Introduction to Machine Learning (WISB 365)
Final Exam

Sjoerd Dirksen

31 January 2023, 17:00-20:00

This exam consists of 4 questions, worth 45 points in total, and a bonus question. Please write your name and student number on every sheet of your solutions.

Question 1 [14 points]

Consider a training dataset $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathbb{R}^d \times \{-1, 1\}$. To find a linear classifier, we consider the optimization problem

$$\min_{w \in \mathbb{R}^d,\ b \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} 1_{\{\operatorname{sign}(\langle w, x_i \rangle + b) \neq y_i\}}, \tag{1}$$

(a) Explain how (1) is related to the problem of finding a linear classifier that has the highest probability of classifying a new datum correctly.

(b) Let $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, let A be the matrix containing $y_j x_j^T$, $j = 1, \ldots, m$, as its rows and let $B = [A\ y] \in \mathbb{R}^{m \times (d+1)}$ be the matrix obtained by appending y as an additional column to A. Consider the optimization problem

$$\max_{w \in \mathbb{R}^d,\ b \in \mathbb{R},\ z \in \mathbb{R}^m} \min\{z_1, \ldots, z_m\} \qquad \text{s.t.} \qquad B \begin{pmatrix} w \\ b \end{pmatrix} - z = 0 \tag{2}$$

and assume that a solution exists. Show that by solving (2) we find either a solution of (1) or can conclude that S is not linearly separable. Does (2) have a solution if all labels in the training dataset are identical?

Question 2 [12 points]

Let $x_1, \ldots, x_m \in \mathbb{R}^d$ and consider the k-means clustering problem

$$\min_{(C_1, \ldots, C_k) \in \mathcal{C}_{k,m}} \min_{\mu_1, \ldots, \mu_k \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{j \in C_i} \|x_j - \mu_i\|_2^2, \tag{3}$$

where $\mathcal{C}_{k,m}$ is the collection of all k-clusterings.

(a) Show that (3) is equivalent to the problem

$$\min_{(C_1, \ldots, C_k) \in \mathcal{C}_{k,m}} \sum_{i=1}^{k} \sum_{j \in C_i} \left\| x_j - \frac{1}{|C_i|} \sum_{\ell \in C_i} x_\ell \right\|_2^2.$$

(b) Show that (3) is equivalent to the problem

$$\min_{(C_1, \ldots, C_k) \in \mathcal{C}_{k,m}} \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{j \in C_i} \sum_{\ell \in C_i} [\langle x_j, x_j \rangle - \langle x_j, x_\ell \rangle]$$

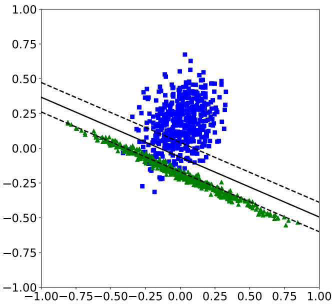
(c) Consider the problem

$$\min_{(C_1,\ldots,C_k)\in\mathcal{C}_{k,m}} \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{j\in C_i} \sum_{\ell\in C_i} [(1+\langle x_j, x_j\rangle)^2 - (1+\langle x_j, x_\ell\rangle)^2].$$

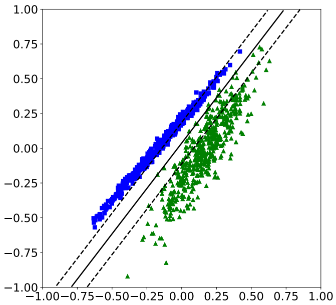
Show that this problem is equivalent to using k-means clustering after transforming the data $x_1, \ldots, x_m$ using a certain feature map.

Question 3 [9 points]

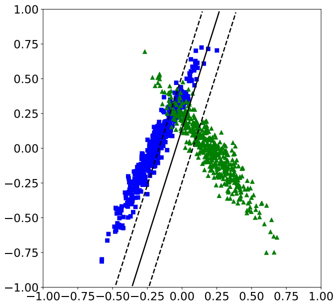
The three pictures (a),(b),(c) below display the output of soft SVM trained on three different 2D datasets. Training samples with label $+1$ are marked by green triangles (▲) while training samples with label -1 are marked by blue squares (■). Match each classifier to the correct ROC curve and explain your reasoning carefully.



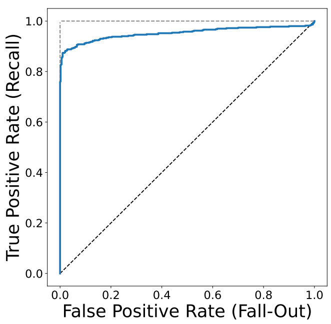
(a)



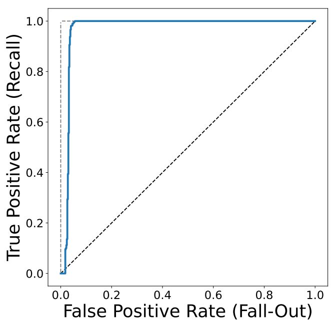
(b)



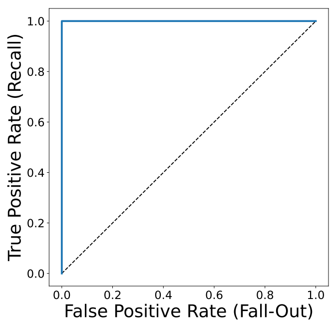
(c)



(1)



(2)



(3)

Question 4 [10 points]

In the lectures we discussed that the ℓ_1-norm is often used as a convex 'substitute' for the sparsity of a vector, e.g., in sparse linear regression. To motivate this, show that the ℓ_1-norm is the largest convex underestimator of the sparsity. Specifically, show that for any convex function $f : [0,1]^d \to \mathbb{R}$ satisfying

$$f(x) \le \|x\|_0, \qquad \text{for all } x \in [0,1]^d,$$

we must have

$$f(x) \le \|x\|_1, \qquad \text{for all } x \in [0,1]^d.$$

What is the largest error that one can make by using $\|x\|_1$ as a proxy for the sparsity of x if $x \in [0,1]^d$?

Hint: You may use (without proof) Jensen's inequality, which states that for any convex function g, $g(\sum_{i=1}^{p} \lambda_i x_i) \leq \sum_{i=1}^{p} \lambda_i g(x_i)$ for any $p \in \mathbb{N}$, any scalars $\lambda_i \geq 0$, $1 \leq i \leq p$, satisfying $\sum_{i=1}^{p} \lambda_i = 1$, and any x_i in the domain of g, $1 \leq i \leq p$.

Bonus Question [4 points]

On 29-01-2023, the following conversation took place:

- S.D.: *Write a confusing explanation of a confusion matrix with a mistake in it.*
- ChatGPT: *A confusion matrix is a table that is used to define the performance of a classification algorithm. It is also known as an error matrix or a match matrix. The basic layout of a confusion matrix is a two-dimensional grid where the rows represent the predicted class and the columns represent the actual class. Each cell in the grid represents the number of observations that were predicted to be in a certain class but are actually in a different class. For example, if a confusion matrix is used to evaluate a binary classification algorithm, it will have two rows and two columns. The top-left cell will contain the number of true positives, which are the observations that were correctly predicted to be in the positive class. The bottom-right cell will contain the number of true negatives, which are the observations that were correctly predicted to be in the negative class.*

  The mistake in it: The top-left cell contains the number of False positive and the bottom-right cell contains the number of False negative.

Identify the mistakes in the reply.