CSCI 567, Spring 2019
Prof. Victor Adamchik

Theory Assignment 3

Due: 11:59 pm, March 24th, 2019

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder named 'Theory Assignment 3' 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 (remember Murphy's Law), 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. It is strongly recommended that you typeset with LaTeX. There are many free integrated LaTeX editors that are convenient to use (e.g Overleaf, ShareLaTeX). Choose the one(s) you like the most. This tutorial Getting to Grips with LaTeX is a good start if you do not know how to use LaTeX yet.

Please also follow the rules below:

- The file should be named as `firstname_lastname_USCID.pdf` e.g., `Don_Quijote_de_la_Mancha_8675309045.pdf`).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of your report as well.

Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit separately. Also in your report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration. Review the rules on academic conduct in the syllabus: a single instance of plagiarism can adversely affect you significantly more than you could stand to gain.

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 Kernel Methods (7 points)

On the midterm, we encountered the following kernel. Let us assume for simplicity that the kernel only operates on distinct data points, i.e. $\mathbf{x}_i \neq \mathbf{x}_j$ if $i \neq j$. In other words, the Gram matrix is the identity matrix.

$$k(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}' \\ 0, & \text{otherwise} \end{cases} \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^D.$$

Question 1 Suppose you are given a training set $\{(\mathbf{x}_n \in \mathbb{R}^D, y_n \in \mathbb{R})\}_{n=1}^N$ for a linear regression problem, where $\mathbf{x}_i \neq \mathbf{x}_j$ if $i \neq j$. Show that by using this kernel, the least square solution (with no regularization) will always lead to a total square loss of 0 – meaning that all the training examples are *predicted accurately* by the least square solution.

What to submit: A less-than-four line derivation of the model's perfect predictions on training data.

Ans: In the following, $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y} = \mathbf{y}$, since $\mathbf{K}$ is the identity matrix. The prediction of the least square solution on any example $(\mathbf{x}_m, y_m)$ is

$$\sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}_m) = \alpha_m = y_m,$$

that is, it correctly predicts the outcome and thus has 0 square loss.

Question 2 Although the least square solution has 0 loss on the training set, it in fact does not generalize to the test data all (that is, this algorithm completely overfits the training data). Specifically, show that for any unseen data point $\mathbf{x}$, that is, $\mathbf{x} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the prediction of the least square solution is always 0.

What to submit: A less-than-three line derivation of the model's 0 predictions on unseen data.

Ans: The prediction is $\sum_{n=1}^N \alpha_n k(\mathbf{x}_n, \mathbf{x})$, and since $\mathbf{x} \neq \mathbf{x}_n$ for all n, by the definition of the kernel, the prediction is $\sum_{n=1}^N \alpha_n 0 = 0$.

Problem 2 Support Vector Machine I (12 points)

In class, we saw that if our data is not linearly separable, then we need to modify our support vector machine (SVM) algorithm by introducing an error margin that must be minimized. In this problem we will consider a method known as the ℓ_2 norm soft margin SVM. This algorithm is given by the following optimization problem with squared slack penalties (ξ^2).

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \end{aligned}$$

Question 1 What is the Lagrangian $\mathcal{L}$ of the ℓ_2 soft margin SVM optimization problem?
What to submit: The lagrangian in terms of $\mathbf{w}, b, \xi_i, \alpha_i, \mathbf{x}_i, y_i$. Please include any constraints on variables.
Ans:

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i]$$

where $\alpha_i \geq 0$ for $i = 1, \ldots, m$.

Question 2 Minimize the Lagrangian with respect to $\mathbf{w}$, b, and $\boldsymbol{\xi}$ by taking the following gradients: $\nabla_{\mathbf{w}}\mathcal{L}$, $\frac{\partial\mathcal{L}}{\partial b}$, and $\nabla_{\boldsymbol{\xi}}\mathcal{L}$, and then setting them equal to 0. Here, $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_n]^T$.
What to submit: The results of setting $\nabla_{\mathbf{w}}\mathcal{L}$, $\frac{\partial\mathcal{L}}{\partial b}$, and $\nabla_{\boldsymbol{\xi}}\mathcal{L}$ to 0. Please simplify expressions.
Ans:

$$\begin{aligned} \nabla_{\mathbf{w}}\mathcal{L} = 0 \quad &\Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \frac{\partial\mathcal{L}}{\partial b} = 0 \quad &\Rightarrow \quad 0 = \sum_{i=1}^{n}\alpha_i y_i \\ \nabla_{\boldsymbol{\xi}}\mathcal{L} = 0 \quad &\Rightarrow \quad C\xi_i = \alpha_i \end{aligned}$$

Question 3 What is the dual of the ℓ_2 soft margin SVM optimization problem? Start with the objective of the dual, i.e. $\min_{\mathbf{w},b,\boldsymbol{\xi}}\mathcal{L}$. Hint: utilize your answers from Questions 1 and 2 to solve this question.
What to submit: The dual formulation of the problem and steps to reach it, starting at $\min_{\mathbf{w},b,\boldsymbol{\xi}}\mathcal{L}$.
Ans:

$$\begin{aligned} W(\alpha) &= \min_{\mathbf{w},b,\xi}\mathcal{L}(\mathbf{w}, b, \xi, \alpha) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i y_i \mathbf{x}_i)^T(\alpha_j y_j \mathbf{x}_j) + \frac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i}{\xi_i}\xi^2 - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\left(\sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j\right)^T\mathbf{x}_i + b\right) - 1 + \xi_i\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \frac{1}{2}\sum_{i=1}^{n}\alpha_i\xi_i - \left(\sum_{i=1}^{n}\alpha_i y_i\right)b + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \frac{1}{2}\sum_{i=1}^{n}\alpha_i\xi_i \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \frac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{C} \end{aligned}$$

Dual

$$\begin{aligned} \max_{\alpha} \quad & \left(\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \frac{1}{2} \sum_{i=1}^{n} \frac{\alpha_i^2}{C} \right) \\ \text{s.t.} \quad & \alpha_i \geq 0, i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

Problem 3 Support Vector Machine II (12 points)

Consider the dataset consisting of points (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. There are only three points $(x_1, y_1) = (-1, 1)$, $(x_2, y_2) = (1, 1)$, $(x_3, y_3) = (0, -1)$. Let the feature mapping $\boldsymbol{\phi}(u) = [u, u^2]^T$, corresponding to the kernel function $k(u, v) = uv + u^2v^2$.

Question 1 Write down the primal and dual formulations of SVM for this dataset in the transformed two-dimensional feature space based on $\boldsymbol{\phi}(\cdot)$. Note that we assume the data points are separable and set the hyperparameter C to be $+\infty$, which forces all slack variables (ξ) in the primal formulation to be 0 (and thus can be removed from the optimization).

What to submit: primal and dual formulations of optimization objectives with dataset values substituted in. The optimization objectives should include constraints.

Ans: General primal formulation of SVM for separable data is:

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

Plugging in the specific dataset gives:

$$\begin{aligned} \min_{w_1,w_2,w_3,b} \quad & \frac{1}{2}(w_1^2 + w_2^2) \\ \text{s.t.} \quad & -w_1 + w_2 + b \geq 1 \\ & w_1 + w_2 + b \geq 1 \\ & -b \geq 1 \end{aligned}$$

General dual formulation of SVM is:

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

Plugging in the specific dataset gives:

$$\begin{aligned} \max_{\alpha_1,\alpha_2,\alpha_3 \geq 0} \quad & \alpha_1 + \alpha_2 + \alpha_3 - \alpha_1^2 - \alpha_2^2 \\ \text{s.t.} \quad & a_1 + a_2 = a_3 \end{aligned}$$

Question 2 Next, solve the dual formulation. Based on that, derive the primal solution.

What to submit: Optimal $\alpha_1^*, \alpha_2^*, \alpha_3^*$ in dual solution, and optimal w_1^*, w_2^*, b^* in primal solution.

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

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

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

The primal solution can be found by

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

$$b^* = y_1 - \mathbf{w}^{*T} \boldsymbol{\phi}(x_1) = -1 \qquad \text{(using any example works in this case)}$$

Problem 4 Boosting (9 points)

Question 1 We discussed in class that AdaBoost minimizes the exponential loss greedily. In particular, Adaboost seeks the optimal β_t that minimizes

$$\epsilon_t(e^{\beta_t} - e^{-\beta_t}) + e^{-\beta_t}$$

where ϵ_t is the weighted classification error of h_t and is fixed. Show that $\beta^* = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ is the minimizer.
What to submit: A less-than-four line derivation of β^*.

Ans: Set the derivative to 0:

$$\epsilon_t(e^{\beta_t} + e^{-\beta_t}) - e^{-\beta_t} = 0$$

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

$$e^{2\beta t} = \frac{1}{\epsilon_t} - 1$$

Solving for β_t finishes the proof.

Question 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

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

What to submit: A less-than-six line derivation of the 1/2 result.

Ans: By the update algorithm, we have

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

and similarly

$$\sum_{n:h_t(\mathbf{x}_n)=y_n} D_{t+1}(n) \propto \sum_{n:h_t(\mathbf{x}_n)=y_n} D_t(n)e^{-\beta t} = \epsilon_t e^{-\beta t} = \sqrt{(1-\epsilon_t)\epsilon_t}$$

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

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