Homework 12

36-465/665, Spring 2021

Due at 6 pm on Thursday, 29 April 2021

Agenda: Filling in some important details on support vector machines

We have a data set $(x_1, y_1), \ldots (x_n, y_n)$. Each class label y_i is either $+1$ or -1. We also have a kernel function $K(x, z)$, resulting in an $[n \times n]$ kernel matrix $\mathbf{K}$ with $K_{ij} = K(x_i, x_j)$. The **margin** of function g at the ith data point, γ_i, is defined as $y_i g(x_i)$. The over-all margin of the function g is $\min_{i \in 1:n} \gamma_i$, written $M(g)$.

We will begin working in feature space, so we are interested in functions of the form

$$g(x) = b_0 + s(x) = b_0 + \sum_{i=1}^{\infty} \beta_j \phi_j(x) = b_0 + \langle \beta, \phi(x) \rangle \tag{1}$$

Here, the transformations ϕ_j are the features (implicitly) defined by the kernel K, via $K(x, z) = \sum_{j=1}^{\infty} \phi_j(x)\phi_j(z)$.

1. *Turning margin maximization into a quadratic program* We'd like to try maximizing the over-all margin. Because that's a minimum over all data points, one way to set up the optimization problem is as follows:

$$\begin{aligned} \min_{b_0, \beta, \gamma} \quad & -\gamma & (2)\\ \text{subject to} \quad & y_i(b_0 + \langle \beta, \phi(x_i) \rangle) \geq \gamma, \ i \in 1:n & (3)\\ \text{and} \quad & \|\beta\| = 1 & (4) \end{aligned}$$

   a. (5) What's the relationship between γ and the margin?
   b. (6) Explain why requiring $\|\beta\| = 1$ doesn't change the performance of the resulting classifier. If it doesn't change the performance, why impose the requirement?
   c. (5) The Lagrangian for this constrained optimization problem is

$$\mathcal{L}(\beta, b_0, \gamma, \alpha, \lambda) = -\gamma - \sum_{i=1}^{n} \alpha_i \left[y_i(b_0 + \langle \beta, \phi(x_i) \rangle) - \gamma \right] + \lambda(\|\beta\|^2 - 1) \tag{5}$$

   Explain what each term in this Lagrangian is doing — which come from the original objective function, and which ones enforce the constraints? Why does $\|\beta\|^2$ appear here, rather than $\|\beta\|$?
   d. (5) Differentiate $\mathcal{L}$ with respect to β, γ and b_0 to show that the following equations hold at the optimum:

$$\begin{aligned} \beta &= \frac{1}{2\lambda} \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) & (6)\\ \sum_{i=1}^{n} \alpha_i &= 1 & (7)\\ \sum_{i=1}^{n} \alpha_i y_i &= 0 & (8) \end{aligned}$$

e. (5) Use Q1d to show that, at the optimum,

$$\mathcal{L}(\beta, b_0, \gamma, \alpha, \lambda) = -\frac{1}{4\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K_{ij} - \lambda \tag{9}$$

Hint: $K_{ij} = K(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$.

f. (5) Show that optimizing Q1e over λ gives

$$\mathcal{L}(\alpha) = -\sqrt{\sum_{i,j}\alpha_i\alpha_j y_i y_j K_{ij}} \tag{10}$$

g. (6) Explain why the original optimization problem is equivalent to

$$\min_{\alpha} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K_{ij} \tag{11}$$

$$\text{subject to} \quad \sum_{i=1}^{n}\alpha_i = 1 \tag{12}$$

$$\text{and} \quad \sum_{i=1}^{n} y_i\alpha_i = 0 \tag{13}$$

$$\text{and} \quad \alpha_i > 0,\ i \in 1:n \tag{14}$$

h. (5) Explain why the margin of the resulting machine is $\gamma^* = \sqrt{\sum_{i,j}\alpha_i\alpha_j y_i y_j K_{ij}}$.

i. (5) A vector x_i is called a **support vector** if $\alpha_i \neq 0$. All the support vectors will haver margin equal to γ^*. Use this to show that

$$b_0 = y_i(\gamma^*)^2 - \sum_{j=1}^{n}\alpha_j y_j K_{ji} \tag{15}$$

The final form of the support-vector-machine classifier is, for reference

$$g(x) = \text{sgn}\left(b_0 + \sum_{i=1}^{n}\alpha_i y_i K(x, x_i)\right) \tag{16}$$

where b_0 is given in Q1i.

2. *Empirical Rademacher complexity of large-margin kernel machines* We'll write the risk of a classifier g, using the 0-1 loss, as $r(g) = \mathbb{P}(Y \neq \text{sgn}\, g(X))$. In this problem, we'll want to give generalization error bounds in terms of the margin for this risk. As an intermediate step, however, we'll introduce another, more continuous loss function. Define the **margin-based loss function** $\ell_\gamma(u)$ as

$$\ell_\gamma(u) = \begin{cases} 1 & u \leq 0 \\ 1 - u/\gamma & 0 < u \leq \gamma \\ 0 & u > \gamma \end{cases} \tag{17}$$

a. (5) Show that $\ell_\gamma(yg(x)) \geq$ the $0-1$ loss of using the sign of $g(x)$ as a classifier.

b. (5) Explain why Q2a implies that

$$r(g) \leq \mathbb{E}\left[\ell_\gamma(Yg(X))\right] \tag{18}$$

c. (6) Explain why, when g was selected from a class of functions $\mathcal{G}$,

$$\mathbb{P}\left(\mathbb{E}\left[\ell_\gamma(Yg(X))\right] \geq \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma(Y_i g(X_i)) + 2\hat{\mathcal{R}}_n(\ell_\gamma \circ \mathcal{G}) + 3\sqrt{\frac{\log 2/\alpha}{2n}}\right) \leq \alpha \tag{19}$$

where $\ell_\gamma \circ \mathcal{G}$ is the class of $f(x, y) \mapsto \ell_\gamma(yg'(x))$ for all the various $g' \in \mathcal{G}$. *Hint:* Go back to our basic results about empirical Rademacher complexity.

d. (5) A function f is called **Lipschitz continuous**, with **Lipschitz constant** L, if $|f(u_1)-f(u_2)| \leq L|u_1 - u_2|$. (If f is differentiable, this means that its derivative is everywhere $\leq L$.) Show that ℓ_γ is Lipschitz-continuous, and find an expression for its Lipschitz constant. Your answer will involve γ (and possibly other things).

e. (5) Here is a fact about Rademacher complexity which we will use but not prove (see extra credit below). Start with a class of functions $\mathcal{G}$, of known (or bounded) empirical Rademacher complexity $\hat{\mathcal{R}}_n(\mathcal{G})$. Fix a function λ which is Lipschitz-continuous, with Lipschitz constant L, and for which $\lambda(0) = 0$. Define a new class of functions $\mathcal{F} = \{\lambda(g(\cdot)) \ : \ g \in \mathcal{G}\}$. Then $\hat{\mathcal{R}}_n(\mathcal{F}) \leq 2L\hat{\mathcal{R}}_n(\mathcal{G})$. Use this fact to show that

$$\hat{\mathcal{R}}_n(\ell_\gamma \circ \mathcal{G}) \leq 2L\hat{\mathcal{R}}_n(\mathcal{G}) \tag{20}$$

You should be able to be explicit about the value of L here from earlier parts of this question.

f. (6) We now assume that $\mathcal{G}$ is the class of kernel machines with $\|\beta\| = 1$. Using earlier results in the course, state a bound on $\hat{\mathcal{R}}_n(\mathcal{G})$. Your answer should involve the sample size n, the kernel matrix $\mathbf{K}$, and possibly other things.

g. (5) Using the earlier parts of this question, show that

$$\mathbb{P}\left(r(g) \geq \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma(Y_i g(X_i)) + (\text{something}) + 3\sqrt{\frac{\log 2/\alpha}{2n}}\right) \leq \alpha \tag{21}$$

Be explicit about the (something) term, which should involve n, γ, and $\mathbf{K}$ (and possibly other things).

h. (5) Using Q2g, state a bound on the mis-classification risk of a kernel-based classifier, with $\|\beta\| = 1$, which classifies the entire training data set perfectly and achieves margin γ. Does your result match the one stated (without proof!) in Lecture 21? If not, are they compatible?

3. (1) How much time did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Extra credit (5 points total): Prove the result about Rademacher complexity of "compositional" function classes used in Q2. That is, assume that we have a fixed function λ which is Lipschitz-continuous, with constant L, and where $\lambda(0) = 0$, and a class of functions $\mathcal{G}$ of empirical Rademacher complexity $\hat{\mathcal{R}}_n(\mathcal{G})$. Define $\mathcal{F}$ as class of all functions f of the form $f(z) = \lambda(g(z))$ for $g \in \mathcal{G}$. Show that $\hat{\mathcal{R}}_n(\mathcal{F}) \leq 2L\hat{\mathcal{R}}_n(\mathcal{G})$.