CSE 6740 A/ISyE 6740, Fall 2023, Georgia Tech

CSE 6740: Midterm I

Due Oct 19th, '23 (1:40 pm ET)

Time limit: 70 minutes; Total: 25 points

- Question 1 has 4 parts, questions 2 and 3 have 1 part each. All parts carry equal weight (5 points). Read Question 1 parts in order.
- Solve **any 5 out of 6 parts** (maximum 25 points)
- Maximum possible bonus points is 5.
- Total number of pages is 10; last 3 pages are for additional work – use them if needed.
- Do not consult any material outside of your cheat sheet.
- Do not copy or collaborate with others.
- Return all 10 pages of this booklet.
- Write complete proofs, including all the assumptions made, for full credit.
- All the best!

Question 1: Kernel Ridge Regression (20 points)

Let $\mathbb{H}$ be an RKHS with inner product $\langle \cdot, \cdot \rangle$, induced norm, $\| \cdot \|$ and an associated kernel $k(\cdot, \cdot)$. Consider the kernel ridge regression problem

$$\min_{h \in \mathbb{H}} \hat{R}_S(h) = \min_{h \in \mathbb{H}} \frac{1}{m} \sum_{i=1}^{m} (y_i - h(x_i))^2 + \lambda \|h\|^2, \tag{1}$$

where $\lambda > 0$ is a regularization parameter, $S = \{(x_i, y_i)\}_{1 \le i \le m}$ is a training set with iid samples.

Part 1 (5 points)

Show that the minimizer of the above problem (1) is given by

$$h_S(x) = \sum_{i=1}^{m} ((K + \lambda I)^{-1} Y)_i k(x_i, x), \tag{2}$$

where K is the $m \times m$ Gram matrix, $K_{ij} = k(x_i, x_j)$ and $Y = (y_1, \ldots, y_m)^\top$. You are allowed to use the representer theorem without proof.

Solution:

- Applying representer theorem to the ERM problem in (1), we obtain that the solution $h_S \in \mathbb{H}$ has the form $h_S(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$. Thus, the ERM problem can be written as the following finite-dimensional ERM:

$$\min_{\alpha \in \mathbb{R}^d} \sum_{i=1}^{m} (y_i - \alpha^\top K e_i)^2 + \lambda \alpha^\top K \alpha, \tag{3}$$

  where K is the associated Gram matrix.

- We set the gradient of (3) with respect to α to 0, to obtain, $-KY + K(K + \lambda I)\alpha = 0$, which gives, $\alpha = (K + \lambda I)^{-1} Y$.

Part 2 (5 points)

Let S_i be the training set S without the i-th sample, i.e., $S_i = S \setminus \{(x_i, y_i)\}$. As done in (2), denote by h_S the minimizer of the problem (1) with training set S. Thus, h_{S_i} is the minimizer of the problem (1) with training set S_i. Let S_i' be the set S with the ith element being $(x_i, h_{S_i}(x_i))$ (instead of (x_i, y_i)). Show that $h_{S_i} = h_{S_i'}$.

Solution: The hypothesis makes 0 error on the ith point of S_i' and is defined as the minimizer over the remaining points. Thus, h_{S_i} is also the minimizer over S_i'.
Thus $h_{S_i} = h_{S_i'}$.

Part 3 (5 points)

Define $Y_i = (y_1, \ldots, y_{i-1}, h_{S_i}(x_i), y_{i+1}, \ldots, y_m)^\top = Y - y_i e_i + h_{S_i}(x_i) e_i$, where e_i is the standard basis element with 1 at the ith entry and 0 elsewhere. Show that $h_{S_i}(x_i) = Y_i^\top (K + \lambda I)^{-1} K e_i$.

Solution: Using Part 2 and the definition of KRR hypothesis with respect to the dual variables we have,

$$h_{S_i}(x_i) = h_{S_i'}(x_i) = \alpha_{S_i'}^T K e_i \tag{4}$$

where $\alpha_{S_i'} = (K + \lambda I)^{-1} Y_i$, from Part 1, and the fact that K is symmetric.

$$h_{S_i}(x_i) = ((K + \lambda I)^{-1} Y_i)^T K e_i$$
$$\therefore h_{S_i}(x_i) = Y_i^T (K + \lambda I)^{-1} K e_i$$

Part 4 (5 points)

Define the following leave-one-out error:

$$\hat{R}_S^{\mathrm{l}} = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_{S_i}(x_i))^2. \tag{5}$$

Show that $\hat{R}_S^{\mathrm{l}}$ admits the expression

$$\hat{R}_S^{\mathrm{l}} = \frac{1}{m} \sum_{i=1}^{m} \Big(\frac{h_S(x_i) - y_i}{1 - ((K + \lambda I)^{-1} K)_{ii}} \Big)^2 \tag{6}$$

(Thus, when the diagonal entries of $(K + \lambda I)^{-1} K$ are all 0, $\hat{R}_S^{\mathrm{l}} = \hat{R}_S(h_S)$. That is, the leave-one-out error can be computed by solving just one ERM problem instead of m-many that (5) suggests.)

Solution: From part 3, $h_{S_i}(x_j) = Y_i^\top (K + \lambda I)^{-1} K e_j$. Since $Y_i = Y + (h_{S_i}(x_i) - y_i) e_i$, taking the dot product with $(K + \lambda I)^{-1} K e_i$, we obtain, $h_{S_i}(x_i) = h_S(x_i) + (h_{S_i}(x_i) - y_i) e_i^\top (K + \lambda I)^{-1} K e_i$. Note that $e_i^\top (K + \lambda I)^{-1} K e_i = ((K + \lambda I)^{-1} K)_{ii}$. Substituting this and rearranging, we obtain,

$$h_{S_i}(x_i) - y_i = (h_S(x_i) - y_i) + (h_{S_i}(x_i) - y_i)((K + \lambda I)^{-1} K)_{ii}.$$

Question 2: Rademacher complexity (5 points)

Define a hypothesis class

$$\mathcal{H} = \{h_w(x) = \gamma(w^\top x) : \|w\| \leq \Lambda, \gamma(x) = 1/(1 + e^{-x})\},$$

where $\|\cdot\|$ is the ℓ^2-norm on $\mathbb{R}^d$. This class represents a neuron, for instance. Let $S \subseteq \{x : \|x\| \leq r\}$ be a sample set of n points. Show that the Rademacher complexity

$$\mathrm{Rad}_S(\mathcal{H}) \leq r\Lambda/\sqrt{n}.$$

You are allowed to use the following facts without proof: 1) Lipschitz constant of the sigmoid function γ is 1 and 2) Talagrand's lemma: if Φ is a Lipschitz function with Lipschitz constant l, $\mathrm{Rad}_S(\Phi \circ \mathcal{H}) \leq l \mathrm{Rad}_S(\mathcal{H})$.

Solution: Using the given lemma:

$$\mathrm{Rad}_S(\mathcal{H}) = \mathrm{Rad}_S(\gamma \circ \mathcal{L}) \leq 1.\mathrm{Rad}_S(\mathcal{L}).$$

where

$$\mathcal{L} = \{h_w(x) = w^\top x : \|w\| \leq \Lambda, \}$$

We need to simply calculate the Rademacher complexity of the Linear Hypothesis class:

$$\text{Rad}_S(\mathcal{L}) := \mathbb{E}\left[\sup_{w,\|w\|_2\leq\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i w^T x_i\right|\right] \tag{7}$$

$$= \frac{1}{n}\mathbb{E}\left[\sup_{w,\|w\|_2\leq\Lambda}\left|w^T\sum_{i=1}^{n}\sigma_i x_i\right|\right] \tag{8}$$

$$\leq \frac{1}{n}\mathbb{E}\left[\sup_{w,\|w\|_2\leq\Lambda}\|w\|_2\left\|\sum_{i=1}^{n}\sigma_i x_i\right\|_2\right] \text{ (Cauchy-Schwartz)} \tag{9}$$

$$= \frac{1}{n}\mathbb{E}\left[\Lambda\|\sum_{i=1}^{n}\sigma_i x_i\|_2\right] \tag{10}$$

$$= \frac{\Lambda}{n}\mathbb{E}\left[\sqrt{\left\langle\sum_{i=1}^{n}\sigma_i x_i, \sum_{i=1}^{n}\sigma_i x_i\right\rangle}\right] \tag{11}$$

$$= \frac{\Lambda}{n}\mathbb{E}\left[\sqrt{\sum_{1\leq i,j\leq n}\sigma_i\sigma_j\left\langle x_i, x_j\right\rangle}\right] \tag{12}$$

$$\leq \frac{\Lambda}{n}\sqrt{\mathbb{E}\left[\sum_{1\leq i,j\leq n}\sigma_i\sigma_j\left\langle x_i, x_j\right\rangle\right]} \text{ (Jensen's inequality since square root is a concave function)} \tag{13}$$

$$= \frac{\Lambda}{n}\sqrt{\sum_{1\leq i,j\leq n}\mathbb{E}\left[\sigma_i\sigma_j\right]\left\langle x_i, x_j\right\rangle} \tag{14}$$

$$= \frac{\Lambda}{n}\sqrt{\sum_{1\leq i\leq n}\|x_i\|_2^2} \quad \because \mathbb{E}[\sigma_i\sigma_j] = \mathbb{1}_{i=j} \tag{15}$$

$$\leq \frac{\Lambda r}{\sqrt{n}} \tag{16}$$

$$\tag{17}$$

Question 3: Positive definite kernels (5 points)

Show that $k(x,y) = e^{\left(\sum_{i=1}^{d} \min(|x_i|,|y_i|)\right)}$ is a positive definite kernel on $\mathbb{R}^d \times \mathbb{R}^d$.

Solution: Define $1_a(t) = \begin{cases} 1 & \text{if } t \leq a \\ 0 & \text{else} \end{cases}$. Observe that

$$\begin{aligned} \min(|x_i|, |y_i|) &= \int_0^\infty 1_{|x_i|}(t) 1_{|y_i|}(t) dt \\ &= \langle 1_{|x_i|}, 1_{|y_i|} \rangle \end{aligned}$$

Let $k_1(x_i, y_j) = \min(x_i, y_j)$, where x_i and y_j are real numbers. For any $m \in \mathbb{N}$ and $v \in \mathbb{R}^m$,

$$\sum_{i,j=1}^{m} v_i v_j k_1(x_i, x_j) = \sum_{i,j=1}^{m} v_i v_j \int_0^\infty 1_{x_i}(t) 1_{x_j}(t) dt \tag{18}$$

$$= \int_0^\infty \sum_{i,j=1}^{m} v_i v_j 1_{x_i}(t) 1_{x_j}(t) = \int_0^\infty (\sum_{i=1}^{m} v_i 1_{x_i}(t))^2 \geq 0, \tag{19}$$

where the integral and summation can be exchanged since the integral is finite. This proves that k_1 is PD on $\mathbb{R}\times\mathbb{R}$. Since PDS kernels are closed under addition, $\sum_{i=1}^{d} \min(|x_i|, |y_i|)$ is a PDS kernel. The Taylor expansion of e^t is $\sum_{n=0}^{\infty} \frac{t^n}{n!}$. Since PDS kernels are closed under composition with a power series with positive coefficients, $k(x,y) = e^{\left(\sum_{i=1}^{d} \min(|x_i|,|y_i|)\right)}$ is a positive definite kernel on $\mathbb{R}^d \times \mathbb{R}^d$.