CSCI 567, Spring 2021
Sirisha Rambhatla

Homework #4

Released: March 5, 2021
Due: 11:59 pm, March 19, 2021
Total scores: 40 points

Instructions

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Decision Trees (15 points)

Consider a binary dataset with 400 examples, where half of them belongs to class A and another half belongs to class B. Next consider two decision stumps (i.e. trees with depth 1) $\mathcal{T}_1$ and $\mathcal{T}_2$, each with two children. For $\mathcal{T}_1$, its left child has 150 examples in class A and 50 examples in class B; for $\mathcal{T}_2$, its left child has 0 example in class A and 100 examples in class B. (You can infer what are in the right child.)

1.1 For *each leaf* of $\mathcal{T}_1$ and $\mathcal{T}_2$, compute the corresponding classification error, entropy (base e) and Gini impurity. **Note**: Prediction for each leaf is the majority class of examples belonging to the leaf. **(10 points)**

1.2 Compare the quality of $\mathcal{T}_1$ and $\mathcal{T}_2$ (that is, the two different splits of the root) based on classification error, conditional entropy (base e), and weighted Gini impurity respectively.

Note: By compare the quality, we mean that in your solution (1) mention what is the value of classification error/ conditional entropy/ weighted Gini impurity for each *tree* $\mathcal{T}_1, \mathcal{T}_2$, and (2) which tree is better *for each* measure based on the calculated values. **(5 points)**

Problem 2 Boosting (8 points)

2.1 We discussed in class that AdaBoost minimizes the exponential loss greedily. In particular, the derivation of β_t is by finding the minimizer of $\epsilon_t(e^{\beta_t} - e^{-\beta_t}) + e^{-\beta_t}$ where ϵ_t is the weighted classification error of h_t and is fixed. Show that $\beta_t^* = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ is the minimizer. **(2 points)**

2.2 Recall that at round t of AdaBoost, a classifier h_t is obtained and the weighting over the training set is updated from D_t to D_{t+1}. Prove that h_t is only as good as random guessing in terms of classification error weighted by D_{t+1}, i.e., **(6 points)**

$$\sum_{n:h_t(\mathbf{x}_n)\neq y_n} D_{t+1}(n) = \frac{1}{2}.$$

in other words, the update is so that D_{t+1} is the "hardest" weighting for h_t.

Problem 3 Support Vector Machines (17 points)

Consider the dataset consisting of points (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. Let's start with three points $(x_1, y_1) = (-1, -1)$, $(x_2, y_2) = (1, -1)$, $(x_3, y_3) = (0, 1)$.

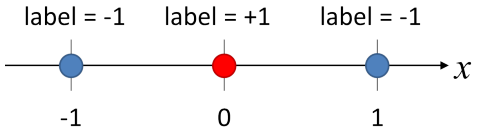

Figure 1: Three data points considered in Problem 3

3.1 Can three points shown in Figure 1, in their current one-dimensional feature space, be perfectly separated with a linear separator? Why or why not? **(1 points)**

3.2 Now we define a simple feature mapping $\phi(x) = [x, x^2]^T$ to transform the three points from one- to two-dimensional feature space. Plot the transformed points in the new two-dimensional feature space. Is there a linear decision boundary that can separate the points in this new feature space? Why or why not? **(2 points)**

3.3 Given the feature mapping $\phi(x) = [x, x^2]^T$, write down the kernel function $k(x, x')$. Moreover, write down the 3×3 kernel (or Gram) matrix $\mathbf{K}$ based on $k(x_i, x_j)$ of the three data points. Verify that $\mathbf{K}$ is a positive semi-definite (PSD) matrix. **(6 points)**

3.4 Write down the dual formulation of this problem (plugging in the kernel function values). **(2 points)**

3.5 Solve the dual form analytically. Then obtain primal solution $\mathbf{w}^*, b^*$ using dual solution. **(4 points)**

3.6 Let $\hat{y} = \mathbf{w}^{*T}\phi(\mathbf{x}) + b^*$ be the SVM prediction. Draw the decision boundary in the new two-dimensional feature space and circle all support vectors. (Set $\hat{y}$ to 0 to get the decision boundary). **(2 points)**