University of TEHRAN

Machine Learning

Problem Set 6

Kaveh Kavousi
Hesam Montazeri
Fahimeh Palizban
Zohreh Toghraee

Aban 14, 1398
(November 5, 2019)

Problem 1: Review Questions
Write a summary of the lectures of this week. Write down all formulas we discussed in the lectures and explain in detail each step of derivations. As a guideline, you may consider the following topics:

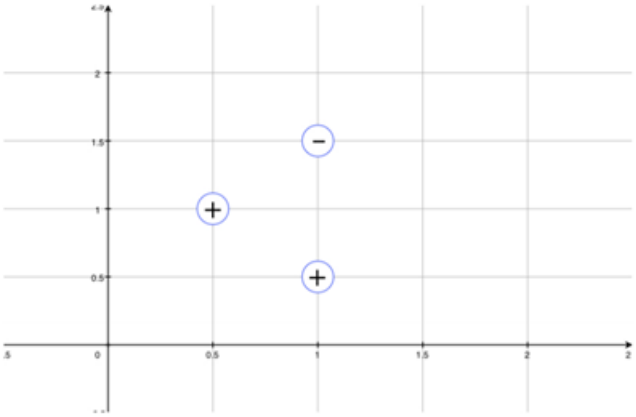
(a) VC dimension; margin concept

(b) Dual Lagrange problem with inequality constraints; KKT conditions

Problem 2: KKT conditions
Using the Karush-Kuhn-Tucker (KKT) conditions, solve the following nonlinear programming problem:

$$\begin{aligned} &\max f(x,y) = xy \\ &\text{s.t.} \quad x + y^2 \leqslant 2 \\ &\qquad x, y \geqslant 0 \end{aligned}$$

Problem 3: Linear SVM
In the following figure, find the optimal linear SVM for separating the classes:



Problem 4: Programming: Liquid Biopsy for Cancer Diagnosis
Early detection of cancer types is a major challenge in cancer diagnosis. Cancer is primarily diagnosed

by clinical presentation, radiology, biochemical tests, pathological analysis and molecular profiling of tumor tissue. The emerging non-invasive blood-based *liquid biopsies* provides a promising alternative diagnostic tool to cancer care. In this exercise, you will investigate whether gene expression profiles of *tumor-educated blood platelets* (TEPs) can be used for identification of six cancer types namely breast, hepatobiliary, colorectal cancer, glioblastoma, pancreatic, and non-small cell lung cancer as well as healthy samples. Your tasks are given below:

(a) Explore the data.

(b) **Two-class classification:** the second task is to determine the presence of cancer, of any type, in the input sample according to the gene expression profile. Assume the response variable $Y = 1$ denotes a cancer sample and $Y = 0$ indicates a healthy donor. Explore the applications of logistic regression, LDA, QDA, SVM (linear kernel and at least two non-linear kernels), KNN on this problem.

(c) **Multi-class classification:** finally you need to perform a multi-class classification where the response variable Y can take seven categories namely six cancer types and the *healthy donor* category. Explore logistic regression, softmax regression, LDA, QDA, SVM (linear kernel and at least two non-linear kernels), KNN for multi-class problem. If needed use one-versus-one and one-versus-all approaches.

Provide a sound statistical analysis. If necessary, avoid the overfitting by performing feature selection or regularization. Assess the performance of your method by appropriate measures.

We encourage discussing the problems with other students, however, similarity between solutions is not allowed. (**Important**) Studying any online solution, no matter to what extent, is strictly forbidden and is considered as a violation of the academic honor code. Please write in the first page of your submission whom you have brainstormed the questions. Submit your solutions (using Easyclass) by Aban 18 for the review part and Aban 25, 1398 for the remaining problems..