MI Exercise 9

2 Seriously Cool Guys

H9.1 (a) Keep in mind: (1) $\|\underline{w}\| = \sqrt{\underline{w}^T\underline{w}}$, (2) $r := d(\underline{x}^\alpha, \underline{w}, b)$

Unit vector $\rightarrow \frac{\underline{w}}{\|\underline{w}\|}$

Constraints:
$$\min_{\alpha=1,2,\dots,p} |\underline{w}^T \underline{x}^\alpha + b| \overset{!}{=} 1$$

Closest point on our hyperplane, $\underline{x}' = \underline{x}^\alpha - y r \frac{\underline{w}}{\|\underline{w}\|}$
w.r.t. a given data point x^α

y is our prediction in set $\{-1,1\}$

Given it's on the hyperplane, $w^T x' + b = 0$

$$\Downarrow$$

$$0 = w^T\left(\underline{x}^\alpha - y r^\alpha \frac{\underline{w}}{\|\underline{w}\|}\right) + b$$

Solving for r...

$$0 = \underline{w}^T \underline{x}^\alpha - y r^\alpha \frac{\underline{w}^T \underline{w}}{\|\underline{w}\|} + b$$

$$0 = \underline{w}^T \underline{x}^\alpha - y r^\alpha \|\underline{w}\| + b$$

$$r^\alpha = \frac{\underline{w}^T \underline{x}^\alpha + b}{y\|\underline{w}\|} = y \frac{\underline{w}^T \underline{x}^\alpha + b}{\|w\|}$$

So, for a minimum constraint $|\underline{w}^T \underline{x}^\alpha + b| \overset{!}{=} 1$

$$r^\alpha = d(\underline{x}^\alpha, \underline{w}, b) \geq \frac{1}{\|w\|},$$

given y is irrelevant, as it only will change the sign of the distance of the point to the hyperplane, but not the euclidean distance itself.

(1a) The shortest distance between a point and a hyperplane is perpendicular to the plane, and therefore parallel to our weight vector. So, a unit vector in this direction is the weight vector / the euclidean distance of the weight vector. We can then define a point on our hyperplane closest to any given x vector input as x', defined here, where y is our prediction {-1,1} and we denote the euclidean distance d(x,w,b) as r, the distance of the input vector to the hyperplane (which we've constrained as being minimum 1). We can then substitue this x' into our equation wTx + b, which equals 0 because the point is on the hyperplane. solving for r we get this equation and we see that y will only change the side of the hyper plane on which the point lies and the distance is dependent on 1/euclidean distance of the weight vector.

H9.1(b)

Primal Problem: $\min_{\underline{w},b} \frac{1}{2}|w|^2$ s.t. $(w^T x^\alpha + b) y_\alpha \geq 1$

in Lagrangian form ... $\min_{\underline{w},b} \max_{\lambda_\alpha \geq 0} L(\underline{w},b,\lambda_\alpha) = \frac{1}{2}|w|^2 - \sum_{\alpha=1}^{P} \lambda_\alpha \{ y_T^\alpha (w^T x^\alpha + b) - 1 \}$

Dual:

$$\max_{\lambda_\alpha \geq 0} \min_{\underline{w},b} L(\underline{w},b,\lambda_\alpha) = \frac{1}{2}|w|^2 - \sum_{\alpha=1}^{P} \lambda_\alpha \{ y_T^\alpha (w^T x^\alpha + b) - 1 \}$$

$$\frac{\partial L}{\partial w} = 0 = w - \sum_{\alpha=1}^{P} \lambda_\alpha y_\alpha x_\alpha \Rightarrow w = \sum_{\alpha} \lambda_\alpha y_T^\alpha x_\alpha$$

$$\frac{\partial L}{\partial b} = 0 = -\sum_{\alpha=1}^{P} \lambda_\alpha y_T^\alpha \Rightarrow \sum_{\alpha=1}^{P} \lambda_\alpha y_T^\alpha = 0$$

Substituting for $\underline{w}$ & b ..

$$\max_{\lambda_\alpha \geq 0, \sum_{\alpha=1}^{P} \lambda_\alpha y_T^\alpha = 0} L(\lambda_\alpha) = \sum_{\alpha=1}^{P} \lambda_\alpha - \frac{1}{2} \sum_{\alpha=1}^{P} \sum_{\beta=1}^{P} \lambda_\alpha \lambda_\beta y_{T\alpha} y_{T\beta} \underline{x}_\alpha^T \underline{x}_\beta$$

To prevent unbounded growth of λ_α for misclassifications, we add the contraint:

$$0 \leq \lambda_\alpha \leq \frac{C}{P}$$

(1b)With the primal problem, we minimize the Lagrangian by weights and biases that perfectly classfy all our data points according to their ground truths. We see that the dual problem just invovles swapping our min and max terms of the primal problem, and that these two problems are actually the same, just defined with different terms.

If at first we minimized our w and b assuming an optimal lambda, we now want to maximize our lambda assuming optimal w and b. First, we have to find optimal w and b. Knowing we have a concave optimization function, we take this to be where the derivative of our lagrangian is 0 wrt to both w and b. This derivative = 0 for the optimal w is intuitive -- it can be constructed as a linear combination of the ground truth and input vectors. Substituting these optimal w and b back into the lagrangian, we have the dual optimization problem, defined only in terms of our lambdas. We can then add an upper bound C/p to allow for missclassifications, otherwise, lambda could grown unbounded when our constraints are violated in the case misclassified points.