Tianxing Lan
tianxinglan@college.harvard.edu
CS181-S16

Assignment #3
Due: 5:00pm March 25, 2016

Collaborators: Kojin OshibaDylan Ho

Homework 3: SVM

There is a mathematical component and a programming component to this homework. Please submit ONLY your PDF to Canvas, and push all of your work to your Github repository. If a question asks you to make any plots, like Problem 3, please include those in the writeup.

Problem 1 (Fitting an SVM by hand, 8pts)
Consider a dataset with the following 6 points in $1D$:

$$\{(x_1, y_1)\} = \{(-3, +1), (-2, +1), (-1, -1), (1, -1), (2, +1), (3, +1)\}$$

Consider mapping these points to 2 dimensions using the feature vector $\phi : x \mapsto (x, x^2)$. The max-margin classifier objective is given by:

$$\min_{\mathbf{w}, w_0} \|\mathbf{w}\|_2^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T c + w_0) \geq 1, \ \forall i \tag{1}$$

Note: the purpose of this exercise is to solve the SVM without the help of a computer, relying instead on principled rules and properties of these classifiers. The exercise has been broken down into a series of questions, each providing a part of the solution. Make sure to follow the logical structure of the exercise when composing your answer and to justify each step.

1. Write down a vector that is parallel to the optimal vector **w**. Justify your answer.
2. What is the value of the margin achieved by **w**? Justify your answer.
3. Solve for **w** using your answers to the two previous questions.
4. Solve for w_0. Justify your answer.
5. Write down the discriminant as an explicit function of x.

Solution

1. **w** needs to be parallel to (0,1). This is because all the data points of $\phi(w_i)$ are symmetric with regard to y-axis, and by the rule of symmetry, the line that separates them with the maximum margin also needs to be symmetric with regard to y axis. Specifically, for an optimal $\mathbf{w_0}$, tilting **w** from being symmetric to y axis would either decrease its margin from (2,4) and (-3,9) or decrease its margin from (-2,4) and (3,9). Thus **w** has to be parallel to (0,1).

2. Just by looking at the graph of all ϕw_i we can see (-3,9) and (3,9) don't matter. The boundary should be between (-2,4), (2,4) and (-1,1),(1,1), the first two with y value 1 and the latter two with y value -1. Since the boundary needs to be parallel to (1,0), it achieves maximum margin when it is 1.5 away from both (2,4),(-2,4) and (1,-1),(-1,1). Thus the margin is 1.5.

3. By law of symmetry, we can solve for **w** just using (2,4) and (1,1). From 2, both of these points restrict **w**, so in both cases we have equality. 1(**w**(2,4)+w_0)=1 -1(**w**(1,1)+w_0)=1 **w**=c(0,1) Thus **w**=$(0, \frac{2}{3})$

4. plugging $\mathbf{w} = \frac{2}{3}$ in either of the equations and we get $w_0 = -\frac{5}{3}$

5. $f(x) = y_i(\frac{2}{3}x^2 - \frac{5}{3} - 1) = y_i(\frac{2}{3}x^2 - \frac{8}{3})$

Problem 2 (Composing Kernel Functions , 7pts)
Prove that

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-||\boldsymbol{x} - \boldsymbol{x}'||_2^2\},$$

where $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^D$ is a valid kernel, using only the following properties. If $K_1(\cdot, \cdot)$ and $K_2(\cdot, \cdot)$ are valid kernels, then the following are also valid kernels:

$$K(\boldsymbol{x}, \boldsymbol{x}') = c\, K_1(\boldsymbol{x}, \boldsymbol{x}') \quad \text{for } c > 0$$
$$K(\boldsymbol{x}, \boldsymbol{x}') = K_1(\boldsymbol{x}, \boldsymbol{x}') + K_2(\boldsymbol{x}, \boldsymbol{x}')$$
$$K(\boldsymbol{x}, \boldsymbol{x}') = K_1(\boldsymbol{x}, \boldsymbol{x}')\, K_2(\boldsymbol{x}, \boldsymbol{x}')$$
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\{K_1(\boldsymbol{x}, \boldsymbol{x}')\}$$
$$K(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})\, K_1(\boldsymbol{x}, \boldsymbol{x}')\, f(\boldsymbol{x}') \quad \text{where } f \text{ is any function from } \mathbb{R}^D \text{ to } \mathbb{R}$$

Solution

$K(x, x')$=$exp(-(x^T x - 2x^T x' + x'^T x'))$ $=$ $exp(-x^T x)exp(2x^T x')exp(x'^T x')$ Let $f(x)$ $=$ $exp(-x^T x)$, then $K(x, x')$ $=$ $f(x)exp(2x^T x')f(x')$ Since $x^T x'$ is a kernel, $2x^T x'$ is a kernel and thus $exp(2x^T x')$ is a kernel. Thus $K(x, x')$ is a kernel.

Problem 3 (Scaling up your SVM solver, 10pts (+7pts with extra credit))

In the previous homework, you studied a simple data set of fruit measurements. We would like you to code up a few simple SVM solvers to classify lemons from apples. To do this, read the paper at http://www.jmlr.org/papers/volume6/bordes05a/bordes05a.pdf and implement the Kernel Perceptron algorithm and the Budget Kernel Perceptron algorithm. The provided code has a base Perceptron class, which you will inherit to write KernelPerceptron and BudgetKernelPerceptron. This has been set up for you in problem3.py. The provided data is linearly separable. Make the optimization as fast as possible.

Additionally, we would like you to do some experimentation with the hyperparameters for each of these models. Try seeing if you can identify some patterns by changing β, N (maximum number of support vectors), or the number of random samples you take. Note the training time, accuracy, shapes/orientations of hyperplanes, and number of support vectors for various setups. We are intentionally leaving this open-ended to allow for experimentation, and so we will be looking for your thought process and not a rigid graph this time. That being said, any visualizations that you want us to grade and refer to in your descriptions should be included in this writeup. You can use the trivial $K(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{x_1}^T\mathbf{x_2}$ kernel for this problem, though you are welcome to experiment with more interesting kernels too. Also, answer the following reading questions in one or two sentences each.

1. In one short sentence, state the main purpose of the paper?
2. Identify each of the parameters in Eq. 1
3. State one guarantee for the Kernel perceptron algorithm described in the paper.
4. What is the main way the budget kernel perceptron algorithm tries to improve on the perceptron algorithm.
5. In simple words, what is the theoretical guarantee of LASVM algorithm? How does it compare to its practical performance?

For extra credit (+7 pts), implement the SMO algorithm and implement the LASVM process and do the same as above.

Solution

1. Using LASVM to evaluate the importance of data, and thus save more time and memory when performing high dimensional classification.
2. w':weight vector/coefficient vector for $\phi(x)$; $\phi(x)$ feature vector resulting from x; b: bias parameter
3. Novikoffs Theorem, which states that the perceptron algorithm converges after a finite number of mistakes, or after inserting a finite number of support vectors.
4. It removes support vectors from the kernel expansion to avoid noisy data and ultimately overfitting
5. It converges to the standard SVM algorithm. It runs faster and saves memory than SVM because it has the flexibility of online algorithm and runs through the data set only once.

Calibration [1pt]

Approximately how long did this homework take you to complete?