In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 04

Spring Semester 1401-02

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

Deadline: To be announced
Late submisstion: To be announced

(∗) starred problems are optional and have a bonus mark!

1 Representer Theorem

1.1

In your own words, explain the meaning of each term below (you don't need to get too technical here. The aim is to ensure that you have enough knowledge to answer the next part. Don't freak out!):

- Hilbert Space
- Reproducing Kernel Hilbert Space
- Reproducing Kernel
- 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 $\mathcal{L}_\mathcal{K}$ 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 $\{(x_i, y_i)\}$

Important: For each of the questions below, explain and motivate your answers!

1.2.1

Use the Mercer's theorem to write $f \in \mathcal{H}_{\mathcal{K}}$ in the following form:

$$f(x_j) = \sum_{k=1}^{N} \alpha_k \Phi(x_k) + v(x_j)$$

where $\Phi_k(.)$ and $v(.)$ are orthogonal i.e. $\langle v(.), \Phi(.) \rangle = 0$

1.2.2

Use the reproducing kernel property to write f as a dot product of Φ_k's.

1.3

Find a lower bound on the regularization term $R(\|f\|)$ using the orthoganality of Φ_k and v and the monotonocity of R.

1.4

Now jointly optimize both the loss terms and the penalty function with respect to v and prove the representor's theorem.

1.5

How does the representor theorem solution compare to the final SVM solution?

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 $\boldsymbol{\theta}$.

2.1.1

Show that the expected output of the network is 0, i.e. $\mathbf{E}_{\boldsymbol{\theta}}[f_k(x)] = 0$.

2.1.2

Show that the covariance of the output for two different inputs is the following:

$$\mathbf{E}_{\boldsymbol{\theta}}[f_k(x) f_k(x')] = \sigma_b^2 + \sigma_v^2 H \mathbf{E}_u[h_j(x) h_j(x')]$$

.

2.1.3

Using the central limit theorem, argue that as $H \to \infty$, the output of the network converges to a multivariate Gaussian distribution with mean and covariance calculated above. This is equivalent to a Gaussian process (this kernel can be computed in close form for certain activation functions such as the ReLU.)

2.2 (*)

The result above can also be extended to arbitrary deep neural networks. Search "Neural tangent kernels" and in two paragraphs, explain your understanding of what they are (Note: it's not necessarily practical to always use GPs instead of neural networks).

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).

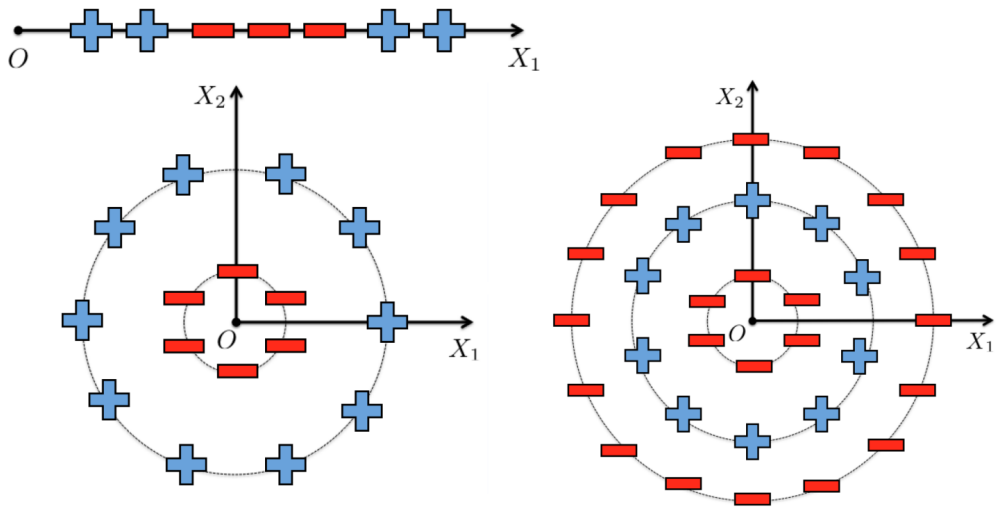


Figure 1: Datasets

4 Interpretation Via Maximum Projection Spread

In this point of view we need some directions that have maximum data projection variance.

Derive sample covariance matrix and show that the linear projection onto an M -dimensional subspace that maximizes the variance of the projected data is defined by the M eigenvectors of the sample covariance matrix S, corresponding to the M largest eigenvalues.

5 Interpretation Via Reconstruction

Prove the following statement:

$$||x_i - \sum_{j=1}^{K} z_{ij} v_j||^2 = x_i^T x_i - \sum_{j=1}^{K} v_j^T x_i x_i^T v_j$$

6 Whitening Using PCA

Whitening is one of the pre-processing techniques which is used in practical ML. By whitening, we mean that for a given feature matrix X, this feature matrix should have zero mean vector and identity matrix as its covariance matrix. Explain that how we could transform X by using its covariance matrix principal components in order to have a whitened data set? After that, prove this new feature matrix has the desired properties, namely, zero mean vector and identity matrix as its covariance matrix.

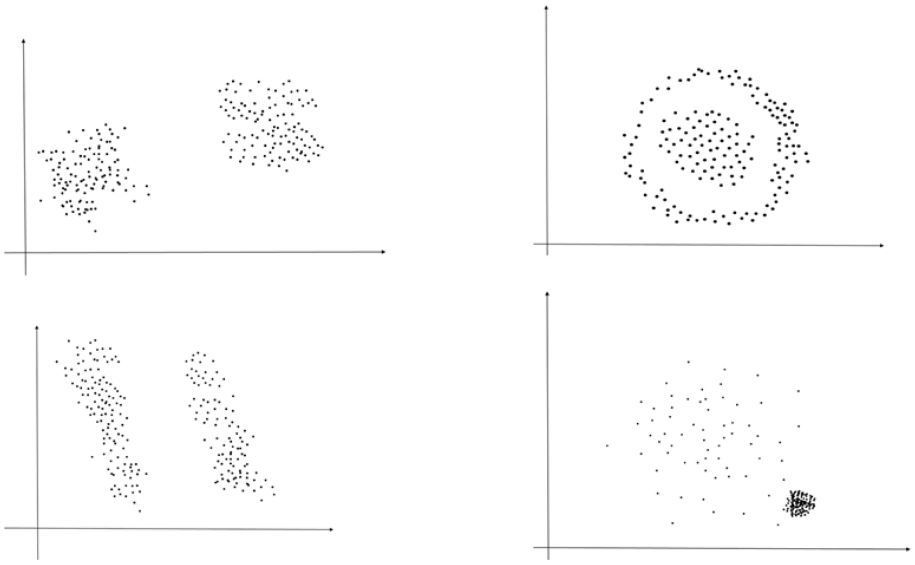
7 *Principal Components with Missing Values

Suppose that some of the observations xij of the data matrix X are missing. Suggest an algorithm that can both impute the missing values and find the principal component at the same time

8 Clustering

8.1

In each sample below, draw the boundry that K-means finds for$K = 2$. Do you think the clusters separated by borders found by K means is meaningfull in each case? If not, what property of data causes this?(*Recommend some solutions for these problems)



8.2

Is it important to choose initiall points carfully in $Kmeans$ clustering?
Illustrate it with some examples.(You can use examples above)

8.3 (*)

Now you have found that initializing the points randomly is not always good.Because of that we should assign initiall points more carefully.
Explain $Kmeans++$ algorithm and define WCSS and elbow method.