \alpha_i \leq \tfrac{1}{m} \;\; i \in \{1, \ldots, m\}$$

- The primal solution $(\boldsymbol{w}^*, b^*)$ is obtained from $\boldsymbol{\alpha}^*$ using KKT conditions

$$\boldsymbol{w}^* = \sum_{i=1}^m y^i \boldsymbol{\phi}(x^i) \alpha_i^* \qquad \text{and} \qquad b^* = y^i - \langle \boldsymbol{w}^*, \boldsymbol{\phi}(x^i) \rangle \;, \; \forall i \in \mathcal{I}_{\text{sv}}^<$$

  where $\mathcal{I}_{\text{sv}}^< = \{i \in \{1, \ldots, m\} \mid 0 < \alpha_i < \frac{1}{m}\}$ are boundary SVs.

- To represent the classifier we need only the support vectors: training examples with indices $\mathcal{I}_{\text{sv}} = \{i \in \{1, \ldots, m\} \mid \alpha_i > 0\}$.

Kernel SVM

- We see that the values of **kernel function** $k(x, x') = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x') \rangle$ are **sufficient to learn parameters**

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\operatorname{argmin}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2\lambda} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^i y^j k(x^i, x^j) \right)$$
$$\text{s.t.} \qquad \sum_{i=1}^{m} \alpha_i \, y^i = 0 \,, \quad 0 \leq \alpha_i \leq \frac{1}{m} \;\; i \in \{1, \ldots, m\}$$

and **to evaluate the SVM classifier**

$$h(x; \boldsymbol{\alpha}^*, b^*) = \operatorname{sign}(\langle \boldsymbol{w}^*, \boldsymbol{\phi}(x) \rangle + b^*) = \operatorname{sign} \left(\sum_{i=1}^{m} y^i \alpha_i^* k(x^i, x) + b^* \right)$$

- When is it useful to represent inputs by $k(x, x')$ instead of $\boldsymbol{\phi}(x)$?
- What choices of $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are reasonable ?

Linear SVM versus kernel SVM

	Linear SVM	Kernel SVM
	Input representation	
comput. time	$\boldsymbol{\psi}\colon \mathcal{X} \to \mathbb{R}^n$ $\mathcal{O}(n)$	$k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ typically $\mathcal{O}(d)$

	Classifier	
linear score	$\langle \boldsymbol{w}, \boldsymbol{\psi}(x) \rangle$	$\sum_{i=1}^{m} \alpha_i y_i k(x^i, x)$
memory/time	$\mathcal{O}(n)$	$\mathcal{O}(m_{\text{sv}} + m_{\text{sv}} \cdot d)$

	Learning (QP solver)	
memory/minimal time	$\mathcal{O}(n \cdot m)$	$\mathcal{O}(m^2)$

d is the size of a single input $x \in \mathcal{X}$

m is the number of training examples

$m_{\text{sv}} = |\{i \in \{1, \ldots, m\} \mid \alpha_i > 0\}|$ number of support vectors

Positive definite kernel

Definition 1. *Let $\mathcal{X}$ be a non-empty set. The function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if it is symmetric and for any finite set of inputs $x^1, \ldots, x^m$, the kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ with elements $K_{i,j} = k(x^i, x^j)$ is positive semi-definite.*

- The kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ represents similarities between each pair of inputs $\{x^1, \ldots, x^m\}$:

$$\mathbf{K} = \begin{pmatrix} k(x^1, x^1), & k(x^1, x^2), & \ldots, & k(x^1, x^m) \\ k(x^2, x^1), & k(x^2, x^2), & \ldots, & k(x^2, x^m) \\ \vdots & & & \\ k(x^m, x^1), & k(x^m, x^2), & \ldots, & k(x^m, x^m) \end{pmatrix}$$

- A matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ is PSD if for every $\boldsymbol{\alpha} \in \mathbb{R}^m$, $\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \geq 0$.

Hilbert space

Definition 2. *A Hilbert space $\mathcal{H}$ is a complete vector space with a dot product $\langle \cdot, \cdot \rangle \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that satisfies the following properties:*

- *Symmetry:* $\langle f, g \rangle = \langle g, f \rangle$, $\forall f, g \in \mathcal{H}$
- *Linearity:* $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$, $\forall f_1, f_2, g \in \mathcal{H}$, $\alpha_1, \alpha_2 \in \mathbb{R}$
- *Positive definiteness:* $\langle f, f \rangle \geq 0$ *with equality iff* $f = 0$

The dot product defines a norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$

Every feature map defines PSD kernel

Theorem 1. *Let $\phi\colon \mathcal{X} \to \mathcal{H}$ be a feature map representing inputs from $\mathcal{X}$ in a Hilbert space $\mathcal{H}$. Then the function $k(x, x') = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x') \rangle$ is a positive definite kernel.*

Proof. *Given $\{x^i \in \mathcal{X} \mid i = 1, \ldots, m\}$, the kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$, with elements $K_{i,j} = k(x^i, x^j)$ is PSD because*

$$
\begin{aligned}
\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} &= \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \langle \boldsymbol{\phi}(x^i), \boldsymbol{\phi}(x^j) \rangle \\
&= \left\langle \sum_{i=1}^{m} \alpha_i \boldsymbol{\phi}(x^i), \sum_{j=1}^{m} \alpha_j \boldsymbol{\phi}(x^j) \right\rangle \\
&= \left\| \sum_{i=1}^{m} \alpha_i \boldsymbol{\phi}(x^i) \right\|^2 \geq 0
\end{aligned}
$$

Every kernel defines a feature space

Theorem 2. *For every positive definitie kernel* $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *there exists a Hilbert spave* $\mathcal{H}$ *and a feature map* $\phi\colon \mathcal{X} \to \mathcal{H}$ *such that* $k(x, x') = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x') \rangle$.

Proof for a kernel defined on a finite input space $\mathcal{X}$:

The kernel matrix $\mathbf{K} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is symmetric and PSD hence the spectral decomposition exists $\mathbf{K} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ where $\mathbf{V} \in \mathbb{R}^{m \times m}$ is orthogonal and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{|\mathcal{X}|})$ is diagonal matrix of non-negative eigenvalues.

Therefore $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$ where $\mathbf{\Phi}^T = \mathbf{V}\mathbf{D}^{\frac{1}{2}}$.

String Subsequence kernel

- Input space $\mathcal{X} = \cup_{d=0}^{\infty} \Sigma^d$ contains all strings on a finite alphabet Σ
- The features measure the number of occurrences of subsequences of length q weighting them according to their length and decay factor $\lambda \in (0, 1]$.

$$\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^{|\Sigma|^q} \qquad \text{and} \qquad \phi_u(s) = \sum_{\boldsymbol{i}\colon\, u = s[\boldsymbol{i}]} \lambda^{l(\boldsymbol{i})}, \quad \forall u \in \Sigma^q$$

- Example for strings "cat", "car", "bat" and "bar" and $q = 2$:

	c-a	c-t	a-t	b-a	b-t	c-r	a-r	b-r
$\boldsymbol{\phi}(\text{"}cat\text{"})$	λ^2	λ^3	λ^2	0	0	0	0	0
$\boldsymbol{\phi}(\text{"}car\text{"})$	λ^2	0	0	0	0	λ^3	λ^2	0
$\boldsymbol{\phi}(\text{"}bat\text{"})$	0	0	λ^2	λ^2	λ^3	0	0	0
$\boldsymbol{\phi}(\text{"}bar\text{"})$	0	0	0	λ^2	0	0	λ^2	λ^3

$k(\text{"}cat\text{"}, \text{"}car\text{"}) = \lambda^4$, $k(\text{"}cat\text{"}, \text{"}bar\text{"}) = \lambda^4$, $k(\text{"}cat\text{"}, \text{"}bar\text{"}) = 0$, ...

- The kernel value $k(s, t)$ can be computed by dynamic programming in time $\mathcal{O}(q\,|s|\,|t|)$ instead of $\mathcal{O}(|\Sigma|^q)$ required by using explicit features.

Polynomial kernel of degree 2

- Consider quadratic function of d-dimensional inputs $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^d$

$$\begin{aligned} f(\boldsymbol{x}) &= w_0 + \sqrt{2}x_1 w_1 + \cdots + \sqrt{2}x_d w_d + x_1^2 w_{1,1} + \cdots + x_d^2 w_{d,d} \\ &\quad + \sqrt{2}x_1 x_2 w_{1,2} \cdots + \sqrt{2}x_1 x_d w_{1,d} + \cdots + \sqrt{2}x_{d-1} x_d w_{d-1,d} \\ &= \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle \end{aligned}$$

  where $\boldsymbol{w} \in \mathbb{R}^n$, $n = \frac{(d+1)d}{2} + d + 1$, and

$$\boldsymbol{\phi}(\boldsymbol{x}) = (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1 x_2, \ldots, \sqrt{2}x_{d-1} x_d)$$

- The dot product of two inputs mapped to the features can be computed via 2nd order polynomial kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^2$$

- The kernel evaluation requires d multiplications/additions instead of $n = \frac{(d+1)d}{2} + d + 1$ when computing the feature maps explicitly.

Examples of kernels

List of frequently used kernels on vectorial inputs $\mathcal{X} = \mathbb{R}^d$:

- Homogeneous polynomial of degree p: $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^p$

  The feature map composed of all monomials of degree p.

  For example, if $d = p = 2$, $\boldsymbol{\phi}(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$

- Inhomogeneous polynomial of degree p: $k(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^p$

  The feature map composed of all monomials up to degree p.

  For example, if $d = p = 2$, $\boldsymbol{\phi}(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$

- Gaussian kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$

  The feature map represents inputs in infinite dimensional space.

Summary

- Lagrange duality
- Dual formulation of Support Vector Machines
- Positive definite kernel
- Examples of kernels