Lecture 17: Learning with Kernels

Isabel Valera

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

16.06.2021

Outline

Main references

- Learning with Kernels - Chapter 2
- Bishop - Chapter 6

Outline

Kernel based learning

Learning with kernels:

- As hypothesis space we use the RKHS $\mathcal{H}_k$ associated to the kernel k,
- As regularization functional we use: $\Omega(f) = \|f\|^2_{\mathcal{H}_k}$ (or more generally a strictly monotonically increasing function of $\|f\|_{\mathcal{H}_k}$)

Regularized empirical risk minimization problem with a RKHS as hypothesis space:

$$f^* = \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda \, \Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big),$$

Important observations

Problems

- The RKHS has often very high dimension or is even infinite dimensional. This means we have a very high dimensional hypothesis space.
- Thus, there is a danger of **overfitting**!

Solution:

- Regularization + **the representer theorem**!
- Effectively we are working in an n-dimensional subspace of $\mathcal{H}_k$!

Representer Theorem

Theorem (Representer Theorem)

Denote by $\Omega : [0, \infty) \to \mathbb{R}$ a strictly monotonically increasing function. Let $\mathcal{X}$ be the input space, $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ an arbitrary loss function and $\mathcal{H}_k$ the reproducing kernel Hilbert space associated to the kernel k. Then, each minimizer $f^* \in \mathcal{H}_k$ of the regularized empirical risk

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda \Omega\Big(\|f\|_{\mathcal{H}_k}^2 \Big),$$

admits a representation as

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

Note also that $\|f^*\|_{\mathcal{H}_k}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$.

Proof I

- $\mathcal{G} = \mathrm{Span}\{k(\mathbf{x}_i, \cdot) \mid i = 1, \ldots, n\}$ is the finite dimensional subspace of $\mathcal{H}_k$ spanned by the data.
- Decompose any $f \in \mathcal{H}_k$ into $f^{\parallel} \in \mathcal{G}$ and the orthogonal part $f^{\perp} \in \mathcal{G}^{\perp}$. Then,

$$f(\mathbf{x}) = f^{\parallel}(\mathbf{x}) + f^{\perp}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + f^{\perp}(\mathbf{x}).$$

- Note that since $k(\mathbf{x}_i, \cdot) \in \mathcal{G}$ and $f^{\perp} \in \mathcal{G}^{\perp}$ we have,

$$\langle f^{\perp}, k(\mathbf{x}_i, \cdot) \rangle_{\mathcal{H}_k} = f^{\perp}(\mathbf{x}_i) = 0,$$

for all $i = 1, \ldots, n$. Therefore,

$$f(\mathbf{x}_j) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j) + f^{\perp}(\mathbf{x}_j) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}_j).$$

Moreover,

$$\Omega\left(\|f\|^2_{\mathcal{H}_k} \right) = \Omega\left(\left\| f^{\parallel} \right\|^2_{\mathcal{H}_k} + \|f^{\perp}\|^2_{\mathcal{H}_k} \right) \geq \Omega\left(\left\| f^{\parallel} \right\|^2_{\mathcal{H}_k} \right)$$

Proof II

In words:

- Any function in the RKHS $\mathcal{H}_k$ decomposes as $f(\mathbf{x}) = f^{\parallel}(\mathbf{x}) + f^{\perp}(\mathbf{x})$.
- The training emprirical risk of any function $f(\mathbf{x})$ in $\mathcal{H}_k$ depends only on $f^{\parallel}(\mathbf{x})$.
- The regularizatiom term $\Omega\left(\|f\|^2_{\mathcal{H}_k}\right)$ is minimized when the optimal solution $f^*(\mathbf{x})$ can be written in terms of only $f^{\parallel}$.
- Thus, the solution to the regularized emprirical risk in the RKHS can always be written as:

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}).$$

Kernelization of algorithms

When? I.e., **which learning methods can be used with kernels?**

- Any regularized empirical risk minimization problem of the form,

$$f^* = \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big).$$

- Any method which can be formulated only using inner products (usually inner product in $\mathbb{R}^d$)

How? **Replace inner product with kernel, or equivalently, use the the representer theorem:**

- Final function: $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$.
- Regularizer: $\|f\|^2_{\mathcal{H}_k} = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$.

Kernelization of algorithms II

- **Optimization point of view:** Transformation of any regularized empirical risk minimization problem of the form,

$$f^* = \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda \Omega\Big(\|f\|^2_{\mathcal{H}_k} \Big)$$

$$\Downarrow$$

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} L\Big(y_i, \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \mathbf{x}_i)\Big) + \lambda \Omega\Big(\sum_{i,j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)\Big)$$

  and $f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* k(\mathbf{x}_i, \mathbf{x})$.
- **Geometric point of view:**
  - Map data to high-dimensional feature space: $\phi : \mathcal{X} \to \mathcal{H}_k$
  - Apply linear algorithm in $\mathcal{H}_k$. Equivalently, replace inner product with kernel function,

$$\langle \mathbf{x}, \mathbf{x}' \rangle_{\mathbb{R}^d} \quad \Longrightarrow \quad k(\mathbf{x}, \mathbf{x}') = \langle \Phi_{\mathbf{x}}, \Phi_{\mathbf{x}'} \rangle_{\mathcal{H}_k} .$$

General Scheme

Replace inner products with kernels:

- any linear method can be kernelized,
- often the dual formulation is more easily accessible and better suited for optimization,
- Kernel Logistic Regression, Kernel Fisher Discriminant Analysis, Kernel PCA, Kernel Perceptron, ...

Outline

SVM I

The soft margin SVM is formulated using **slack variables** $\xi_i \geq 0$.

$$\min_{\mathbf{w}\in\mathbb{R}^d,\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\text{subject to: } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, \quad \forall\, i = 1, \ldots, n, \quad \xi_i \geq 0,$$

- the geometric margin is given by $\frac{2}{\|\mathbf{w}\|_2}$,
- maximizing the margin corresponds to minimizing $\|\mathbf{w}\|_2$,
- slack variables allow points to get inside the margin - soft margin

SVM II

SVM = RERM with Hinge loss and squared regularizer:

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

- error parameter C is inverse to the regularization parameter $\lambda = \frac{1}{C}$.

Dual problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle\,,$$

$$\text{subject to: } 0 \le \alpha_i \le \frac{C}{n}, \quad i = 1,\dots,n, \quad \sum_{i=1}^{n} y_i\alpha_i = 0.$$

Kernalized SVM

SVM = RERM with Hinge loss and squared regularizer:

$$\min_{f\in\mathcal{H}_k,\ b\in\mathbb{R}} C\,\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1-y_i(\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle + b)\right) + \|\mathbf{w}\|_{\mathcal{H}_k}^2\,,$$

becomes with the representer theorem,

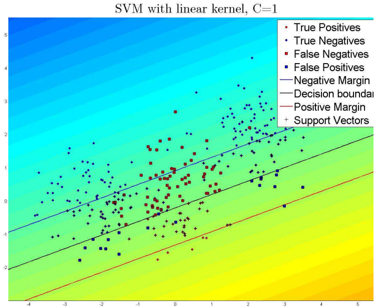
$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n,\ b\in\mathbb{R}} C\,\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1-y_i\left(\sum_{j=1}^{n}\alpha_j k(\mathbf{x}_j,\mathbf{x}_i)+b\right)\right)+\sum_{i,j=1}^{n}\alpha_i\alpha_j k(xb_i, xb_j),$$

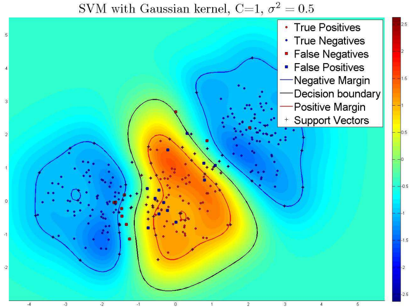
The dual problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\, k(\mathbf{x}_i,\mathbf{x}_j),$$

$$\text{subject to: } 0\le\alpha_i\le\frac{C}{n},\quad i=1,\dots,n,\quad \sum_{i=1}^{n}y_i\alpha_i = 0.$$

Example of Kernalized SVM





Left: the result of the linear SVM with error parameter C - clearly no linear hyperplane can solve this problem. **Right:** the result of the SVM with a Gaussian kernel with $\sigma^2 = \frac{1}{2}$ and $C = 1$. We observe that the Gaussian kernel can nicely identify the class structure. (Image by Prof. Hein)

Outline

Regularization

What is the purpose of regularization?

- penalize functions which are not smooth, i.e., functions where small changes in the data lead to large changes in the prediction.
- regularization functional should measure complexity of the function.

How can we measure smoothness of a function?

- Penalize the derivatives of a function e.g. $\Omega(f) = \int_{\mathbb{R}^d} \|\nabla f\|_2^2 \, dx$.
- How can we achieve that using a RKHS? Can we see directly from a kernel what kind of regularization functional it induces?

Regularization II

Translation invariant kernels in $\mathbb{R}^d$

$$k(x, y) = k(x - y).$$

What does translation invariant mean?

- *What?* Translating all feature vectors by a constant vector $c \in \mathbb{R}^d$, $x \mapsto x + c$, does not change the kernel.

$$k\big(x+c, y+c\big) = k\big((x+c)-(y+c)\big) = k\big(x+c-y-c\big) = k\big(x-y\big) = k\big(x, y\big).$$

- *When?* Use them if only **relative** properties of the features are important, but not **absolute** ones.

Translation and rotation invariant kernels

A **translation and rotation invariant kernel** has the form

$$k(x, y) = \phi(\|x - y\|^2).$$

Such kernels are called **radial**.

What means rotational invariance?
Let R be an orthogonal matrix, that is $RR^T = R^T R = \mathbb{1}$, then

$$\begin{aligned} k(Rx, Ry) &= \phi\big(\,\|Rx - Ry\|^2\,\big) = \phi\big(\,\langle R(x-y), R(x-y)\rangle\,\big) \\ &= \phi\big(\,\langle (x-y), R^T R(x-y)\rangle\,\big) = \phi\big(\,\langle x-y, x-y\rangle\,\big) = \phi\big(\,\|x-y\|^2\,\big) \\ &= k(x, y). \end{aligned}$$

Applying a rotation on the whole space does not change the kernel.

Translation and rotation invariant kernels II

Standard radial kernels:

$$\text{Gaussian kernel:} \qquad k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right),$$

$$\text{Laplace kernel:} \qquad k(x,y) = \exp\Big(-\lambda\,\|x-y\|\Big).$$

Outline

Summary

Summary via example:

Go to Jupyter notebook on Kernalized Ridge Regression.

Homeworks:

Kenalized Logistic Regression.