CS446 / ECE 449: Machine Learning, Fall 2020, Homework 2

Name: Saud Alrasheed (Sauda2)

Worked individually

Problem (2):

Deriving Support Vector Machines (SVMs) from the primal formulations.

Solution:

Problem 1.a:
Case 1: $z \leq 0$ Thus,

$$max(0, 1-z) \geq 1 = 1[z \leq 0] = 1[y \neq f(x, y)]$$

Case 2: $z > 0$ Thus,

$$max(0, 1-z) \geq 0 = 1[z \leq 0] = 1[y \neq f(x, y)]$$

Thus, the hinge loss upper bounds the 0-1 loss. Now, we show that the hinge loss is convex. For any z_1, z_2 we have that if both $z_1, z_2 \leq 1$ or $z_1, z_2 \geq 1$. Then, trivially, the following inequality holds, (The are both on same straight line),

$$f(tz_1 + (1-t)z_2) \leq tf(z_1) + (1-t)f(z_2)$$

Where f is the hinge loss, and $0 \leq t \leq 1$. The only case worth proving is when $z_1 < 1$ and $z_2 > 1$. Thus,

$$f(tz_1 + (1-t)z_2) \leq t(1-z_1) = t(1-z_1) + f(z_2) = tf(1-z_1) + (1-t)f(z_2)$$

Thus, the hinge loss is convex, and therefore is a surrogate loss.

Problem 1.b:
The hinge loss gives the points that are farther away from the decision margins greater loss value, and therefore reducing the chance of accepting miss-classified points by heavily penalizing them. Also, the hinge loss is differential-able almost every where.

Problem 2:

We have that the hinge loss is defined to be,

$$\ell(z) = \begin{cases} 1-z & z \leq 1 \\ 0 & Otherwise \end{cases}$$

Thus, the sub-gradient of the hinge loss with respect to the linear predictor function times the label is defined to be,

$$c_i(w) = \begin{cases} -1 & y^{(i)}w^T x^{(i)} < 1 \\ [-1, 0] & y^{(i)}w^T x^{(i)} = 1 \\ 0 & y^{(i)}w^T x^{(i)} > 1 \end{cases}$$

Problem 3:

One can easily observe that $\frac{\partial \ell}{\partial w} = y^{(i)}x^{(i)}c_i(w)$. Therefore, we have the following,

$$\partial_w \mathcal{E}^C = 0 \implies \frac{1}{n}\sum_{i=1}^{n} y^{(i)}x^{(i)}c_i(w) + 2Cw = 0 \implies w = \frac{-1}{2Cn}\sum_{i=1}^{n} y^{(i)}x^{(i)}c_i(w)$$

Problem 4:

Using the result from problem 3, we get,

$$f(x, w) = w^T x = \frac{-1}{2Cn}\sum_{i=1}^{n} y^{(i)}c_i(w)\phi(x^{(i)})^T\phi(x) = \frac{-1}{2Cn}\sum_{i=1}^{n} y^{(i)}c_i(w)k(x^{(i)}, x)$$

Problem 6.a: The fact that we only consider if $sgn(y) = sgn(f(w, x))$, then the we have a loss value of 0, gives rise to the sparsity. Although the Hinge loss leads to better accuracy and some sparsity, it fails to produce accurate sensitivity regarding probabilities.
Problem 6.b: Yes, $f(w, x)$ gives all the weight to miss-classified points, as it gives the points that are farther away from the decision margins greater loss value.
Problem 7.a:
For each training data point, the slack variable measure the distance of the point to its marginal hyper-plane. Thus, correctly classifying points corresponds to slack variable equal to 0, and positive slack variable otherwise.
Problem 7.b:
One can define the missclassification error to be $\sum_i 1[\xi_i > 0]$

Problem 8.a:

We have that,

$$\tilde{C} = \frac{1}{C}$$

Problem 8.b:

Clearly, as C goes to 0, the loss function put emphasis on allowing large slack variables. Thus, we will end uo with a softer margin. However, as C increases, we expect to have huge emphasis on the slack variables. Thus, we will end up with set of weights that almost perfectly seperate the data.