Homework 2
Non-parametric Regression, Deep Density Estimation

CMU 10-716: Advanced Machine Learning (Spring 2020)

OUT: Feb. 25, 2020
DUE: **March 5, 2020, 3:00 PM**.

Instructions:

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Bob explained to me what is asked in Question 4.3"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submission can be written in LaTeX. Upon submission, label each question using the tempate provided by Gradescope.

1 Local Polynomial Regression

In this question we will analyze the bias of local polynomial regression. We will borrow the notation from the lecture notes, so they should be familiar, but we redefine them here for convenience. In 1-D local polynomial regression, given a set of n points $\{(X_i, Y_i)\}$, we let $\hat{m}(x) = \hat{\beta}_{x,0} + \sum_{j=1}^{k} \hat{\beta}_{x,j} x^j$ where $\hat{\beta}_{x,0}, \ldots, \hat{\beta}_{x,k}$ minimize

$$\sum_{i=1}^{n} K\left(\frac{|x - X_i|}{h}\right)\left(Y_i - \beta_0 - \sum_{j=1}^{k} \beta_j X_i^j\right)^2$$

over all $\beta_0, \ldots, \beta_k \in \mathbb{R}$. Let $\Omega \in \mathbb{R}^{n \times n}$ be a diagonal matrix with the ith diagonal entry $K(|x - X_i|/h)$ and let $B \in \mathbb{R}^{n \times (k+1)}$ be defined as $B_{ij} = X_i^{j-1}$.

1. (5 points) Let $b(x) \in \mathbb{R}^{k+1}$ be defined as $b_i(x) = x^{i-1}$ for $i = 1, \ldots, k+1$. Derive an expression for $w(x) \in \mathbb{R}^n$ where $\hat{m}(x) = w(x)^T Y$ in terms of Ω, B and b.

2. (5 points) Show that under the assumption $Y_i = m(X_i) + \epsilon_i$ for some true m and ϵ_i independent mean 0 and variance σ^2 noise, the variance of the estimator can be written as

$$\mathrm{Var}(\hat{m}(x)) = \sigma^2 ||w(x)||^2$$

3. (10 points) Assume that the diagonal matrix Ω is identity. Prove that the bias of the local polynomial regression increases with the degree k.

   Hint: Look up Schur's complement to invert matrices.

2 RKHS

Given a Mercer Kernel $K(\cdot, \cdot)$, denoting the corresponding RKHS by $\mathcal{H}_K$, consider the RKHS regression estimator:

$$\widehat{f} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \tag{1}$$

Letting $K \in \mathbb{R}^{n \times n}$ denote the kernel matrix over the samples, in this problem, you will derive that the problem 1 is equivalent to the simpler, finite dimensional problem of:

$$\widehat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha \tag{2}$$

Recall the following key properties of an RKHS:

- Functions $K(\cdot, x_i) = K_{x_i}$ for $i = 1, ..., n$ are the representers
- $\langle f, K_{x_i} \rangle_{\mathcal{H}_K} = f(x_i)$ for any function $f \in \mathcal{H}_K$

1. (5 points) Let $f = \sum_{i=1}^{n} \alpha_i K_{x_i}$ and define $\tilde{f} = f + \rho$, where $\rho \in \mathcal{H}_K$ is any function orthogonal to K_{x_i} for $i = 1, ..., n$. Prove that $\tilde{f}(x_i) = f(x_i)$ for $i = 1, ..., n$, and $\|\tilde{f}\|_{\mathcal{H}_K} \geq \|f\|_{\mathcal{H}_K}$.

2. (10 points) Using part (a), show that in the infinite-dimensional problem (1), we only need to consider functions of the form $f = \sum_{i=1}^{n} \alpha_i K_{x_i}$, and this reduces problem (1) to problem (2).

3 Deep Density Estimation

3.1 Variational Inference (10 Points)

Recall the setup of variational auto-encoders from the lecture notes, where we have a standard Gaussian latent random vector $Z \sim N(0, I)$, and we parameterize the distribution of X given Z as $p_\theta(X|Z)$ which in turn specifies a marginal distribution $p_\theta(X)$. Since the resulting log-likelihood is intractable, we wish to approximate it using the so-called ELBO variational objective $L(\theta, \phi, x)$ (which uses a separate parametric distribution $q_\phi(Z|X)$) defined as

$$L(\theta, \phi, x) = -\mathrm{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right].$$

It was claimed in the lecture notes that:

$$\log p_\theta(x) - L(\theta, \phi, x) = \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)).$$

Prove this claim.

3.2 GANs (10 Points)

Consider a density p over $\mathcal{X}$ and a parameterized density q_θ we would like to train to be close to p. Let $D : \mathcal{X} \to [0, 1]$ be the discriminator of our GAN. Recall we define $V(p, q_\theta, D)$ to be the following:

$$V(p, q_\theta, D) = \mathbb{E}_{x\sim p}\left[\log D(x)\right] + \mathbb{E}_{x\sim q_\theta}\left[\log(1 - D(x))\right]$$

1. (5 Points) For fixed known p and q_θ, derive the optimal discriminator D that maximizes $V(p, q_\theta, D)$.

2. (5 Points) Show that

$$\max_D V(p, q_\theta, D) = -\log(4) + 2\mathrm{JSD}(p, q_\theta)$$

   What is the value of $\min_\theta \max_D V(p, q_\theta, D)$?