In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 04

Spring Semester 1402-03

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. R.Amiri

Due on Khordad 18th, 1403 at 23:59

(∗) starred problems are just optional!

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).

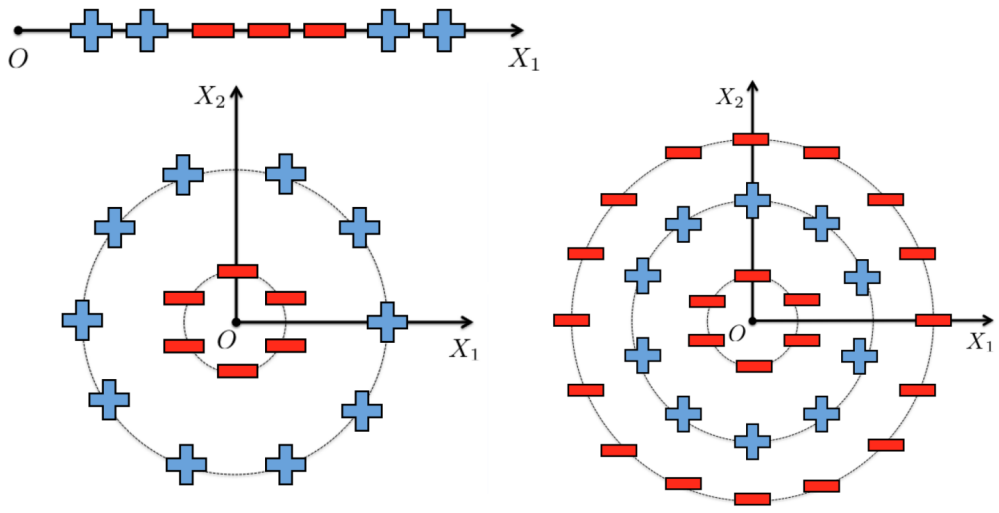
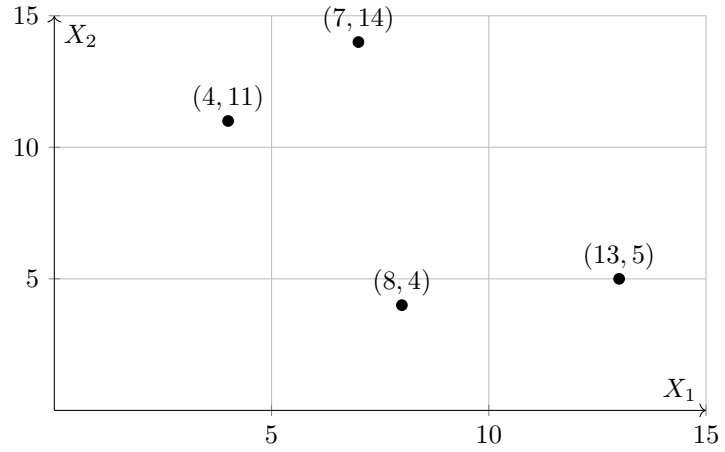


Figure 1: Datasets

4 Dimensionality Reduction using PCA

For the following data points, find the first principal component and then project the data points onto it.



5 Reconstruction Error

We want to perform PCA. Each sample $x_i \in \mathbb{R}^p$ is projected onto the new coordinate system using $z_i = V_{1:k}^T x_i$. Here, $V_{1:k}$ is the matrix of the first k principal components ($V_{1:k} = [v_1|v_2|...|v_k]$). We can reconstruct x_i from z_i using the equation $\hat{x}_i = V_{1:k} z_i$.

5.1

Prove:

$$\|\hat{x}_i - \hat{x}_j\|_2 = \|z_i - z_j\|_2$$

5.2

Prove:

$$\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = (n-1) \sum_{i=k+1}^{p} \lambda_i$$

What can be inferred from this equation regarding the reconstruction error?

6 $(*)$ Clustering

6.1

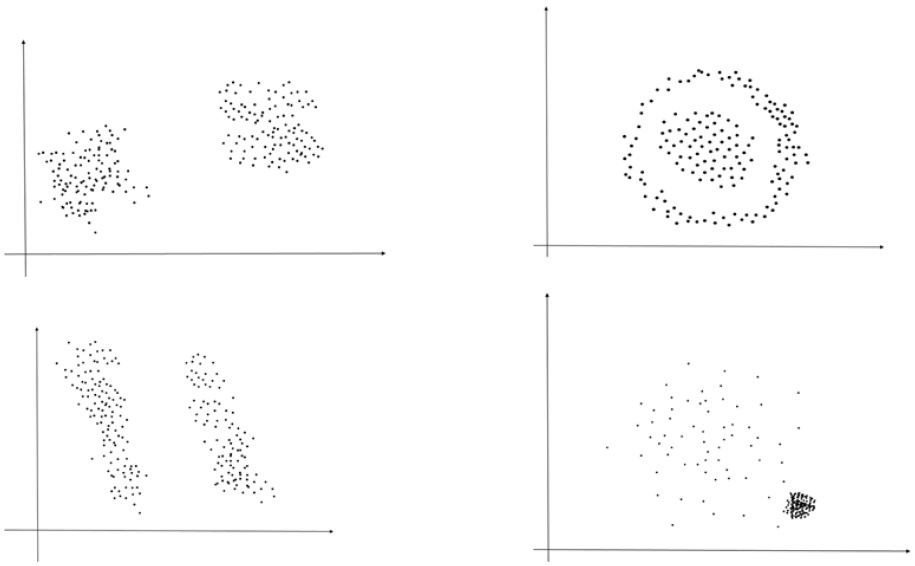
In each sample below, draw the boundry that K-means finds for$K = 2$. Do you think the clusters separated by borders found by K means is meaningfull in each case? If not, what property of data causes this?(*Recommend some solutions for these problems)

6.2

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

6.3

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.



7 Mixture Models

In this question we get introduced to mixture models.

7.1 Introduction to Mixture Models

In your own words, explain how the MM algorithm can deal with non convex optimization objective functions by considering simpler convex objective functions.

7.2 Mixture Models for specific distribution

You are given a data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}, \mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}, i = 1, \ldots, N$. The data points accumulate on m different lines, $\mathbf{a}_j^T \mathbf{x}_i = y_i$, for $\mathbf{a}_j \in \mathbb{R}^d, j = 1, \ldots, m$.

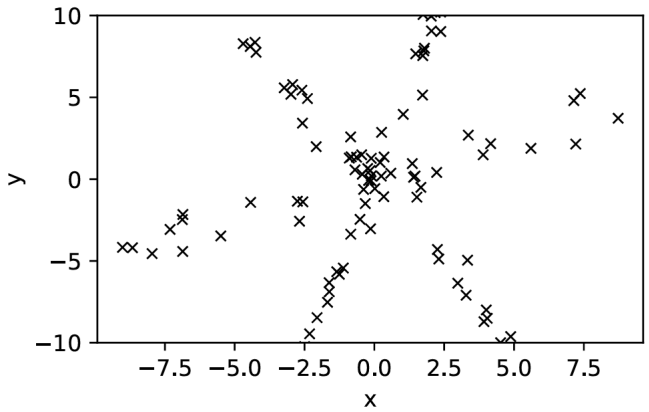


Figure 2: d=1 and m=3

Then

$$p(\mathbf{x}, y \mid \theta) = \sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\mathbf{a}_j^T \mathbf{x} - y\right)^2}{2\sigma^2}\right),$$

where $\theta = (\pi_{1:m}, \mathbf{a}_{1:m}), \sum \pi_j = 1, \pi_j \geq 0$ and $\sigma > 0$ is given and fixed.

- Find the responsibilities in the E-step of Soft EM,

$$r_{nk}^{(t)} = p\left(z_n = k \mid \mathbf{x}_n, y_n, \theta^{(t-1)}\right).$$

- Write down the class predictions $z_n^{(t)}$ for $(\mathbf{x}_n, z_n)$ in the E-step of Hard EM in terms of $r_{nk}^{(t)}$.
- Assume that we observe the true labels $z_1, \ldots, z_\ell$ for the first ℓ datapoints, $\ell < N$. How can we modify the E-step of Soft EM to incorporate the additional information?

8 EM Algorithm for GMM and CMM

In this part, we want you to apply EM algorithm to learn (estimate) parameters for two different mixture model and find closed-form solution of their parameters.

8.1 EM for Gaussian Mixture Model

Compute estimate of parameters for Gaussian Mixture Models for N observed data $\{\mathbf{x}_i\}_{i=1}^N$.

- Determine model parameters and initialize them.
- Compute complete dataset likelihood[1].
- Find closed-form solution for parameters using EM algorithm.

8.2 EM for Categorical Mixture Model

Compute estimate of parameters for Categorical Mixture Models for N observed data $\{\mathbf{x}_i\}_{i=1}^N$.

- Determine model parameters and initialize them.
- Compute complete dataset likelihood.
- Find closed-form solution for parameters using EM algorithm.

9 Advanced Hard-Margin SVM with Dual Problem and Kernel Methods

9.1 Primal Formulation and Geometric Interpretation

Define the primal optimization problem for a hard-margin SVM. Clearly state the objective function and the constraints. Provide a geometric interpretation of the margin and explain why maximizing the margin is important in the context of classification.

9.2 Derive the Dual Problem from the Primal Formulation

Starting from the primal formulation, derive the dual optimization problem for the hard-margin SVM. Introduce Lagrange multipliers and formulate the Lagrangian. Show detailed steps to obtain the dual problem by minimizing the Lagrangian with respect to the primal variables $\mathbf{w}$ and b. Ensure to explain the mathematical properties and assumptions used in the derivation.

9.3 KKT Conditions and Support Vectors

State the Karush-Kuhn-Tucker (KKT) conditions for the hard-margin SVM. Use these conditions to explain the significance of support vectors in the context of the dual problem.

[1] $\mathrm{p}(\mathrm{D}\ ;\ \boldsymbol{\theta}) = \mathrm{p}\Big(\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{z}_i\}_{i=1}^N ; \boldsymbol{\theta}\Big)$

9.4 Kernel Trick and Dual Problem Reformulation

Using the kernel trick, replace the inner product in the dual problem with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. Derive the dual optimization problem for a hard-margin SVM with a specific kernel, such as the Gaussian RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$. Discuss the computational benefits of using the kernel trick and how it allows handling non-linear separable data.

9.5 Solve the Dual Problem and Interpret the Results

Solve the dual problem and determine the support vectors. Explain how the support vectors are used to construct the decision boundary in both the original and transformed feature spaces. Provide a geometric interpretation of the decision boundary in the context of the kernelized SVM.

10 Advanced Soft-Margin SVM with Regularization and Kernel Methods

10.1 Primal Formulation with Regularization and Slack Variables

Define the primal optimization problem for a soft-margin SVM, incorporating regularization and slack variables.

10.2 Derive the Dual Problem with Regularization

Derive the dual form of the optimization problem for the soft-margin SVM. Introduce Lagrange multipliers for both the margin constraints and the slack variables. Formulate the Lagrangian and show the detailed steps to obtain the dual problem by minimizing the Lagrangian with respect to the primal variables.

10.3 Real-World Application and Parameter Selection

Discuss how a soft-margin SVM can be applied to a real-world classification problem, such as spam email detection. Describe the process of selecting the regularization parameter C and kernel parameters through cross-validation. Explain the impact of these parameters on the model's performance and generalization ability.

10.4 Kernelized Soft-Margin SVM with Polynomial Kernel

Formulate the dual problem for a kernelized soft-margin SVM using the polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$. Derive the necessary mathematical expressions and constraints.

10.5 Model Evaluation and Interpretation

Explain how to evaluate the performance of a soft-margin SVM model on a test dataset. Discuss metrics such as accuracy, precision, recall, and F1-score. Provide an interpretation of the model's decision boundary and the influence of support vectors in the context of the chosen kernel and regularization parameters.