Practice Problems on Linear Classifiers

February 21, 2015

Note: Please do not copy answers from your friends. Please do not submit your solutions to us, as we do not plan to evaluate them. Please feel free to discuss solutions with course instructor.

1. Show that the hard-margin SVM formulation i.e., equation (2.3) in Summary notes, is equivalent[1] to the following:

$$\begin{aligned} \max_{w \in \mathbb{R}^n} \quad & \left(\min_{i:y_i=1} w^\top \phi(x_i) - \max_{i:y_i=-1} w^\top \phi(x_i)\right), \\ \text{s.t.} \quad & \|w\| = 1, \end{aligned} \tag{1}$$

with $b = \frac{\min_{i:y_i=1}(w^*)^\top \phi(x_i) + \max_{i:y_i=-1}(w^*)^\top \phi(x_i)}{\min_{i:y_i=1}(w^*)^\top \phi(x_i) - \max_{i:y_i=-1}(w^*)^\top \phi(x_i)}$, where w^* is the optimal solution of (1).

2. Motivated by above, which says "maximize the distance/margin between the closest scores, $w^\top \phi(x)$, in the positive and negative subsets", one might want to alternatively say "maximize the distance between the average scores with the positive and negative points":

$$\begin{aligned} \max_{w \in \mathbb{R}^n} \quad & \left(\frac{\sum_{i:y_i=1} w^\top \phi(x_i)}{m_+} - \frac{\sum_{i:y_i=-1} w^\top \phi(x_i)}{m_-}\right)^2, \\ \text{s.t.} \quad & \|w\| = 1, \end{aligned} \tag{2}$$

where m_+, m_- are the number of positive and negative points in the training set. Show that the above problem can be solved by simply calling a sub-routine that returns the highest eigen-value and the corresponding eigen-vector. Please write-down[2] the matrix for which you need to compute this largest eigen-value. Hint: Write down the objective in (2) as a (homogeneous) quadratic function and then write the Eigen-Value-Decomposition (EVD) of its hessian. Now try to argue things out with the Hessian written in its EVD form (use the fact that the eigen vectors are unit vectors orthogonal to each other).

[1] More specifically, show that (2.3) and (1) will induce the same hyperplane given the same training set.

[2] Your proof will show that (2) is an example of a non-convex optimization problem that has efficient (polynomial time) algorithm.

3. Let $\mu_+ \equiv \frac{\sum_{i:y_i=1}\phi(x_i)}{m_+}$ and $\mu_- \equiv \frac{\sum_{i:y_i=-1}\phi(x_i)}{m_-}$. While it is clear that (2) relaxes (1) and allow points to lie on the incorrect sides, it will work well only if we include a term that penalizes such mistakes. So, one might want to say "do the above while restricting the sum of variances in the scores of the two sets to be small":

$$\begin{aligned} \max_{w\in\mathbb{R}^n} \quad & \left(w^\top\mu_+ - w^\top\mu_-\right)^2, \\ \text{s.t.} \quad & \frac{\sum_{i:y_i=1}\left(w^\top\phi(x_i)-w^\top\mu_+\right)^2}{m_+} + \frac{\sum_{i:y_i=-1}\left(w^\top\phi(x_i)-w^\top\mu_-\right)^2}{m_-} \leq 1, \end{aligned} \quad (3)$$

Show that this problem also can be solved by simply calling a sub-routine that returns the highest eigen-value and the corresponding eigen-vector. Please write-down the matrix for which you need to compute this largest eigen-value. (3) is popularly known as Fisher Linear Classifier[3]. Hint: Write the constraint as a quadratic one and perform a change of variables that makes this constraint same as the constraint in (2).

4. Compute the conjugates for the various binary classfication ($\mathcal{Y} = \{-1, 1\}$) losses defined in lectures. More specifically, show that:

   Square-hinge-loss: Let $l_i(z) \equiv \max(0, 1-y_iz)^2$. Show that its conjugate is given by: $l_i^*(\alpha) = \frac{\alpha^2}{4} + y_i\alpha,\ \forall\ y_i\alpha \leq 0$.

   Logistic-loss: Let $l_i(z) \equiv \log(1+e^{-y_iz})$. Show that its conjugate is the negative entropy: $l_i^*(\alpha) = (1+\alpha y_i)\log(1+\alpha y_i) - \alpha y_i \log(-\alpha y_i),\ \forall\ \alpha y_i \in [0, -1]$.

   Exponential-loss: Let $l_i(z) \equiv e^{-y_iz}$. Show that its conjugate is given by: $l_i^*(\alpha) = -y_i\alpha\log(-y_i\alpha) + y_i\alpha,\ \forall\ y_i\alpha \leq 0$.

5. Using the above conjugates, and the fact that equation (2.7), in lecture notes, is the dual of equation (2.8), in lecture notes, derive the duals of MMLC with all the above three loss functions. Also write-down all the optimality conditions (using the same methodology as in lectures). Simplify the dual as well as optimality conditions as much as possible[4].

6. Exercise 4.3 in "Foundations of ML" book by Mohri et.al.

[3] Note that there is no elegant loss function that when substituted in MMLC will give Fisher LC.

[4] After doing this exercise do a search on internet to verify if all your derivations are correct.