CSCI 567, Fall 2021
Haipeng Luo

Written Assignment #3

Due: Oct 26, 2021, 11:59 pm, PT

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder set up in which you can submit your files. Please be sure to follow all directions outlined here.

You can submit multiple times, but only *the last submission* counts. That means if you finish some problems and want to submit something first and update later when you finish, that's fine. In fact you are encouraged to do this: that way, if you forget to finish the homework on time or something happens, you still get credit for whatever you have turned in.

Problem sets must be typewritten or neatly handwritten when submitted. In both cases, your submission must be a single PDF. Please also follow the rules below:

- The file should be named as `firstname_lastname_USCID.pdf` (e.g., Joe_Doe_1234567890.pdf).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of the report as well.

Total points: 40 points

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 Support Vector Machines (19 points)

Consider a dataset consisting of points in the form of (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. There are only three points $(x_1, y_1) = (-1, -1)$, $(x_2, y_2) = (1, -1)$, and $(x_3, y_3) = (0, 1)$, shown in Figure 1.

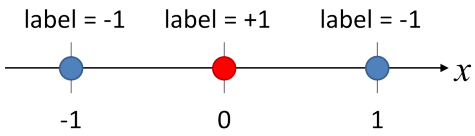

Figure 1: Three data points considered in Problem 1

1.1 Can these three points in their current one-dimensional feature space be perfectly separated with a linear classifier? Why or why not? **(2 points)**

No. A one-dimensional linear model $\text{SGN}(wx + b)$ is equivalent to a simple threshold function of the form

$$f(x) = \begin{cases} +1 & \text{if } x \geq \theta \\ -1 & \text{else} \end{cases} \quad \text{or} \quad f(x) = \begin{cases} -1 & \text{if } x \geq \theta \\ +1 & \text{else} \end{cases}$$

for some threshold $\theta \in \mathbb{R}$. It is thus clear that if we want points x_1 and x_2 to be correctly classified, then x_3 must be incorrectly classified. (Other correct reasoning gets full points as well.)

1.2 Now we define a simple feature mapping $\boldsymbol{\phi}(x) = [x, x^2]^T$ to transform the three points from one-dimensional to two-dimensional feature space. Plot the transformed points in the new two-dimensional feature space. Is there a linear model $\mathbf{w}^\top \mathbf{x} + b$ for some $\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$ that can correctly separate the three points in this new feature space? Why or why not? **(3 points)**

See Figure 2 for the plot. **(1 point)**
Yes. For example, any horizontal line with an intercept between 0 and 1 can correctly separate the data, which corresponds to $\mathbf{w} = (0, 1)$ and any $b \in (-1, 0)$. (It is enough to give one correct example.) **(2 points)**

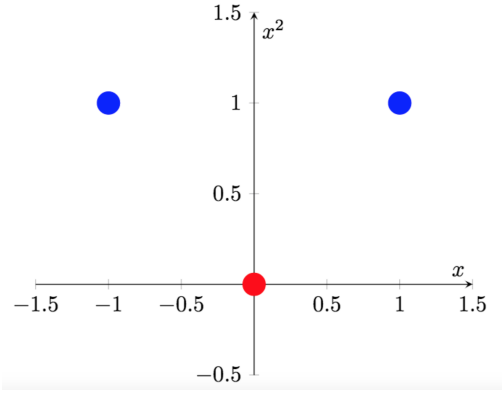

Figure 2: Plot for Q1.2: data points in new 2D space

1.3 Given the feature mapping $\boldsymbol{\phi}(x) = [x, x^2]^T$, write down the 3×3 kernel/Gram matrix $\mathbf{K}$ for this dataset. **(2 points)**

The kernel function is $k(x, x') = \boldsymbol{\phi}(x)^T \boldsymbol{\phi}(x) = xx' + (xx')^2$, so the Gram matrix is $\mathbf{K} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

1.4 Now write down the primal and dual formulations of SVM for this dataset in the two-dimensional feature space. Note that when the data is separable, we set the hyperparameter C to be $+\infty$ which makes sure that all slack variables (ξ) in the primal formulation have to be 0 (and thus can be removed from the optimization). **(4 points)**

General primal formulation of SVM for separable data is:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||_2^2 \\ \text{s.t.} \quad & y_n[\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) + b] \geq 1, \ \forall n \end{aligned}$$

Plugging in the specific dataset gives: **(2 points)**

$$\begin{aligned} \min_{w_1,w_2,b} \quad & \frac{1}{2}(w_1^2 + w_2^2) \\ \text{s.t.} \quad & w_1 - w_2 - b \geq 1 \\ & w_1 + w_2 + b \leq -1 \\ & b \geq 1 \end{aligned}$$

General dual formulation of SVM for separable data is:

$$\begin{aligned} \max_{\alpha} \quad & \sum_n \alpha_n - \frac{1}{2} \sum_{m,n} y_m y_n \alpha_m \alpha_n k(\mathbf{x}_m, \mathbf{x}_n) \\ \text{s.t.} \quad & \alpha_n \geq 0, \ \forall n \\ & \sum_n \alpha_n y_n = 0 \end{aligned}$$

Plugging in the specific dataset gives: **(2 points)**

$$\begin{aligned} \max_{\alpha_1,\alpha_2,\alpha_3 \geq 0} \quad & \alpha_1 + \alpha_2 + \alpha_3 - \alpha_1^2 - \alpha_2^2 \\ \text{s.t.} \quad & \alpha_1 + \alpha_2 = \alpha_3 \end{aligned}$$

Rubrics:

- okay to write the solution directly without first writing down the general form.
- for each formulation, 1 point for the objective and 1 point for the constraints.
- do not deduct points if the mistake is solely due to the incorrect Gram matrix from the last question.

1.5 Next, solve the dual formulation exactly (note: while this is not generally feasible as discussed in the lecture, the simple form of this dataset makes it possible). Based on that, calculate the primal solution. **(5 points)**

Eliminating the dependence on α_3 using the constraint $\alpha_1 + \alpha_2 = \alpha_3$, we arrive at the objective

$$\max_{\alpha_1,\alpha_2 \geq 0} 2\alpha_1 - \alpha_1^2 + 2\alpha_2 - \alpha_2^2.$$

Clearly we can maximize over α_1 and α_2 separately, which gives $\alpha_1^* = \alpha_2^* = 1$ and thus $\alpha_3^* = 2$. **(3 points)**

The primal solution can be found by

$$(w_1^*, w_2^*)^T = \sum_{n=1}^{3} y_n \alpha_n^* \boldsymbol{\phi}(x_n) = (0, -2)^T, \quad \text{(1 point)}$$

$$b^* = y_1 - \mathbf{w}^{*T} \boldsymbol{\phi}(x_1) = 1. \quad \text{(1 point)}$$

(Note that to calculate b^*, one can use any of the three examples, since they all satisfy $0 < \alpha_n < C = +\infty$.)

1.6 Plot the decision boundary (which is a line) of the linear model $\mathbf{w}^{*T}\mathbf{x} + b^*$ in the two-dimensional feature space, where $\mathbf{w}^*$ and b^* are the primal solution you got from the previous question. Then circle all support vectors. Finally, plot the corresponding decision boundary in the original one-dimensional space (which are just all the points x such that $\mathbf{w}^{*T}\boldsymbol{\phi}(x) + b^* = 0$). **(3 points)**

The decision boundary for the two-dimensional space is a horizontal line with intercept $1/2$. All three training points are support vectors. **(2 points)**

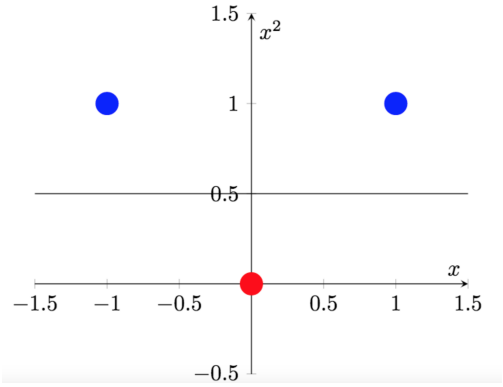

Figure 3: Plot for Q1.6: decision boundary in 2D space

The decision boundary for the one-dimensional space consists of two points $\frac{\sqrt{2}}{2}$ and $-\frac{\sqrt{2}}{2}$ (obtained by solving $\mathbf{w}^{*T}\boldsymbol{\phi}(\mathbf{x}) + b^* = -2x^2 + 1 = 0$). **(1 point)**

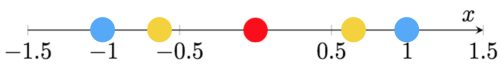

Figure 4: Plot for Q1.6: decision boundary in 1D space

Rubrics: similarly, for Q1.5 and Q1.6, do not deduct points for mistakes inherited from previous questions.

Problem 2 Decision trees (12 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.)

2.1 For each leaf of $\mathcal{T}_1$ and $\mathcal{T}_2$, compute the corresponding classification error, entropy (base e) and Gini impurity. You can either exactly express the final numbers in terms of fractions and logarithms, or round them to two decimal places. (Note: the value/prediction of each leaf is the majority class among all examples that belong to that leaf.) **(6 points)**

Classification error:

$$\epsilon_{1,L} = \frac{50}{150+50} = 0.25 \quad \text{(0.5 point)}$$

$$\epsilon_{1,R} = \frac{50}{50+150} = 0.25 \quad \text{(0.5 point)}$$

$$\epsilon_{2,L} = \frac{0}{0+100} = 0 \quad \text{(0.5 point)}$$

$$\epsilon_{2,R} = \frac{100}{200+100} \approx 0.33 \quad \text{(0.5 point)}$$

Entropy:

$$H_{1,L} = -\frac{150}{150+50}\ln(\frac{150}{150+50}) - \frac{50}{150+50}\ln(\frac{50}{150+50}) \approx 0.56 \quad \text{(0.5 point)}$$

$$H_{1,R} = -\frac{50}{150+50}\ln(\frac{50}{150+50}) - \frac{150}{150+50}\ln(\frac{150}{150+50}) \approx 0.56 \quad \text{(0.5 point)}$$

$$H_{2,L} = -\frac{0}{0+100}\ln(\frac{0}{0+100}) - \frac{100}{0+100}\ln(\frac{100}{0+100}) = 0 \quad \text{(0.5 point)}$$

$$H_{2,R} = -\frac{200}{200+100}\ln(\frac{200}{200+100}) - \frac{100}{100+200}\ln(\frac{200}{200+100}) \approx 0.64 \quad \text{(0.5 point)}$$

Gini impurity:

$$G_{1,L} = 1 - (\frac{150}{150+50})^2 - (\frac{50}{150+50})^2 = 0.375 \approx 0.38 \quad \text{(0.5 point)}$$

$$G_{1,R} = 1 - (\frac{50}{150+50})^2 - (\frac{150}{150+50})^2 = 0.375 \approx 0.38 \quad \text{(0.5 point)}$$

$$G_{2,L} = 1 - (\frac{0}{0+100})^2 - (\frac{100}{0+100})^2 = 0 \quad \text{(0.5 point)}$$

$$G_{2,R} = 1 - (\frac{200}{200+100})^2 - (\frac{100}{200+100})^2 \approx 0.44 \quad \text{(0.5 point)}$$

2.2 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. **(6 points)**

Let $p_1 = \frac{150+50}{400} = 0.5$ be the fraction of examples that belong to left leaf of $\mathcal{T}_1$, and $p_2 = \frac{0+100}{400} = 0.25$ be the fraction of examples that belong to left leaf of $\mathcal{T}_2$. Then the total classification error for $\mathcal{T}_1$ and $\mathcal{T}_2$ are respectively:

$$\epsilon_1 = p_1\epsilon_{1,L} + (1-p_1)\epsilon_{1,R} = 0.25 \quad \text{(0.5 point)}$$
$$\epsilon_2 = p_2\epsilon_{2,L} + (1-p_2)\epsilon_{2,R} = 0.25 \quad \text{(0.5 point)}$$

So they are as good in terms of classification error. **(1 point)**

The conditional entropy for $\mathcal{T}_1$ and $\mathcal{T}_2$ are respectively:

$$\epsilon_1 = p_1 H_{1,L} + (1-p_1)H_{1,R} \approx 0.56 \quad \text{(0.5 point)}$$
$$\epsilon_2 = p_2 H_{2,L} + (1-p_2)H_{2,R} \approx 0.48 \quad \text{(0.5 point)}$$

So $\mathcal{T}_2$ is better in terms of conditional entropy. **(1 point)**

The weighted Gini impurity for $\mathcal{T}_1$ and $\mathcal{T}_2$ are respectively:

$$\epsilon_1 = p_1 G_{1,L} + (1-p_1)G_{1,R} \approx 0.38 \quad \text{(0.5 point)}$$
$$\epsilon_2 = p_2 G_{2,L} + (1-p_2)G_{2,R} \approx 0.33 \quad \text{(0.5 point)}$$

So $\mathcal{T}_2$ is also better in terms of weighted Gini impurity. **(1 point)**

Problem 3 Boosting (9 points)

3.1 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. (You can use the fact that the function above is convex.) **(3 points)**

Set the derivative to 0:

$$\epsilon_t(e^{\beta_t} + e^{-\beta_t}) - e^{-\beta_t} = 0. \quad \text{(2 points)}$$

Multiplying both sides by e^{β_t} and rearranging give

$$e^{2\beta_t} = \frac{1}{\epsilon_t} - 1.$$

Solving for β_t finishes the proof. **(1 points)**

3.2 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}. That is

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

In other words, the update is so that D_{t+1} is the "hardest" weighting for h_t. **(6 points)**

By the algorithm we have

$$\sum_{n:h_t(\mathbf{x}_n)\neq y_n} D_{t+1}(n) \propto \sum_{n:h_t(\mathbf{x}_n)\neq y_n} D_t(n)e^{\beta_t} = \epsilon_t e^{\beta_t} = \sqrt{\epsilon_t(1-\epsilon_t)} \quad \text{(2 points)}$$

and similarly

$$\sum_{n:h_t(\mathbf{x}_n)=y_n} D_{t+1}(n) \propto \sum_{n:h_t(\mathbf{x}_n)=y_n} D_t(n)e^{-\beta_t} = (1-\epsilon_t)e^{-\beta_t} = \sqrt{(1-\epsilon_t)\epsilon_t}. \quad \text{(2 points)}$$

Note that $\sum_{n:h_t(\mathbf{x}_n)\neq y_n} D_{t+1}(n) + \sum_{n:h_t(\mathbf{x}_n)=y_n} D_{t+1}(n) = 1$. We thus have

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