2022

Lecture 4: SVM (wrap-up) and Kernels

Week 4

Lecturer: Tianyu Wang

1 SVM with Soft Margin

Last time, we considered linearly separable datasets $\{(x_i, y_i)\}_{i=1}^n$. If the dataset is not linearly separable, we can allow the constraints to be violated and suffer a penalty when the constraints are violated. Recall the optimization problem from last lecture is

$$\min_{w,b} \frac{1}{2}\|w\|_2^2, \quad \text{such that} \quad y_i\left(w^\top x_i + b\right) \geq 1, \quad \forall i = 1, 2, \cdots, n.$$

We can introduce penalty variables ξ_i, one for each data point, and rewrite the optimization problem as

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \xi_i, \quad \text{subject to} \quad y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \quad \forall i = 1, 2, \cdots, n, \;\; (1)$$

for some *hyperparameter* C.

Note that ξ_i satisfies

$$\xi_i \geq \min\{0, 1 - y_i(w^\top x_i + b)\}.$$

When $y_i(w^\top x_i + b) \geq 1$ (i.e., the margin is large), we have $\xi_i = 0$. Otherwise $\xi_i \geq 1 - y_i(w^\top x_i + b)$. Since we need the minimize the sum of ξ_i, the optimal ξ_i (otherwise it's not optimal) satisfies

$$\xi_i = \begin{cases} 0, & \text{if } y_i(w^\top x_i + b) \geq 1, \\ 1 - y_i(w^\top x_i + b), & \text{otherwise.} \end{cases}$$

In other words, $\xi_i = \lfloor 1 - y_i(w^\top x_i + b)\rfloor_+$, where $\lfloor z\rfloor_+ = \max\{0, z\}$. Recall $\lfloor z\rfloor_+$ is the *hinge loss* discussed before.

The unconstrained objective for SVM (with soft margin) is

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \lfloor 1 - y_i(w^\top x_i + b)\rfloor_+,$$

which is minimizing the hinge loss plus a regularization term. Note that there is the C is not a variable in the optimization problem. Such quantities are called hyperparameters and are often pre-determined. Next we discuss some standard methods for finding good hyperparameters.

2 Hyperparameter Selection by Cross Validation

I'll draw some illustrations on board/screen.

3 Kernels

The dual problem for SVM with soft margin is

$$\Theta_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \quad \text{subject to } \begin{cases} 0 \le \alpha_i \le C, \ i = 1, 2, \cdots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases}$$

The detailed derivation for this dual is left as a homework exercise.

In this dual problem, one can replace the dot product $\boldsymbol{x}_i^\top x_j$ with a more general inner product.

Definition 3.1 (Hilbert space)**.** A Hilbert space is a complete inner product space.

Note. We will only work with real inner product spaces (and thus real Hilbert spaces).

Examples of Hilbert spaces. 1. The Euclidean space with the dot product. 2. The space of real square integrable functions $L_2(\mathbb{R}^d)$ (functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{x \in \mathbb{R}^d} (f(x))^2 dx < \infty$) with inner product $\langle f, g \rangle_{L_2(\mathbb{R}^d)} = \int_{x \in \mathbb{R}^d} f(x) g(x) dx$.

One can think of the Hilbert space as an infinite-dimensional generalization of the Euclidean space.

3.1 Reproducing Kernel Hilbert Spaces

Before proceeding to reproducing kernel Hilbert spaces, we review some concepts from linear algebra. Recall that any vector $x \in \mathbb{R}^d$ is associated with a linear map from $\mathbb{R}^d$ to $\mathbb{R}$ that send $y \in \mathbb{R}^d$ to $x^\top y \in \mathbb{R}^d$. This association between the vectors and the linear maps is isomorphic. This property generalizes to Hilbert spaces as well.

Definition 3.2 (Linear functionals)**.** Consider a real Hilbert space $\mathcal{H}$. A (continuous) linear functional ϕ is a map from $\mathcal{H}$ to $\mathbb{R}$ such that

1. $\phi(f + g) = \phi(f) + \phi(g)$, for all $f, g \in \mathcal{H}$.
2. $\phi(\alpha f) = \alpha \phi(f)$, for all $f \in \mathcal{H}, \alpha \in \mathbb{R}$.

Theorem 3.3 (Corollary of the Riesz representation theorem)**.** *Let $\mathcal{H}$ be a Hilbert space with inner product $\langle , \rangle$. For every continuous linear functional $\varphi : \mathcal{H} \to \mathbb{R}$, there exists a unique $f_\varphi \in \mathcal{H}$ such that*

$$\varphi(g) = \langle g, f_\varphi \rangle \quad \textit{for all } g \in \mathcal{H}.$$

Definition 3.4 (Reproducing Kernel Hilbert Space (RKHS))**.** The evaluation functional over the Hilbert space of functions $\mathcal{H}$ is a linear functional that evaluates each function at a point x,

$$L_x : f \mapsto f(x) \quad \forall f \in H.$$

We say that $\mathcal{H}$ is a reproducing kernel Hilbert space if, for all $x \in X$, L_x is continuous at every f in $\mathcal{H}$.

The Riesz representation theorem implies that for all $f \in \mathcal{H}$, there exists a unique element $K_x \in \mathcal{H}$ with the *reproducing property*, i.e.,

$$f(x) = L_x(f) = \langle f, K_x \rangle .$$

Since K_x is itself an element of $\mathcal{H}$, we can apply the Riesz representation theorem again. For $K_x \in \mathcal{H}$, there exists a unique element $K_y \in \mathcal{H}$ such that

$$K_x(y) = L_y(K_x) = \langle K_y, K_x \rangle .$$

This allows us to define a reproducing kernel.

Definition 3.5 (Reproducing kernel)**.** Let $\mathcal{H}$ be a reproducing kernel Hilbert space whose functions are defined on $\mathcal{X}$. A *reproducing kernel* is a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$k(x, y) = \langle K_x, K_y \rangle ,$$

where $\langle , \rangle$ is the inner product of $\mathcal{H}$.

Examples of kernels.

- Gaussian kernel, radial basis function kernel $k(x, y) = \exp(-\|x - y\|_2^2/\sigma)$ for $x, y \in \mathbb{R}^d$, where σ is a hyperparameter.
- linear kernel $k(x, y) = x^\top y$ for $x, y \in \mathbb{R}^d$.
- polynomial kernel $k(x, y) = (1 + ax^\top y)^p$ for $x, y \in \mathbb{R}^d$, where a and p are hyperparameters.

3.2 The other way around

For machine learning and data science tasks, usually the kernel comes before the Hilbert space. We can use kernels to define reproducing kernel Hilbert spaces as well.

Definition 3.6. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is

- symmetric, if $k(x, y) = k(y, x)$ for all $x, y \in \mathcal{X}$.

- positive definition, if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

holds for any $n \in \mathbb{N}_{>0}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $c_1, \ldots, c_n \in \mathbb{R}$. In addition, strict equality holds only when $c_1 = \cdots = c_n = 0$.

Theorem 3.7 (Moore-Aronszajn)**.** *Suppose k is a symmetric, positive definite kernel on a set $\mathcal{X}$. Then there is a unique Hilbert space of functions on $\mathcal{X}$ for which k is a reproducing kernel.*

Intuitively, this Hilbert space is the completion of the span of the functions $\{k(x, \cdot) : x \in \mathcal{X}\}$. Next we provide a formal proof.

Proof (Optional). For all $x \in \mathcal{X}$, define $K_x = k(\cdot, x)$. Let $\mathcal{H}_0$ be the linear span of $\{K_x : x \in \mathcal{X}\}$. Define an inner product on $\mathcal{H}_0$ by

$$\left\langle \sum_{j=1}^{n} b_j K_{y_j}, \sum_{i=1}^{m} a_i K_{x_i} \right\rangle_{H_0} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j k(y_j, x_i),$$

which implies $k(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}_0}$. The symmetry and positive-definiteness of this inner product follows from the symmetry and positive-definiteness of the kernel k. Linearity of this inner product follows from definition.

Let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ with respect to this inner product. Then $\mathcal{H}$ consists of functions of the form

$$f(x) = \sum_{i=1}^{\infty} a_i K_{x_i}(x) \quad \text{where} \quad \lim_{n\to\infty} \sup_{p\geq 0} \left\| \sum_{i=n}^{n+p} a_i K_{x_i} \right\|_{\mathcal{H}_0} = 0.$$

Now we can check the reproducing property:

$$\langle f, K_x \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} a_i \langle K_{x_i}, K_x \rangle_{\mathcal{H}_0} = \sum_{i=1}^{\infty} a_i k(x_i, x) = f(x).$$

To prove uniqueness, let $\mathcal{G}$ be another Hilbert space that contains span of $\{k(\cdot, x_i)\}$ for which k is a reproducing kernel. Since k is a reproducing kernel, for every x and y in $\mathcal{X}$, it holds that

$$\langle K_x, K_y \rangle_{\mathcal{H}} = k(x, y) = \langle K_x, K_y \rangle_{\mathcal{G}}.$$

By linearity, $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{G}}$ on the linear span of $\{K_x : x \in X\}$. Then $\mathcal{H} \subseteq \mathcal{G}$ because $\mathcal{G}$ is complete and contains $\mathcal{H}_0$ and hence contains its completion.

Now we need to prove that every element of $\mathcal{G}$ is in $\mathcal{H}$. Let f be an element of $\mathcal{G}$. Since $\mathcal{H}$ is a closed subspace of $\mathcal{G}$, we can write $f = f_{\mathcal{H}} + f_{\mathcal{H}^\perp}$ where $f_{\mathcal{H}} \in \mathcal{H}$ and $f_{\mathcal{H}^\perp} \in \mathcal{H}^\perp$. Now if $x \in \mathcal{X}$ then, since k is a reproducing kernel of $\mathcal{G}$ and $\mathcal{H}$:

$$f(x) = \langle K_x, f \rangle_{\mathcal{G}} = \langle K_x, f_{\mathcal{H}} \rangle_{\mathcal{G}} + \langle K_x, f_{\mathcal{H}^\perp} \rangle_{\mathcal{G}} = \langle K_x, f_{\mathcal{H}} \rangle_{\mathcal{G}} = \langle K_x, f_{\mathcal{H}} \rangle_{\mathcal{H}} = f_{\mathcal{H}}(x),$$

where we have used the fact that K_x belongs to $\mathcal{H}$ so that its inner product with $f_{\mathcal{H}^\perp}$ in $\mathcal{G}$ is zero. This shows that $f = f_{\mathcal{H}}$ in G and concludes the proof. □

Reproducing kernel properties summary

For most machine learning tasks, we pick a kernel k defined on $\mathcal{X}$. Then in some Hilbert space $\mathcal{H}$ of real-valued functions defined on $\mathcal{X}$, the kernel k is reproducing:

$$\begin{aligned} \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}} &= k(x,y) \quad \forall x, y \in \mathcal{X}, \\ \langle k(x,\cdot), f\rangle_{\mathcal{H}} &= f(x), \quad \forall f \in \mathcal{H}. \end{aligned}$$

3.3 The representer theorem

The representer theorem is a machine learning version of the Riesz representation theorem.

Theorem 3.8 (Representer theorem)**.** *Consider a symmetric positive-definite real-valued kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ with a corresponding reproducing kernel Hilbert space H. Let there be given*

- *a training dataset $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$,*
- *a strictly increasing real-valued function $\Omega : [0, \infty) \to \mathbb{R}$,*
- *an arbitrary error function $E : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$,*

which together define the following regularized empirical risk functional on H:

$$f \mapsto E\left((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n))\right) + \Omega\left(\|f\|\right),$$

where $\|\cdot\|$ is the norm of $\mathcal{H}$. Then, any minimizer of the regularized empirical risk

$$f^* = argmin_{f \in H} \left\{E\left((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n))\right) + \Omega\left(\|f\|\right)\right\},$$

admits a representation of the form:

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),$$

for some $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{R}$.

Proof. Given any $x_1, ..., x_n$, one can use orthogonal projection to decompose any $f \in H$ into a sum of two functions, one lying in $span\left\{k(\cdot, x_1), ..., \varphi(\cdot, x_n)\right\}$, and the other lying in the orthogonal complement:

$$f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + v,$$

where $\langle v, k(\cdot, x_i)\rangle = 0$ for all i.

The above orthogonal decomposition and the reproducing property together show that applying f to any training point x_j produces

$$f(x_j) = \left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + v, k(\cdot, x_j) \right\rangle = \sum_{i=1}^{n} \alpha_i \langle k(\cdot, x_i), k(\cdot, x_j) \rangle,$$

which is independent of v. Consequently, the value of the error function E (the first term in the empirical risk objective) is independent of v. For the regularization term, since v is orthogonal to $\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and Ω is strictly monotonic, we have

$$\begin{aligned}
\Omega\left(\|f\|\right) &= \Omega\left(\|\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + v\|\right) \\
&= \Omega\left(\sqrt{\|\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)\|^2 + \|v\|^2}\right) \\
&\geq \Omega\left(\|\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)\|\right).
\end{aligned}$$

Therefore setting $v = 0$ strictly decreases the second term, and does not affect the first term. Consequently, any minimizer f^* of the regularized empirical risk must satisfy $v = 0$, i.e., it must be of the form

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),$$

which finishes the proof.

□

3.4 Feature Maps

Theorem 3.9 (Corollary of Mercer's theorem)**.** *Suppose k is a continuous symmetric positive-definite kernel on $\mathcal{X}$. If $\int_{x \in \mathcal{X}} k(x, x)dx < \infty$, then there is an orthonormal set $\{e_i\}_i$ of $L_2(\mathcal{X})$ such that the kernel k admits the following representation*

$$k(s, t) = \sum_{j=1}^{\infty} \lambda_j \, e_j(s) e_j(t), \quad \forall s, t \in \mathcal{X},$$

where λ_j are nonnegative for all j.

This theorem bring us from an uncountable world to a countable world. This shows that, for any kernel k, there exists a *feature map* ϕ, such that $\phi(x) = [\sqrt{\lambda_1} e_1(x), \sqrt{\lambda_2} e_2(x), \cdots]$ and $k(x, y) = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_j(y)$. Such maps ϕ are called *feature maps*, and they fully specify the reproducing kernel and thus the reproducing kernel Hilbert space.

Homework. Find a feature map for the Gaussian kernel.

Acknowledgement

The content is a reorganization of works of many researchs, including Schölkopf, Herbrich, Smola, Kimeldorf, Wahba. TW used lecture notes by Cynthia Rudin to compile this notes. A thank you to wikipedia contributors.