Assignment 4 (Sol.)

Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following are convex functions?

   (a) $f(x) = (\sum_{i=1}^{n} x_i^p)^{1/p}$, where $x \in R^n$ and $p \geq 0$

   (b) $f(x) = \log(\sum_{i=1}^{n} \exp x_i)$ where $x_i \in R^n$

   (c) $f(x) = \sum_{i=1}^{n} \sin(x_i)$, where $x \in R^n$

   (d) $f(x) = \sum_{i=1}^{n} x_i \log x_i$, where $x \in R^n$

   Solution - (b), (c)
   a - This would hold true for $p \geq 1$.
   b - Log-Sum-Exponential function can be proven to be convex by showing that the hessian is psd.
   d - You can use the hessian to prove convexity. Note that the domain restricts the values x_i can take, hence making it obviously a psd

2. We discussed two approaches to classification, one that learns the discriminant functions, and the other that is based on modeling hyperplanes. Which of these approaches is more suitable for multi-class problems and why?

   (a) Discriminant functions; because they allow for a probabilistic interpretation of the predictions.

   (b) Hyperplane methods; because they allow for a probabilistic interpretation of the predictions.

   (c) Discriminant functions; because an appropriate set of functions will allow us to efficiently disambiguate class predictions.

   (d) Hyperplane methods; because we can use basis expansion to transform the input to a space where class-boundaries are linear.

   Solution - c

3. Consider the following optimization problem

$$\begin{array}{lrcl} \min & x^2 + 1 & & \\ \text{s.t.} & (x-2)(x-4) & \leq & 0 \end{array}$$

   Select the correct options regarding this optimization problem.

   (a) Strong Duality holds

   (b) Strong duality doesn't hold.

   (c) The Lagrangian can be written as $L(x, \lambda) = (1 + \lambda)x^2 - 6\lambda x + 1 + 8\lambda$

   (d) The dual objective will be $g(\lambda) = \frac{-9\lambda^2}{1+\lambda} + 1 + 8\lambda$

Solution - a, c

$L(x, \lambda) = (x^2 + 1) + \lambda(x^2 - 6x + 8)$
$L(x, \lambda) = (1 + \lambda)x^2 - 6\lambda x + 1 + 8\lambda$
Hence, c is correct.
Now we will find the dual objective, $g(\lambda)$.
$\frac{\partial L}{\partial x} = 2x(1 + \lambda) - 6\lambda = 0$
$\hat{x} = \frac{3\lambda}{1+\lambda}$

$g(\lambda) = -\frac{9\lambda^2}{1+\lambda} + 1 + 8\lambda$ if $\lambda > -1; g(\lambda) = -\infty$ otherwise
Note that $\lambda \leq -1$ will mean the function will have a minima instead of maxima.

Now we want to maximize g, which will occur at $\lambda = 2$. $\hat{x} = 2$. The given optimization problem is convex and we can see that there are points in the relative interior of the domain. Hence strong duality holds.

4. Which of the following is/are true about the Perceptron classifier?

   (a) It can learn a OR function

   (b) It can learn a OR function

   (c) The obtained separating hyperplane depends on the order in which the points are presented in the training process.

   (d) For a linearly separable problem, there exists some initialization of the weights which might lead to non-convergent cases.

   Solution - a, b, c
   OR is a linear function, hence can be learnt by perceptron.
   XOR is non linear function which cannot be learnt by a perceptron learning algorithm which can learn only linear functions.
   The perceptron learning algorithm dependent on the order on which the data is presented, there are multiple possible hyperplanes, and depending on the order we will converge to any one of them.
   We can also prove that the algorithm always converges to a separating hyperplane if it exists. Hence d is false.

5. Which of the following is/are true regarding an SVM?

   (a) For two dimensional data points, the separating hyperplane learnt by a linear SVM will be a straight line.

   (b) In theory, a Gaussian kernel SVM can model any complex separating hyperplane.

   (c) For every kernel function used in a SVM, one can obtain a equivalent closed form basis expansion.

   (d) Overfitting in an SVM is a function of number of support vectors.

   Solution- a, b, d
   b - Gaussian kernel can be written as a taylor expansion and seen as a basis expansion of infinite dimensions, hence in theory giving it the ability to model any separating hyperplane.
   d - More the number of Support Vectors, higher the chance of the classifier being over fit.

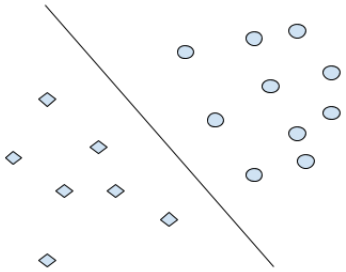

Figure 1: Q6

6. Consider a two class problem, whose training points are distributed in the figure below. One possible separating hyperplane is shown in the figure.

   (a) A classifier can be learnt using the perceptron training algorithm.

   (b) A linear SVM will not work well.

   (c) A linear SVM is sufficient for this data.

   (d) A non zero C value is essential for this data.

   Solution - b.
   Perceptron algorithm can learn linear classifier for linearly separable data. Even a linear SVM will work for the same reason. d is false because, C represents the cost of miss-classifying a point, and here irrespective of the cost, you can find a solution for the optimization problem.

7. For a two-class classification problem, we use an SVM classifier and obtain the following separating hyperplane. We have marked 4 instances of the training data. Identify the point which will have the most impact on the shape of the boundary on it's removal.

   (a) 1

   (b) 2

   (c) 3

   (d) 4

   Solution - a
   We need to identify support vectors on which the hyperplane is supported. The support vectors lie on a margin at a fixed distance from the separating hyperplane. By removing point 1, we separating hyperplane will change, since it is on the boundary.

8. For the dataset 1, train linear and radial basis function kernel SVMs. What are the number of Support Vectors in each of the case?

   (a) 100, 100

   (b) 10, 105

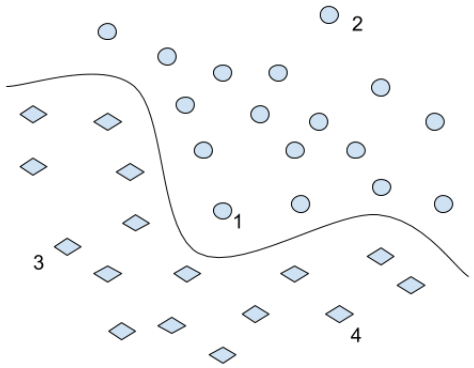

Figure 2: Q7

(c) 3, 104

(d) 500, 50

Solution - c
The point of observing the number of support vectors is to see the quality of the model learnt in some sense. If you get a model with very high number of support vectors, it means that the model has overfitted to the given data, and might not work very well on unseen data. This can be used as an indicator to decide on the kernel to be used, as you will notice in a later question.

9. For dataset 2, train 5 degree polynomial (5 degree, coef0 = 0), 10 degree polynomial (10 degree, coef0 = 0) and radial basis kernel functions. What are the number of support vectors for each?

   (a) 10, 300, 56

   (b) 324, 20, 27

   (c) 43, 98, 76

   (d) 12, 27, 20

   Solution - b

10. Based on the previous experiments, which would you think is the ideal classifier for Dataset 1.

    (a) Linear SVM

    (b) Polynomial SVM

    (c) Radial basis SVM

    Solution - a
    Linear SVM gives the classifier with the least number of support vectors which means the least overfitting thus, it would be the best.