EECS 598: Statistical Learning Theory, Winter 2014 **Topic 13**

Kernel Methods and the Representer Theorem

Lecturer: Clayton Scott *Scribe: Mohamad Kazem Shirani Faradonbeh*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

These notes describe kernel methods for supervised learning problems. We have an input space $\mathcal{X}$, an output space $\mathcal{Y}$, and training data $(x_1, y_1), ..., (x_n, y_n)$. Keep in mind two important special cases: binary classification where $\mathcal{Y} = \{-1, 1\}$, and regression where $\mathcal{Y} \subseteq \mathbb{R}$.

2 Loss Functions

Definition 1. *A* loss function *(or just* loss*) is a function $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. For a loss L and joint distribution on (X, Y), the L-risk of a function $f : \mathcal{X} \to \mathbb{R}$ is $R_L(f) := \mathbb{E}_{XY} L(Y, f(X))$.*

Examples. (a) In regression with $\mathcal{Y} = \mathbb{R}$, a common loss is the *squared error loss*, $L(y, t) = (y - t)^2$, in which case $R_L(f) = \mathbb{E}_{XY}(Y, f(X))^2$ is the mean squared error.

(b) In classification with $\mathcal{Y} = \{-1, 1\}$, the *0-1 loss* is $L(y, t) = \mathbf{1}_{\{\text{sign}(t) \neq y\}}$ in which case $R_L(f) = P_{XY}(\text{sign}(f(X)) \neq Y)$ is the probability of error.

(c) The 0-1 loss $L(y, t)$ is neither differentiable nor convex in its second argument, which makes the empirical risk difficult to optimize in practice. A *surrogate loss* is a loss that serves as a proxy for another loss, usually because it possesses desirable qualities from a computational perspective. Popular convex surrogates for the 0-1 loss are the *hinge loss*

$$L(y, t) = \max(0, 1 - yt)$$

and the *logistic loss*

$$L(y, t) = \log(1 + e^{-yt}).$$

Remarks. (a) In classification we associate $f : \mathcal{X} \to \mathbb{R}$ to the classifier $h(x) = \text{sign}(f(x))$ where $\text{sign}(t) = 1$ for $t \geq 0$ and $\text{sign}(t) = -1$ for $t < 0$. The convention for sign(0) is not important.

(b) To be consistent with our earlier notation, we write $R(f)$ for $R_L(f)$ when L is the 0-1 loss.

(c) In the classification setting, if $L(y, t) = \phi(yt)$ for some function ϕ, we refer to L as a *margin loss*. The quantity $yf(x)$ is called the *functional margin*, which is different from but related to the geometric margin, which is the distance from a point x to a hyperplane. We'll discuss the functional margin more later.

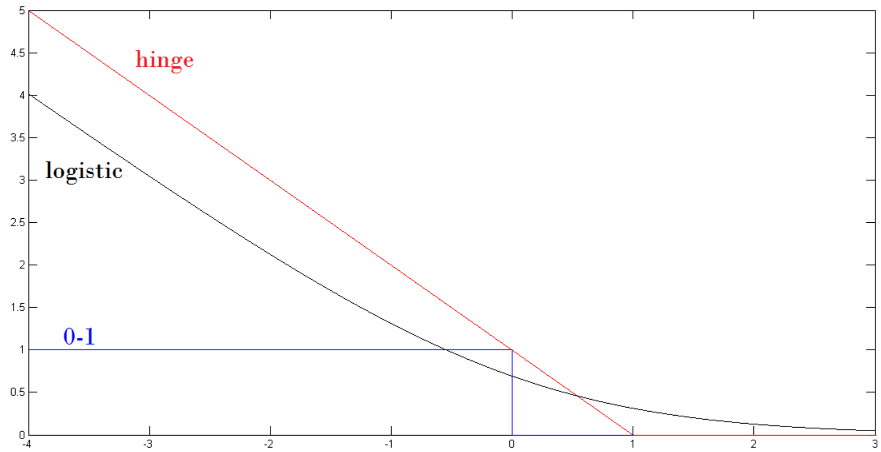

Figure 1: The logistic and hinge losses, as functions of yt, compared to the loss $\mathbf{1}_{\{ty \leq 0\}}$, which upper bounds the 0-1 loss $\mathbf{1}_{\{\text{sign}(t) \neq y\}}$.

3 The Representer Theorem

Let k be a kernel on $\mathcal{X}$ and let $\mathcal{F}$ be its associated RKHS. A *kernel method (or kernel machine)* is a discrimination rule of the form

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{F}}^2 \tag{1}$$

where $\lambda \geq 0$. Since $\mathcal{F}$ is possibly infinite dimensional, it is not obvious that this optimization problem can be solved efficiently. Fortunately, we have the following result, which implies that (2) reduces to a finite dimensional optimization problem.

Theorem 1 (The Representer Theorem)**.** *Let k be a kernel on $\mathcal{X}$ and let $\mathcal{F}$ be its associated RKHS. Fix $x_1, \ldots, x_n \in \mathcal{X}$, and consider the optimization problem*

$$\min_{f \in \mathcal{F}} \; D(f(x_1), \ldots, f(x_n)) + P(\|f\|_{\mathcal{F}}^2), \tag{2}$$

where P is nondecreasing and D depends on f only though $f(x_1), \ldots, f(x_n)$. If (2) has a minimizer, then it has a minimizer of the form $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ where $\alpha_i \in \mathbb{R}$. Furthermore, if P is strictly increasing, then every solution of (2) has this form.

Proof. Denote $J(f) = D(f(x_1), ..., f(x_n)) + P(\|f\|_{\mathcal{F}}^2)$. Consider the subspace $S \subset \mathcal{F}$ given by $S = \text{span}\{k(\cdot, x_i) : i = 1, \ldots, n\}$. S is finite dimensional and therefore closed. The projection theorem then implies $\mathcal{F} = S \oplus S^{\perp}$, i.e., every $f \in \mathcal{F}$ we can uniquely written $f = f_{\|} + f_{\perp}$ where $f_{\|} \in S$ and $f_{\perp} \in S^{\perp}$. Noting that $\langle f_{\perp}, k(\cdot, x_i) \rangle = 0$ for each i, the reproducing property implies

$$\begin{aligned} f(x_i) &= \langle f, k(\cdot, x_i) \rangle \\ &= \langle f_{\|}, k(\cdot, x_i) \rangle + \langle f_{\perp}, k(\cdot, x_i) \rangle \\ &= f_{\|}(x_i). \end{aligned}$$

Then

$$\begin{aligned} J(f) &= D(f(x_1), ..., f(x_n)) + P(\|f\|_{\mathcal{F}}^2) \\ &= D(f_\|(x_1), ..., f_\|(x_n)) + P(\|f\|_{\mathcal{F}}^2) \\ &\geq D(f_\|(x_1), ..., f_\|(x_n)) + P(\|f_\|\|_{\mathcal{F}}^2) \\ &= J(f_\|). \end{aligned}$$

The inequality holds because P is non-decreasing and $\|f\|_{\mathcal{F}}^2 = \|f_\|\|_{\mathcal{F}}^2 + \|f_\perp\|_{\mathcal{F}}^2$. Therefore if f is a minimizer of $J(f)$ then so is $f_\|$. Since $f_\| \in S$, it has the desired form. The second statement holds because if P is strictly increasing then for $f \notin S, J(f) > J(f_\|)$. □

4 Kernel Ridge Regression

Now let's use the representer theorem in the context of regression with the squared error loss, so that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. The kernel method solves

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2,$$

so the representer theorem applies with $D(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^{n} (f(x_i) - y_i)^2$ and $P(t) = \lambda t$, and we may assume $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$. So it suffices to solve

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{n} \alpha_j k(x_j, x_i))^2 + \lambda \| \sum_{j=1}^{n} \alpha_j k(\cdot, x_j)\|^2.$$

Denoting $K = [k(x_i, x_j)]_{i,j=1}^n$ and $y = (y_1, ..., y_n)^T$, the objective function is

$$J(\alpha) = \alpha^T K \alpha - 2y^T K \alpha + y^T y + \lambda \alpha^T K \alpha.$$

Since this objective is strongly convey, it has a unique minimizer. Assuming K is invertible, $\frac{\partial J}{\partial \alpha} = 0$ gives $\alpha = (K + \lambda I)^{-1} y$ and $\widehat{f}(x) = \alpha^T \underline{k}(x)$ where $\underline{k}(x) = (k(x, x_1), ..., k(x, x_n))^T$.

This predictor is kernel ridge regression, which can alternately be derived by kernelizing the linear ridge regression predictor. Assuming x_i, y_i have zero mean, consider linear ridge regression:

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|^2.$$

The solution is

$$\beta = (XX^T + \lambda I)^{-1} X y$$

where $X = [x_1 \ \cdots \ x_n] \in \mathbb{R}^{d \times n}$ is the data matrix. Using the matrix inversion lemma one can show

$$\beta^T x = y^T X^T (XX^T + \lambda I)^{-1} x = y^T (X^T X + \lambda I)^{-1} (\langle x, x_1 \rangle, \ldots, \langle x, x_n \rangle)^T$$

where the inner product is the dot product. Note that $X^T X$ is a Gram matrix, so the above predictor uses elements of $\mathcal{X}$ entirely via inner products. If we replace the inner products by kernels,

$$\langle x, x' \rangle \mapsto k(x, x') = \langle \Phi(x), \Phi(x') \rangle,$$

it is as if we are performing ridge regression on the transformed data $\Phi(x_i)$, where Φ is a feature map associated to k. The resulting predictor is now nonlinear in x and agrees with the predictor derived from the RKHS perspective.

5 Support Vector Machines

A support vector machine (without offset) is the solution of

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i f(x_i)) + \frac{\lambda}{2} \|f\|^2. \tag{3}$$

By the representer theorem and strong convexity, the unique solution has the form $f = \sum_{i=1}^{n} r_i k(\cdot, x_i)$. Plugging this into (3) and applying Lagrange multiplier theory, it can be shown that the optimal r_i have the form $r_i = y_i \alpha_i$ where α_i solve

$$\begin{aligned} \min_{\alpha} \quad & -\sum_i \alpha_i + \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \\ \text{s.t} \quad & 0 \le \alpha_i \le \frac{1}{n\lambda}, \ i = 1, \ldots, n. \end{aligned}$$

This classifier is usually derived from an alternate perspective, that of maximizing the geometric (soft) margin of a hyperplane, and then applying the kernel trick as was done with kernel ridge regression. This derivation should be covered in EECS 545 Machine Learning.

Exercises

1. In some kernels methods it is desirable to include an offset term. Prove an extension of the representer theorem where the class being minimized over is $\mathcal{F} + \mathbb{R}$, the set of all functions of the form $f(x) + b$ where $f \in \mathcal{F}$ and $b \in \mathbb{R}$