In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 04 Solution

Spring Semester 1401-02

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

1 Representer Theorem

1.1

In your own words, explain the meaning of each term below **(For the sake of completeness, the solution here has more details than required of your answers. It's okay if your answers weren't as technical as this.)**:

- Hilbert Space: A function space $\mathcal{H}$ with an inner product operator is called a Hilbert space. (We aso require that the function space be complete, which means that every Cauchy sequence of functions $f_i \in \mathcal{H}$ has a limit that is also in H.)
- Reproducing Kernel Hilbert Space: Let $\mathcal{H}$ be the space $L^2(\mathcal{X})$ (square integrable) of functions $\mathcal{X} \to \mathbb{R}$. For each $x \in \mathcal{X}$, let $\delta_x : \mathcal{H} \to \mathbb{R}$ be the evaluation functional that takes as input a function $f \in \mathcal{H}$ and evaluates it at x, i.e., $\delta_x(f) = f(x)$. We say that a Hilbert space $\mathcal{H}$ is a Reproducing Kernel Hilbert Space or RKHS if for each δ_x there is a unique element $\mathcal{K}_x \in \mathcal{H}$ such that for every $f \in \mathcal{H}$

$$\delta_x f = f(x) = \langle f, \mathcal{K}_x \rangle$$

- Reproducing Kernel: Now consider the function $f = \mathcal{K}_y \in \mathcal{H}$ evaluated at x. We have

$$\delta_x \mathcal{K}_y = \mathcal{K}_y(x) = \langle \mathcal{K}_y, \mathcal{K}_x \rangle \triangleq \mathcal{K}(x, y)$$

  This is called the reproducing property and $\mathcal{K}(x, y)$ is called the RKHS kernel.

- Mercer's Theorem: The reproducing kernel above is positive semidefinite, in the sense that it can be represented as:

$$\mathcal{K}(x, y) = \sum_{j \geq 1} \lambda_j \psi_j(x) \psi_j(y)$$

  where λ_j is a countable sequence of non-negative numbers decreasing to 0 , and the ψ_j are orthonormal functions in $L^2(\mathcal{X})$. This result is known as Mercer's theorem.

1.2

Theorem 1.1. *(Representer Theorem). Consider The following optimization problem:*

$$f^* = \arg\min_{f \in \mathcal{H}_\mathcal{K}} \mathcal{L}(f)$$

$$\mathcal{L}_\mathcal{K} = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(x_i)) + R(\|f\|)$$

where $\mathcal{H}_\mathcal{K}$ is an RKHS with kernel $\mathcal{K}$, $\ell(y, \hat{y}) \in R$ is a loss function, $R(c) \in R$ is a strictly monotonicly increasing penalty function, and

$$\|f\|_\mathcal{H} = \sqrt{\langle f, f \rangle_\mathcal{H}}$$

is the norm of the function. Then we have

$$f^*(x) = \sum_{k=1}^{N} \alpha_k \mathcal{K}(x, x_k)$$

where $\alpha_k \in R$ are some coefficients that depend on the training data $\{(x_i, y_i)\}$

Proof. Let us define the feature map $\Phi(x) = \mathcal{K}(\cdot, x)$. Given any $x_1, \ldots, x_N$, we can use orthogonal projection to decompose any $f \in \mathcal{H}_\mathcal{K}$ into a sum of two functions, one lying in the span of $\{\Phi(x_1), \ldots, \Phi(x_N)\}$, and the other lying in the orthogonal complement; let this latter function be denoted by $v(x)$. Thus

$$f = \sum_{i=1}^{N} \alpha_i \Phi(x_i) + v$$

where $\langle v, \Phi(x_i) \rangle = 0$ for all i. Using this result, together with the reproducing property, we have

$$f(x_j) = \left\langle \sum_{i=1}^{N} \alpha_i \Phi(x_i) + v, \Phi(x_j) \right\rangle = \sum_{i=1}^{N} \alpha_i \langle \Phi(x_i), \Phi(x_j) \rangle$$

Since f is independent of v, we see that the first term in $\mathcal{L}(f)$ is independent of v. Now let us consider the second (regularization) term in $\mathcal{L}(f)$. Since v is orthogonal to $\sum_{i=1}^{N} \alpha_i \Phi(x_i)$ and R is strictly monotonic, we have

$$\begin{aligned} R(\|f\|) &= R\left(\left\| \sum_{i=1}^{N} \alpha_i \Phi(x_i) + v \right\| \right) \\ &= R\left(\sqrt{\left\| \sum_{i=1}^{N} \alpha_i \Phi(x_i) \right\|^2 + \|v\|^2} \right) \\ &\geq R\left(\left\| \sum_{i=1}^{N} \alpha_i \Phi(x_i) \right\| \right) \end{aligned}$$

Thus we can minimize the second term of $\mathcal{L}(f)$ by setting $v = 0$. Hence we can minimize the overall regularized loss by using a solution of the form

$$f^*(\cdot) = \sum_{i=1}^{N} \alpha_i \Phi(x_i) = \sum_{i=1}^{N} \alpha_i \mathcal{K}(\cdot, x_i)$$

□

1.3

How does the representor theorem solution compare to the final SVM solution?
answer
From the representer theorem we know that to find the optimal solution for f, it suffices to find the optimal coefficients α_i's. We also observe that the SVM optimization problem can be formulated in the framework of the representer theorem by setting $R(\|f\|) = \frac{1}{2}\|w\|_2^2$ and $\mathcal{L}_\mathcal{K} = C \sum \xi_i$.
This allows for the application of the representer theorem, which reduces the problem to optimizing over coefficients α_i. This agrees with the solution to the dual form of the SVM optimization problem where the optimal function is represented by a linear combination of kernels. Note that the SVM optimization problem puts constraints on these coefficients, which might complicate the problem in the general case but representer theorem works on SVM nevertheless.

2 Neural Networks Can be Seen as (almost) GPs!

In this problem, we explore an interesting property of Gaussian processs:

2.1

Consider an MLP with one hidden layer and activation functions $h_j(x), j \in 1, 2, ..., H$:

$$f_k(x) = b_k + \sum_{j=1}^{H} v_{jk} h_j(x)$$
$$h_j(x) = h(u_{0j} + x^T u_j)$$

where H is the number of hidden units, and $h()$ is some nonlinear activation function, such as the ReLU. Assume Gaussian prior on the parameters (each set of parameters below are independent from the other sets):

$$b_k \sim \mathcal{N}(0, \sigma_b), v_{jk} \sim \mathcal{N}(0, \sigma_v), b_k \sim \mathcal{N}(0, \sigma_b), u_j \sim \mathcal{N}(0, \Sigma),$$

Denote all the parameters by $\boldsymbol{\theta}$.

Proof. The expected output from unit k when applied to one input vector is given by

$$\mathbb{E}_{\boldsymbol{\theta}}\left[f_k(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[b_k + \sum_{j=1}^{H} v_{jk} h_j(\boldsymbol{x})\right] = \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[b_k\right]}_{=0} + \sum_{j=1}^{H} \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[v_{jk}\right]}_{=0} \mathbb{E}_{\boldsymbol{u}}\left[h_j(\boldsymbol{x})\right] = 0$$

The covariance in the output for unit k when the function is applied to two different inputs is given by the following:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[f_k(\boldsymbol{x}) f_k\left(\boldsymbol{x}'\right)\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\left(b_k + \sum_{j=1}^{H} v_{jk} h_j(\boldsymbol{x})\right)\left(b_k + \sum_{j=1}^{H} v_{jk} h_j(\boldsymbol{x})\right)\right] \\ &= \sigma_b^2 + \sum_{j=1}^{H} \mathbb{E}_{\boldsymbol{\theta}}\left[v_{jk}^2\right] \mathbb{E}_{\boldsymbol{u}}\left[h_j(\boldsymbol{x}) h_j\left(\boldsymbol{x}'\right)\right] = \sigma_b^2 + \sigma_v^2 H \mathbb{E}_{\boldsymbol{u}}\left[h_j(\boldsymbol{x}) h_j\left(\boldsymbol{x}'\right)\right]\end{aligned}$$

Now consider the limit $H \to \infty$. We scale the magnitude of the output by defining $\sigma_v^2 = \omega / H$. Since the input to k' th output unit is an infinite sum of random variables (from the hidden units $h_j(\boldsymbol{x})$), we can use the central limit theorem to conclude that the output converges to a Gaussian with mean and variance given by

$$\mathbb{E}\left[f_k(\boldsymbol{x})\right] = 0, \mathbb{V}\left[f_k(\boldsymbol{x})\right] = \sigma_b^2 + \omega \mathbb{E}_{\boldsymbol{u}}\left[h(\boldsymbol{x})^2\right]$$

Furthermore, the joint distribution over $\left\{f_k\left(\boldsymbol{x}_n\right) : n = 1 : N\right\}$ for any $N \geq 2$ converges to a multivariate Gaussian with covariance given by

$$\mathbb{E}\left[f_k(\boldsymbol{x}) f_k\left(\boldsymbol{x}'\right)\right] = \sigma_b^2 + \omega \mathbb{E}_{\boldsymbol{u}}\left[h(\boldsymbol{x}) h\left(\boldsymbol{x}'\right)\right] \triangleq \mathcal{K}\left(\boldsymbol{x}, \boldsymbol{x}'\right)$$

□

3 SVM

3.1

In the soft-margin SVM problem, the slack value ξ_i takes three possible values for the ith sample: $(\xi_i = 0, 0 < \xi \leq 1, 1 \leq \xi)$. For each of these scenarios, where does the point lie relative to the margin? Is the point classified correctly?

3.2

Each of the datasets below contains points beloning to two classes {-1,1} (positives and negatives correspond to 1 and -1). For each dataset, find a transformation of the features X_1 and X_2 such that the points are linearly separable (can be separated by a line in the new expanded feature space).

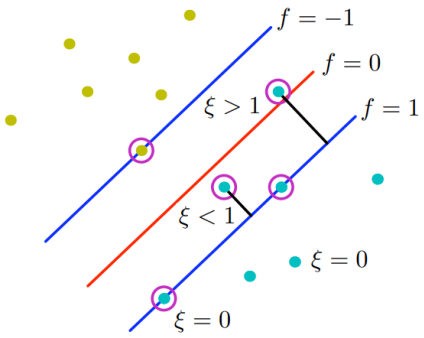
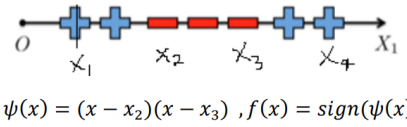
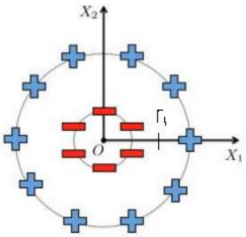


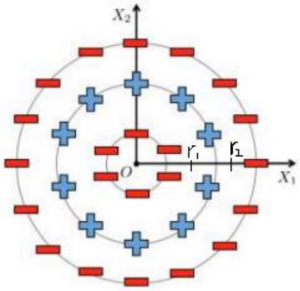
Figure 1: (From *Pattern Recognition and Machine Learning* by Chris Bishop.



$$\psi(x) = (x - x_2)(x - x_3)\ , f(x) = sign(\psi(x))$$



$$\psi(x) = \sqrt{x_1^2 + x_2^2} - r_1, \qquad f(x) = sign(\varphi(x))$$



$$\psi(x) = \left(\sqrt{x_1^2 + x_2^2} - r_1\right)\left(\sqrt{x_1^2 + x_2^2} - r_2\right), \qquad f(x) = sign(-\varphi(x))$$

Figure 2: Classifiers as functions of augmented features ϕ.