Non-linear SVM & Introduction to Kernels

Nirupam Gupta

Department of Computer Science

UNIVERSITY OF COPENHAGEN

Recap of optimization theory & linear SVM

Mathematical optimization (Recap)

Target optimization problem:

$$
\begin{aligned}
\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad & f(w) \\
\text{Subject to} \quad & f_i(w) \leq 0 \quad i = 1, \ldots, p \\
& g_i(w) = 0 \quad i = 1, \ldots, q
\end{aligned}
\tag{Primal}
$$

The **Lagrangian** $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ of the above problem is defined to be

$$\mathcal{L}(w, \lambda, \nu) = f(w) + \sum_{i=1}^{p} \lambda_i f_i(w) + \sum_{i=1}^{q} \nu_i g_i(w).$$

$\lambda = (\lambda_1, \ldots, \lambda_p)$ and $\nu = (\nu_1, \ldots, \nu_q)$ are called **dual variables** or **Lagrange multipliers**.

Lagrange dual problem (Recap)

The **Lagrange dual function** $\phi : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is defined to be

$$\phi(\lambda, \nu) := \min_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \nu) = \min_{w \in \mathbb{R}^d} \left(f(w) + \sum_{i=1}^{p} \lambda_i f_i(w) + \sum_{i=1}^{q} \nu_i g_i(w) \right).$$

The **Lagrange dual problem** is defined to be

$\lambda_{t+1} = \Pi_{\lambda \geq 0} [\lambda_t - \gamma$

$$\begin{aligned} &\underset{\lambda \in \mathbb{R}^p,\ \nu \in \mathbb{R}^q}{\text{Maximize}} \quad \phi(\lambda, \nu) \\ &\text{Subject to} \quad \lambda \succeq 0 \end{aligned} \tag{Dual}$$

Note: (Dual) is a convex optimization problem even when (Primal) is not convex.

Strong duality and KKT conditions (Recap)

Let p^* and d^* be the optimal values of (Primal) and (Dual), resp. Then, $p^* \geq d^*$.

Strong duality: when $p^* = d^*$.

KKT conditions: Necessary conditions for optimality (under strong duality and differentiability).

$$\begin{aligned} f_i(w^*) &\leq 0\ , \quad && i = 1, \dots, p \\ g_i(w^*) &= 0\ , \quad && i = 1, \dots, q \\ \lambda^* &\succeq \mathbf{0} && \\ \lambda_i^* f_i(w^*) &= 0\ , \quad && i = 1, \dots, p \\ \nabla_w f(w^*) + \textstyle\sum_{i=1}^p \lambda_i \nabla_w f_i(w^*) + \sum_{i=1}^q \nu_i \nabla_w g_i(w^*) &= \mathbf{0} && \end{aligned} \tag{KKT}$$

KKT conditions are also **sufficient if the primal is convex and differentiable.** Strong duality is for free in that case.

Soft-margin linear SVM (Recap)

Dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, with $x_i \in \mathcal{X} \subset \mathbb{R}^m$ and $y_i \in \{-1, +1\}$.

Primal optimization problem:

$$\begin{aligned} \underset{w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Minimize}} \quad & \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n} \xi_i^r \\ \underset{i=1,\ldots,n}{\text{Subject to}} \quad & 1 - y_i(w^\intercal x_i + b) - \xi_i \leq 0 \\ & \xi_i \geq 0 \end{aligned}$$

$c \in \mathbb{R}_{++}$ is the **misclassification penalty** and $r \geq 1$.

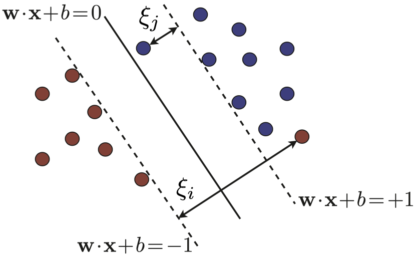
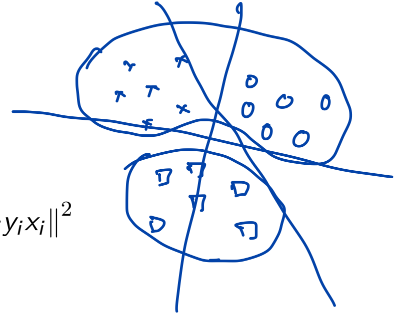


Figure: Linear classification with soft-margin.

Solution to linear SVM (Recap)



Dual optimization problem (for $r = 1$):

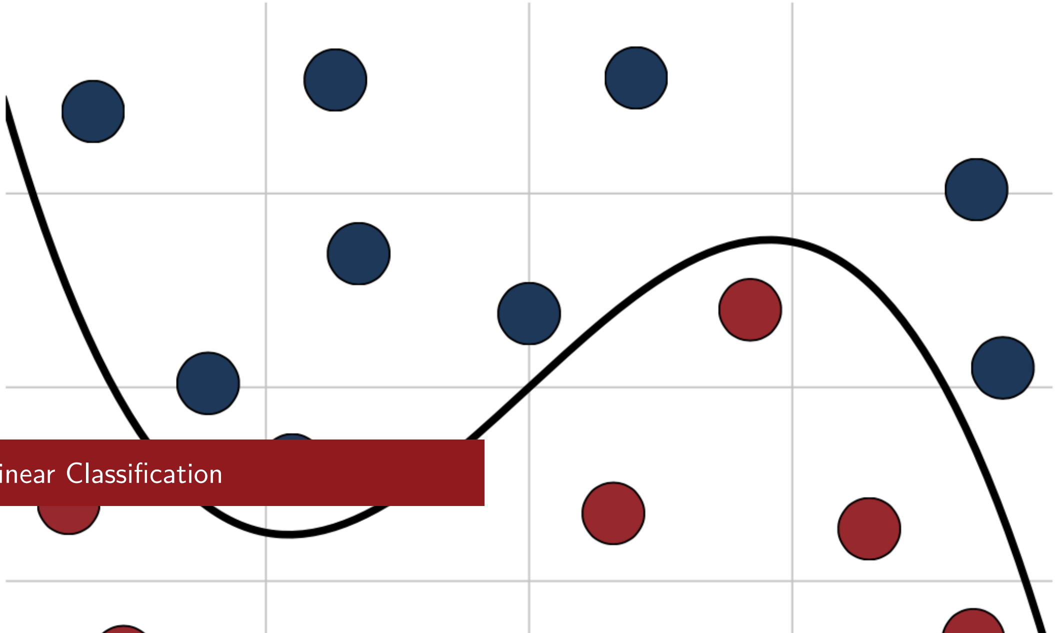
what is the advantage of this formulation over the primal ?

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^n}{\text{Maximize}} \quad & \phi(\lambda) := \sum_{i=1}^n \lambda_i - \frac{1}{2} \left\| \sum_{i=1}^n \lambda_i y_i x_i \right\|^2 \\ \underset{i=1,\dots,n}{\text{Subject to}} \quad & 0 \le \lambda_i \le c \\ & \sum_{i=1}^n \lambda_i y_i = 0 \end{aligned}$$

Support vectors: set of points (x_i, y_i) for which $\lambda_i^* > 0$.

Optimal weights. $w^* = \sum_{i \in SV} \lambda_i^* y_i x_i$, where $SV \subseteq [n]$ is indices of support vectors.

For $i \in SV$ with $\lambda_i^* < c$ (i.e., x_i lies on the *marginal hyperplane*) , $\langle w^*, x_i \rangle + b^* = y_i$.
Thus, $b^* = y_i - \sum_{j=1}^n \lambda_j^* y_j \langle x_j, x_i \rangle$.



Nonlinear Classification

Linear classifiers can be suboptimal

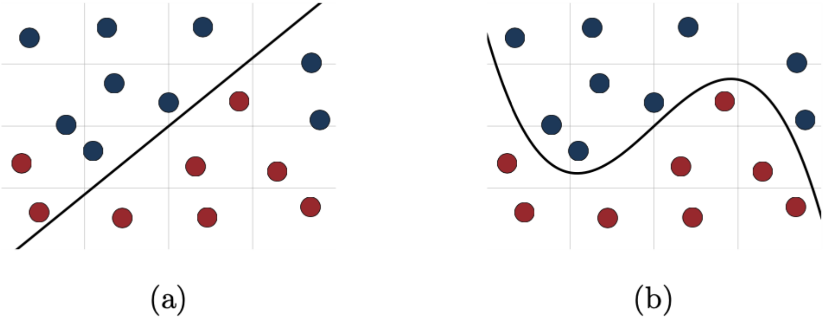
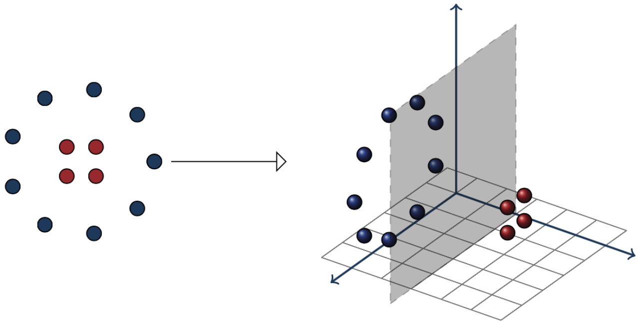


Figure: (a) Linear classifier. (b) Nonlinear classifier.

Nonlinear transformation for inducing linear separability

Nonlinear mapping to a higher dimensional space.
For example, in the case below with input space $\mathcal{X} \subset \mathbb{R}^2$, by mapping x to $\Psi(x) = ([x]_1, [x]_2, \|x\|)$ we obtain linear separability.



Linear SVM with input space transformation

Consider **feature mapping** $\Psi : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is referred to as the **feature space**.

Dual problem of **linear SVM over transformed data points:**

$$\begin{aligned} &\underset{\lambda \in \mathbb{R}^n}{\text{Maximize}} && \phi(\lambda) := \sum_{i=1}^n \lambda_i - \frac{1}{2} \left\| \sum_{i=1}^n \lambda_i y_i \Psi(x_i) \right\|^2 \\ &\underset{i=1,\ldots,n}{\text{Subject to}} && 0 \le \lambda_i \le c \quad \text{and} \quad \sum_{i=1}^n \lambda_i y_i = 0 \end{aligned}$$

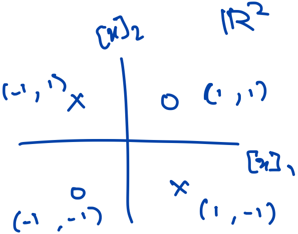
By analogy to original SVM problem, $w^* = \sum_{i=1}^n \lambda_i^* y_i \Psi(x_i) \in \mathcal{Z}$. For i such that $0 < \lambda_i^* < c$ we obtain $b^* = y_i - \sum_{j=1}^n \lambda_i^* y_j \langle \Psi(x_i), \Psi(x_j) \rangle$.

Hypothesis: $h(x) = \text{Sign}\left(\langle w^*, \Psi(x) \rangle + b^*\right)$.

Caveat: Computational cost for $\Psi(x)$ is in $\mathcal{O}(\dim(\mathcal{Z}))$ and can be prohibitively high in practice.

Linear SVM on XOR-type data

Consider **feature mapping** $\Psi(x) = ([x]_1, [x]_2, [x]_1[x]_2)$.



$(1, 1) \longmapsto (1, 1, 1)$

$(1, -1) \longmapsto (1, -1, -1)$

$(-1, -1) \longmapsto (-1, -1, 1)$

$(-1, 1) \longmapsto (-1, 1, -1)$

$[\Psi(x)]_3 > 0 \longrightarrow o$

$< 0 \longrightarrow x$

Kernels: Efficient incorporation of nonlinear transformation

We can write $\phi(\lambda) := \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \boxed{\langle \Psi(x_i), \Psi(x_j) \rangle}$. $\rightarrow K(x_i, x_j)$

$= K(x_j, x_i)$

Moreover, the hypothesis $h(x) = \text{Sign}\left(\sum_{i=1}^{n} \lambda_i^* y_i \langle \Psi(x_i), \Psi(x) \rangle + b^*\right)$

These computations involving inner-products can be performed without explicitly computing Ψ.

Kernels: A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. $\rightarrow$ symmetric.

K is **positive definite symmetric** (PDS) if for any $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the *Gram matrix* $\mathbf{K} = [K(x_i, x_j)]_{ij}$ is (symmetric) positive semi-definite.

Theorem. If K is **PDS** then K defines an inner product in a Hilbert space $\mathcal{Z}$, and

$$\text{there exists } \Psi : \mathcal{X} \to \mathcal{Z} \text{ such that } K(x, x') = \langle \Psi(x), \Psi(x') \rangle.$$

Linear SVM with kernel K

Replacing $\langle \Psi(x_i), \Psi(x_j) \rangle$ by $K(x_i, x_j)$ in the dual SVM problem we obtain:

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^n}{\text{Maximize}} \quad & \phi(\lambda) := \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,j=1}^n \lambda_i \lambda_j y_i y_j \boxed{K(x_i, x_j)} \\ \underset{i=1,\dots,n}{\text{Subject to}} \quad & 0 \le \lambda_i \le c \\ & \sum_{i=1}^n \lambda_i y_i = 0 \end{aligned}$$

$\langle \Psi(x_i), \Psi(x_j) \rangle$

The resulting hypothesis is given by (why?)

$$h(x) = \text{Sign}\left(\sum_{i=1}^n \lambda_i^* y_i K(x_i, x) + b^* \right),$$

where $b^* = y_i - \sum_{j=1}^n \lambda_i^* y_j K(x_i, x_j)$ with $i \in [n]$ such that $0 < \lambda_i^* < c$.

Linear SVM with kernel K (derivation)

The resulting hypothesis is given by:

$$h(x) = \mathrm{Sign}\left(\sum_{i=1}^{n} \lambda_i^* y_i K(x_i, x) + b^*\right),$$

Examples of PDS kernels

$$\psi(x_1) = \begin{bmatrix} x_1^T x_1 \\ \vdots \\ x_N^T x_1 \end{bmatrix} \qquad \binom{m+k}{k}$$

$\mathbb{R}_+$

- **Polynomial.** For $a \in \cancel{\mathbb{R}}$, polynomial kernel of degree $k \geq 1$ is $K(x, x') = \underline{\underline{(x^\intercal x' + a)}}^k$.
- **Exponential.** For $a \in \mathbb{R}$, exponential kernel is $K(x, x') = \exp\left(\frac{x^\intercal x'}{a^2}\right)$.
- **Normalized.** For a PDS K, its normalized kernel $\widehat{K}$ (defined below) is also PDS.

$$* \qquad \widehat{K}(x, x') = \begin{cases} 0 & , \quad K(x, x) = 0 \vee K(x', x') = 0 \\ \frac{K(x, x')}{\sqrt{K(x, x)\, K(x', x')}} & , \quad \text{o.w.} \end{cases}$$

- **Gaussian.** For $a \in \mathbb{R}$, <u>Gaussian kernel</u> (or *radial basis function* (RBF)) is

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2a^2}\right).$$

- **Sigmoid.** For $a, b \in \mathbb{R}_+$, a sigmoidal kernel is $K(x, x') = \underline{\underline{\tanh\left(a\, x^\intercal x' + b\right)}}$.

Gaussian versus exponential kernels

Gaussian. For $a \in \mathbb{R}$, Gaussian kernel (or *radial basis function* (RBF)) is $K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2a^2}\right)$.

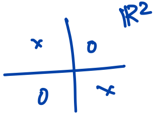
Exponential. For $a \in \mathbb{R}$, exponential kernel is $K_{exp}(x, x') = \exp\left(\frac{x^\top x'}{a^2}\right)$.

$$\exp\left(-\|x - x'\|^2\right) = \exp\left(-2x^\top x' + \|x\|^2 + \|x'\|^2\right) = \exp(-2x^\top x') \cdot \exp(\|x\|^2) \cdot \exp(\|x'\|^2)$$

* RBF is normalized exp. kernel.

$$= \frac{K_{exp}(x, x')}{\sqrt{K_{exp}(x, x) \cdot K_{exp}(x', x')}}$$

Example of XOR



$[x]_1, [x]_2$

Polynomial kernel. $K(x, x') = (x^\intercal x' + a)^2$.

$x, x' \in \mathbb{R}^2$.

$\hookrightarrow \langle \underline{\underline{\varphi(x)}}, \varphi(x') \rangle$

- $\dim(z)$?

Reproducing property of PDS kernels

Kernels can be used to define a large class of functions on $\mathcal{X}$.

If $K : \mathcal{X} \times \mathcal{X} \to \mathcal{Z}$ is PDS, then

Hilbert space

- For all $x \in \mathcal{X}$, $K(x, \cdot) \in \mathcal{Z}$.
- $\mathcal{Z}$ is a **reproducing kernel Hilbert space** associated to K. Specifically, any $z \in \mathcal{Z}$ defines a mapping from $\mathcal{X}$ to $\mathbb{R}$ whose value at any $x \in \mathcal{X}$ is given by a linear combination:

$$z(x) = \langle z, K(x, \cdot)\rangle .$$

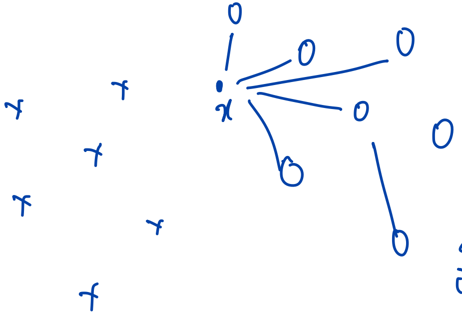
Therefore, for a PDS K we can define $\Psi(x) = K(x, \cdot)$.

For SVM with K, an optimal hyperplane (ignoring the offset) is given by $z^* := \sum_{i=1}^{n} \lambda_i y_i K(x_i, \cdot)$.

Geometric interpretation

$$[K(x,y)]_{y \in \mathcal{X}}$$

For a PDS K we can define $\Psi(x) = K(x,\cdot)$ and $z(x) = \langle z, K(x,\cdot)\rangle$.
For SVM with K, an optimal hyperplane (ignoring the offset) is given by $z^* := \sum_{i=1}^{n} \lambda_i y_i K(x_i,\cdot)$.



$$K(x,x_i) = \langle \Psi(x), \Psi(x_i) \rangle$$

$$x_i \in S$$

$$\sum_{i=1}^{n} \xi_i \, y_i \langle \Psi(x), \Psi(x_i) \rangle \qquad \forall\, i \in S$$

Beyond SVM: wider application of kernels

We can also apply kernels in other machine learning tasks like regression, dimensionality reduction or clustering.

Consider the following optimization problem associated with a mapping z induced over $\mathcal{X}$ by a **PDS kernel** $K : \mathcal{X} \times \mathcal{X} \to \mathcal{Z}$.

$$\underset{z \in \mathcal{Z}}{\text{Minimize}} \; M(\|z\|) + \mathcal{L}(z(x_1), \ldots, z(x_n)), \tag{Opt K}$$

where $M : \mathbb{R} \to \mathbb{R}$ is monotonically non-decreasing and $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$.

Representer theorem: (Opt K) admits a solution $z^* = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$.

What is M and $\mathcal{L}$ for SVM with PDS K?

Limitations of kernel trick

- **Kernel selection is challenging.** Requires domain expertise and lot of experimentation.
- **Not very scalable.** Can be expensive to implement on large datasets.
  Can be tackled to certain extent through *approximate kernel feature maps*.
- **High sensitivity to kernel parameters.** A small change in kernel parameters can drastically change SVM's performance.

References & further readings

The lecture notes are based on Chapter 6 of "Foundations of Machine Learning" by M. Mohri, A. Rostamizadeh, and A. Talwalkar.

Additional reading:

- **Learning guarantee of a PDS kernel based method:** Section 6.3.3.
- **Approximate kernel feature maps:** Section 6.6.