10-715 Advanced Introduction to Machine Learning: Homework 3
Kernels, VC dimension, Rademacher Complexity

Released: Wednesday, September 26, 2018
Due: 11:59 p.m. Wednesday, October 3, 2018

Instructions

- **Late homework policy:** Homework is worth full credit if submitted before the due date, half credit during the next 48 hours, and zero credit after that.
- **Collaboration policy:** Collaboration on solving the homework is allowed. Discussions are encouraged but you should think about the problems on your own. When you do collaborate, you should list your collaborators! Also cite your resources, in case you got some inspiration from other resources (books, websites, papers). If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution. We will be assuming that, as participants in a graduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.
- **Online submission:** You must submit your solutions online on Autolab (link: https://autolab.andrew.cmu.edu/courses/10715-f18/assessments). Please use LaTeX to typeset your solutions, and submit a single pdf called **hw3.pdf**.

Problem 1: Construction of kernels [20 points]

1.1 Construction of the Gaussian Kernel function [10 points]

Suppose that $K_1(\boldsymbol{u}, \boldsymbol{v}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $K_2(\boldsymbol{u}, \boldsymbol{v}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are both valid kernel functions and that c_1, $c_2 \geq 0$ are real constants.

1. **[1 point]** Prove that the function $K(\boldsymbol{u}, \boldsymbol{v}) = c_1 K_1(\boldsymbol{u}, \boldsymbol{v}) + c_2 K_2(\boldsymbol{u}, \boldsymbol{v})$ is a valid kernel function.
2. **[1 point]** Prove that the function $K(\boldsymbol{u}, \boldsymbol{v}) = K_1(\boldsymbol{u}, \boldsymbol{v}) K_2(\boldsymbol{u}, \boldsymbol{v})$ is a valid kernel function.
3. **[1 point]** Prove that the function $K(\boldsymbol{u}, \boldsymbol{v}) = p(K_1(\boldsymbol{u}, \boldsymbol{v}))$, where p is a polynomial with positive coefficients, is a valid kernel function.
4. **[1 point]** Prove that $K(\boldsymbol{u}, \boldsymbol{v}) = e^{K_1(\boldsymbol{u}, \boldsymbol{v})}$ is a valid kernel function.
5. **[4 points]** Let $h : \mathbb{R}^d \to \mathbb{R}$, and $K(\boldsymbol{u}, \boldsymbol{v}) = h(\boldsymbol{u})h(\boldsymbol{v})$. Prove that the matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, with $\boldsymbol{K}_{ij} = K(\boldsymbol{u}_i, \boldsymbol{u}_j)$ for $i, j = 1, 2, \ldots n$ and any set of vectors $\boldsymbol{u}_i$, is positive semi-definite (directly, without using Mercer's Theorem).
6. **[2 points]** Prove that $K(\boldsymbol{u}, \boldsymbol{v}) = e^{-\frac{||\boldsymbol{u}-\boldsymbol{v}||^2}{\sigma^2}}$ is a valid kernel function (the Gaussian kernel function).

1.2 Construction of more kernels [10 points]

By using the properties of the previous subsection,

1. **[4 points]** Prove that $K(\boldsymbol{u}, \boldsymbol{v}) = ||\boldsymbol{u}||^2 ||\boldsymbol{v}||^2 e^{(||\boldsymbol{u}||^2 + ||\boldsymbol{v}||^2)}$ is a valid kernel function.
2. **[6 points]** Prove that $K(\boldsymbol{u}, \boldsymbol{v}) = \sinh\left(e^{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}\right)$ is a valid kernel function.

Problem 2: VC Dimension [50 points + 10 points EXTRA CREDIT]

2.1 Warmup [10 points]

For each one of the following function classes, state what the VC dimension is and explain how you found that number.

1. **[2 points]** Circles in $\mathbb{R}^2$. An example is labeled positive if it lies within the circle, and negative otherwise. Assume you may shift the circle and change its size.
2. **[2 points]** Consider the same scenario as in part 1, but now assume the circle is centered around the origin.
3. **[6 points]** Consider $X = \mathbb{R}^1$, where we want to learn $f : X \to \{-1, +1\}$. What is the VC dimension of $H = \{Y = sign(sin(w_1 x + w_2))\}$ (Figure 1)

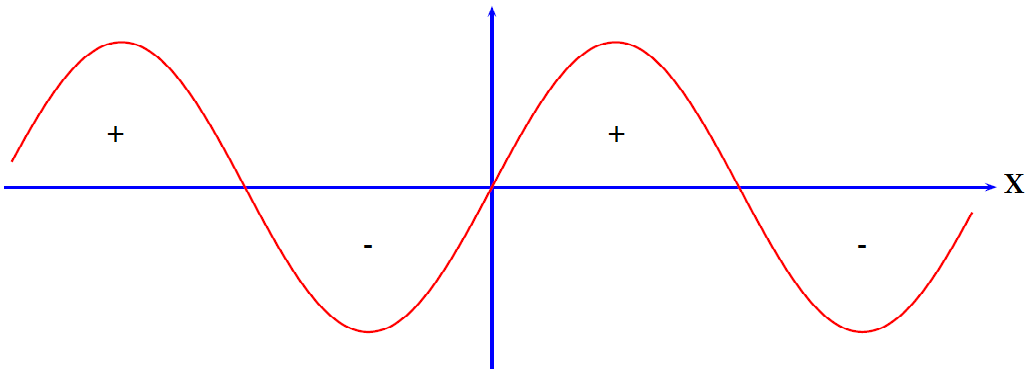

Figure 1: Sine wave classifier

2.2 Linear Separators [30 points]

In this problem, we will derive the VC dimension of linear separators. Let H_n be the set of linear separators in $\mathbb{R}^n$. That is, H_n consists of functions f_a of the following form

$$f_a(\mathbf{X}) = \begin{cases} +1 & \text{if } \sum_{i=1}^{n} a_i X_i > a_0, \\ -1 & \text{otherwise} \end{cases},$$

where $\mathbf{X} = (X_1, \ldots X_n) \in \mathbb{R}^n$ is an n dimensional vector.

1. **[10 points] Lower bound.** Prove that VC-dim$(H_n) \geq n + 1$ by presenting a set of $n + 1$ points in n-dimensional space such that one can partition that set with linear separators in all possible ways. (And, show how one can partition the set in any desired way.)

2. **[10 points] Upper bound.** The following is "Radon's Theorem," from the 1920's.

   Theorem. *Let S be a set of $n + 2$ points in n dimensions. Then S can be partitioned into two (disjoint) subsets S_1 and S_2 whose convex hulls intersect.*

   Show that Radon's Theorem implies that the VC-dimension of halfspaces is *at most* $n + 1$. Conclude that VC-dim$(H_n) = n + 1$.

3. **[10 points]** Now we will prove Radon's Theorem. We will need the following standard fact from linear algebra. If $X_1, \ldots, X_{n+1}$ are $n + 1$ points in n-dimensional space, then they are linearly dependent. That is, there exist real values $\lambda_1, \ldots, \lambda_{n+1}$ *not all zero* such that $\lambda_1 X_1 + \ldots + \lambda_{n+1} X_{n+1} = 0$.

   You may now prove Radon's Theorem however you wish. However, as a suggested first step, prove the following. For any set of $n + 2$ points $X_1, \ldots, X_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ *not all zero* such that $\sum_i \lambda_i X_i = 0$ and $\sum_i \lambda_i = 0$. (This is called *affine dependence*.)

2.3 Majority Voting [10 points]

Show that if hypothesis class H has VC-dimension d, then the class MAJ$_k(H)$ has VC-dimension $O(kd \log kd)$. Here, we define MAJ$_k(H)$ to be the class of functions achievable by taking majority votes over k functions in H. Note that we are only asking for an upper bound here, not a lower bound.

2.3 Boxes [10 points EXTRA CREDIT]

What is the VC-dimension V of the class H of axis-parallel boxes in $\mathbb{R}^d$? That is, $H = \{h_{\mathbf{a},\mathbf{b}} : \mathbf{a}, \mathbf{b} \in \mathbb{R}^d\}$ where $h_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = 1$ if $a_i \leq x_i \leq b_i$ for all $i = 1, ..., d$ and $h_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = 1$ otherwise.

Problem 3: Rademacher Complexity [30 points]

3.1 Empirical Rademacher Complexity [25 points]

Let $\mathcal{F}$ and $\mathcal{G}$ each denote an arbitrary class of functions.

1. **[2 points]** Prove that for any $c \in \mathbb{R}$,
$$\hat{R}_m(c\mathcal{F}) = |c|\hat{R}_m(\mathcal{F})$$

2. **[3 points]** Prove that if $\mathcal{F} \subset \mathcal{G}$ then
$$\hat{R}_m(\mathcal{F}) \leq \hat{R}_m(\mathcal{G})$$

3. **[10 points]** Prove that
$$\hat{R}_m(\text{Conv}(\mathcal{F})) = \hat{R}_m(\mathcal{F}),$$
where $\text{Conv}(\mathcal{F}) = \left\{ \sum_{i=1}^{n} \lambda_i h_i \,\middle|\, n \in \mathbb{N}, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, h_i \in \mathcal{F} \right\}$ is the convex hull function.

4. **[10 points]** Let X the domain of the samples be defined as $X = \{\mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}||_\infty \leq L\}$, and, $H = \{f : X \to \mathbb{R} : f(\mathbf{x}) = \mathbf{b}^T\mathbf{x}, \mathbf{b} \in \mathbb{R}^d, ||\mathbf{b}||_1 \leq C\}$. Prove that:
$$\hat{R}_m(H) \leq L \times C\sqrt{\frac{2\log 2d}{m}}.$$

   Hint: You may want to use Theorem 4 in the notes: http://www.cs.cmu.edu/~ninamf/courses/806/lect-10-05.pdf and any results of the previous questions.

3.2 Computing the Rademacher Complexity [5 points]

[5 points] Consider 1-dimensional data (each example is a point on the real line). For real-valued a, define the function $h_a(x) = 1$ if $x \leq a$ and $h_a(x) = 0$ otherwise. Let $H = \{h_a\}$. Consider a set S of m distinct examples on the line. What is the empirical Rademacher complexity $\hat{R}_m(H)$ of H on S?

There are several ways to analyze this. If you want, you may use the following interesting fact about gambling. Suppose at each time $t = 1, 2, 3, \ldots$ you bet \$1 on a fair game (with probability 1/2 you win \$1 and with probability 1/2 you lose \$1). After T total rounds, by linearity of expectation, your expected total winnings is \$0. However, if you look back and imagine you had stopped at the best possible time in hindsight (the time $t \leq T$ at which your total winnings were highest), the expected value of your winnings then is $\Theta(\sqrt{T})$; i.e., this is the expected maximum, over all $t \leq T$ of your winnings by time T.