ELE 535: Machine Learning and Pattern Recognition
Homework 8

Zachary Hervieux-Moore

Friday 7th December, 2018

Exercise 1: Let $\{(x_i, y_i)\}_{i=1}^m$ with $x_i \in \mathbb{R}^n$ and $y_i \in \{\pm 1\}$, $i \in [1:m]$, be a linearly separable set of training data. Show that if C is sufficiently large, the solution of the primal SVM problem will give the unique maximum margin separating hyperplane. How large does C need to be?

Answer: We first note that if the data is linearly separable, then there exists a feasible solution to the primal SVM problem where $s_i = 0$ for all support vectors $i \in A$. Now, let α_{simple} be the solution to the simple linear SVM introduced in class. We will show that if $\max_i \alpha_{simple_i} < C$, that is, the largest component of the solution to the simple linear SVM is smaller than C, then the primal SVM problem returns the unique maximum margin separating hyperplane.

First, to extract the maximum margin separating hyperplane, we will need to force all the $s_i = 0$, $i \in A$. To accomplish this, by complementary slackness, we need $\mu_i > 0$ or equivalently $0 < \alpha_i^* < C$. Thus, if C is sufficiently large, we extract the maximum margin separating hyperplane. However, as noted before, α_{simple} solves the maximum margin separating hyperplane, thus if we set $\max_i \alpha_{simple_i} < C$, then we will guarantee that C is sufficiently large. α_{simple} can easily be retrieve by solving the dual problem of the simple linear SVM.

Exercise 2: Let $\{(x_i, y_i)\}_{i=1}^m$ with $x_i \in \mathbb{R}^n$ and $y_i \in \{\pm 1\}$, $i \in [1 : m]$ be a training dataset. For a fixed value of C, let the corresponding SVM classifier have parameters w^*, b^*.

a) Let $h \in \mathbb{R}^n$ and $Q \in \mathcal{O}_n$, and form the second training set: $\{(Q(x_i - h), y_i)\}_{i=1}^m$. Show that the SVM classifier for this second dataset using the same value of C has parameters Qw^*, ${w^*}^T h + b^*$.

b) If we first center the training examples, how does this change the SVM classifier?

Answer:

a) We have that the primal problem is:

$$\begin{gathered}\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} \frac{1}{2} w^T w + C\mathbf{1}^T s \\ Z^T w + by + s - \mathbf{1} \geq \mathbf{0} \\ s \geq \mathbf{0}\end{gathered}$$

where $Z = [y_1 x_1, \dots, y_m x_m]$. Suppose that w^* and b^* is the solution. Now we transform the input data to be $\tilde{x}_i = Q(x_i - h)$. Then we have $\tilde{Z} = [y_1 Q(x_1 - h), \dots, y_m Q(x_m - h)]$. The Primal problem for this shifted problem is

$$\begin{gathered}\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} \frac{1}{2} w^T w + C\mathbf{1}^T s \\ \tilde{Z}^T w + by + s - \mathbf{1} \geq \mathbf{0} \\ s \geq \mathbf{0}\end{gathered}$$

Expanding the first constraint yields

$$\begin{aligned} & \tilde{Z}^T w + by + s - \mathbf{1} \geq \mathbf{0} \\ \iff & [y_1 Q(x_1 - h), \dots, y_m Q(x_m - h)]^T w + by + s - \mathbf{1} \geq \mathbf{0} \\ \iff & Z^T Q^T w - h^T Q^T w y + by + s - \mathbf{1} \geq \mathbf{0} \\ \iff & Z^T Q^T w - (h^T Q^T w + b) y + s - \mathbf{1} \geq \mathbf{0} \end{aligned}$$

Now we make the substitution $\tilde{w} = Q^T w$ and $\tilde{b} = h^T \tilde{w} + b$ to get the primal problem

$$\begin{gathered}\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} \frac{1}{2}\tilde{w}^T\tilde{w} + C\mathbf{1}^T s \\ \tilde{Z}^T\tilde{w} + \tilde{b}y + s - \mathbf{1} \geq \mathbf{0} \\ s \geq 0\end{gathered}$$

Notice that the orthonormal matrices canceled each other out in the objective. Also, this problem is identical to what we had before. Thus we have that $\tilde{w} = w^*$ and $\tilde{b} = b^*$. Or undoing the transformations, we have that the solution to the shifted problem is $w'^* = Qw^*$ and $b'^* = b^* + h^T w^*$.

b) If we center the data, this is equivalent to the previous problem with $Q = I_n$ and $b = \mu_x = \sum_{i=1}^m x_i$. Thus, the solution is

$$w'^* = w^* \quad \text{and} \quad b'^* = \mu_x^T w^* + b^*$$

So it only shifts the intercept which makes sense as all the points get the same translation and so the normal of the separating hyperplane does not change.

Exercise 3: Give a clear and concise derivation of the dual of the primal linear SVM problem shown below and explain the origin of each of the constraints in the dual problem.

$$\begin{gathered}\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} \frac{1}{2}w^T w + C\mathbf{1}^T s \\ \text{s.t. } Z^T w + by + s - \mathbf{1} \geq \mathbf{0} \\ s \geq \mathbf{0}\end{gathered}$$

Answer: We have that the Lagrangian for this problem is:

$$L(w,b,s,\alpha,\mu) = \frac{1}{2}w^T w + C\mathbf{1}^T s - \alpha^T(Z^T w + by + s - \mathbf{1}) - \mu^T s$$

Now, following the procedure in Chapter 14, we have that the dual objective is

$$g(\alpha,\mu) = \min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} L(w,b,s,\alpha,\mu)$$

To solve this, we take the derivatives and set them to 0. We also note the domain of $g(\cdot)$ is $\alpha \geq \mathbf{0}$ and $\mu \geq \mathbf{0}$.

$$\begin{gathered}\nabla_w L = 0 \implies w - Z\alpha = \mathbf{0} \text{ or } w = Z\alpha \\ \nabla_b L = 0 \implies \alpha^T y = 0 \\ \nabla_s L = 0 \implies C\mathbf{1} - \alpha - \mu = \mathbf{0} \text{ or } \mu = C\mathbf{1} - \alpha\end{gathered}$$

Putting these substitutions into the objective yields

$$\begin{aligned}g(\alpha,\mu) &= \frac{1}{2}\alpha^T Z^T Z\alpha + C\mathbf{1}^T s - \alpha^T(Z^T Z\alpha + by + s - \mathbf{1}) - (C\mathbf{1} - \alpha)^T s \\ &= -b\alpha^T y + \alpha^T\mathbf{1} - \frac{1}{2}\alpha^T Z^T Z\alpha\end{aligned}$$

But our derivatives above show that $\alpha^T y = 0$ so our dual objective is.

$$g(\alpha, \mu) = \alpha^T \mathbf{1} - \frac{1}{2}\alpha^T Z^T Z \alpha$$

This objective coupled with our derivative constraints and domain requirements yields the dual problem

$$\begin{gathered}
\max_{\alpha \in \mathbb{R}^m, \mu \in \mathbb{R}^m} \alpha^T \mathbf{1} - \frac{1}{2}\alpha^T Z^T Z \alpha \\
\text{s.t. } \alpha^T y = 0 \\
C\mathbf{1} - \alpha - \mu = \mathbf{0} \\
\alpha \geq \mathbf{0} \\
\mu \geq \mathbf{0}
\end{gathered}$$

Where we can do the same trick done in the notes and remove μ from the constraints (as it is a direct function of α)

$$\begin{gathered}
\max_{\alpha \in \mathbb{R}^m} \alpha^T \mathbf{1} - \frac{1}{2}\alpha^T Z^T Z \alpha \\
\text{s.t. } \alpha^T y = 0 \\
\alpha \leq C\mathbf{1} \\
\alpha \geq \mathbf{0}
\end{gathered}$$

Exercise 4: Suppose that instead of using $C\sum_{i=1}^m s_i$ as the penalty term in the objective of the primal SVM problem we use the quadratic penalty $\frac{1}{2}C\sum_{i=1}^m s_i^2$, while maintaining the constraint $s_i \geq 0$

a) Formulate the new primal problem in vector form. When is the primal problem feasible?

b) Does strong duality hold for this problem? Justify your answer.

c) Write down the KKT conditions.

d) Find the dual problem.

Answer:

a) The vector formulation of this problem is

$$\begin{aligned} \min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} \; & \frac{1}{2}w^T w + \frac{C}{2}s^T s \\ \text{s.t. } & Z^T w + by + s - \mathbf{1} \geq \mathbf{0} \\ & s \geq \mathbf{0} \end{aligned}$$

The primal problem is always feasible. For example, take $w = \mathbf{0}$, $b = 0$, and $s = \mathbf{1}$. These always satisfy the constraints and so the primal problem is always feasible. If one doesn't like the use of zeros, take any w and b. Then $Z^T w + by$ will have some entry that is the smallest (most negative). Simply set $s = \mathbf{1}|\min_i(Z^T w + by)_i|$ and this will also be feasible.

b) Strong duality holds if Slater's condition is satisfied. First note that the problem is convex as the objective is quadratic and the constraints are affine. As the constraints are affine, Slater's condition just requires a feasible point in the primal problem. As we have shown this already, then strong duality holds for this problem.

c) The Lagrangian for the problem is

$$L(w, b, s, \alpha, \mu) = \frac{1}{2}w^T w + \frac{C}{2}s^T s - \alpha^T(Z^T w + by + s - \mathbf{1}) - \mu^T s$$

Thus the KKT conditions are (taking the requisite derivatives, primal constraints, dual constraints, and complementary slackness):

$$\begin{aligned}
w - Z\alpha &= 0 \quad && (\nabla_w L = \mathbf{0}) \\
\alpha^T y &= 0 \quad && (\nabla_b L = 0) \\
Cs - \alpha - \mu &= \mathbf{0} \quad && (\nabla_s L = \mathbf{0}) \\
Z^T w + by + s - \mathbf{1} &\geq \mathbf{0} \quad && \text{primal constraint} \\
s &\geq \mathbf{0} \quad && \text{primal constraint} \\
\alpha &\geq \mathbf{0} \quad && \text{dual constraint} \\
\mu &\geq \mathbf{0} \quad && \text{dual constraint} \\
\alpha \otimes (Z^T w + by + s - \mathbf{1}) &= \mathbf{0} \quad && \text{complementary slackness} \\
\mu \otimes s &= \mathbf{0} \quad && \text{complementary slackness}
\end{aligned}$$

d) We again follow the steps of deriving the dual objective from the previous question. Namely just substituting the derivatives from the KKT conditions into the primal objective.

$$g(\alpha, \mu) = \min_{w\in\mathbb{R}^n, b\in\mathbb{R}, s\in\mathbb{R}^m} L(w, b, s, \alpha, \mu)$$

$$= \frac{1}{2}\alpha^T Z^T Z\alpha + \frac{C}{2}\frac{1}{C^2}(\alpha+\mu)^T(\alpha+\mu) - \alpha^T(Z^T Z\alpha + by + \frac{1}{C}(\alpha+\mu) - \mathbf{1})$$
$$-\frac{1}{C}\mu^T(\alpha+\mu)$$
$$= \alpha^T\mathbf{1} + \frac{1}{2C}(\alpha+\mu)^T(\alpha+\mu) - \frac{1}{C}(\alpha+\mu)^T(\alpha+\mu) - \frac{1}{2}\alpha^T Z^T Z\alpha$$
$$= \alpha^T\mathbf{1} - \frac{1}{2C}(\alpha+\mu)^T(\alpha+\mu) - \frac{1}{2}\alpha^T Z^T Z\alpha$$

Thus, the dual problem becomes

$$\begin{aligned}
\max_{\alpha\in\mathbb{R}^m, \mu\in\mathbb{R}^m} \ & \alpha^T\mathbf{1} - \frac{1}{2C}(\alpha+\mu)^T(\alpha+\mu) - \frac{1}{2}\alpha^T Z^T Z\alpha \\
\text{s.t. } & \alpha^T y = 0 \\
& \alpha \geq \mathbf{0} \\
& \mu \geq \mathbf{0}
\end{aligned}$$

Exercise 5: You are provided with $m > 1$ data points $\{x_j \in \mathbb{R}^n\}_{j=1}^m$ of which at least d, with $1 < d \leq m$ are distinct. Let $X = [x_1, \ldots, x_m]$ and consider the one class SVM problem:

$$\begin{aligned} \min_{R\in\mathbb{R}, a\in\mathbb{R}^n, s\in\mathbb{R}^m} \quad & R^2 + C\mathbf{1}^T s \\ \text{s.t.} \quad \|x_i - a\|_2^2 &\leq R^2 + s_i, \quad i = 1, \ldots, m \\ s &\geq \mathbf{0} \end{aligned}$$

a) Show that this is a feasible convex program and that strong duality holds. [Hint: let $r = R^2$]

b) Write down the KKT conditions.

c) Show that $\alpha^* \neq \mathbf{0}$ and that if $C > 1/(d-1)$ then $(R^2)^* > 0$ (harder).

d) What are the support vectors for this problem?

e) Derive the dual problem.

f) Assume $C > 1/(d-1)$. Given the dual solution, how should a and R^2 be selected?

Answer:

a) Following the hint, let $r = R^2$ and add the additional constraint that $r \geq 0$ to enforce positivity. Thus, the problem becomes

$$\begin{aligned} \min_{r\in\mathbb{R}, a\in\mathbb{R}^n, s\in\mathbb{R}^m} \quad & r + C\mathbf{1}^T s \\ \text{s.t.} \quad \|x_i - a\|_2^2 &\leq r + s_i, \quad i = 1, \ldots, m \\ s &\geq \mathbf{0} \\ r &\geq 0 \end{aligned}$$

In which case this has a linear (convex) objective and convex constraints. The main one being quadratic and hence convex. This implies the problem is convex. Again, this is feasible if we set $a = \mathbf{0}$, $r = 0$, and $s_i = \|x_i\|_2^2$. To show strong duality, we need Slater's condition to hold.

Picking $a = \mathbf{0}$, $r = 1$, and $s_i = \|x_i\|_2^2$ means that $r > 0$, $\|x_i\|_2^2 < r + s_i$ and thus we have a point that satisfies Slater's condition. We conclude strong duality holds.

b) The Lagrangian for this problem is

$$L(R, a, s, \alpha, \mu, \lambda) = R^2 + C\mathbf{1}^T s + \sum_{i=1}^{m} \alpha_i(\|x_i - a\|_2^2 - R^2 - s_i) - \mu^T s$$

Thus the KKT conditions are (taking the requisite derivatives, primal constraints, dual constraints, and complementary slackness):

$$\begin{aligned}
R \cdot (1 - \sum_{i=1}^{m} \alpha_i) &= 0 \quad (\nabla_R L = 0) \\
a \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i x_i &= \mathbf{0} \quad (\nabla_a L = \mathbf{0}) \\
C\mathbf{1} - \alpha - \mu &= \mathbf{0} \quad (\nabla_s L = \mathbf{0}) \\
\|x_i - a\|_2^2 &\leq R^2 + s_i \quad \text{primal constraint} \\
s &\geq \mathbf{0} \quad \text{primal constraint} \\
\alpha &\geq \mathbf{0} \quad \text{dual constraint} \\
\mu &\geq \mathbf{0} \quad \text{dual constraint} \\
\alpha_i \cdot (\|x_i - a\|_2^2 - R^2 - s_i) &= \mathbf{0} \quad \text{complementary slackness} \\
\mu \otimes s &= \mathbf{0} \quad \text{complementary slackness}
\end{aligned}$$

c) We first show that $\alpha^* \neq \mathbf{0}$. Suppose by contradiction that $\alpha^* = \mathbf{0}$. Then by complementary slackness, we have that

$$\|x_i - a\|_2^2 \leq R^2 + s_i \quad \forall i$$

As $\alpha = \mathbf{0}$, we also have that $R = 0$ from $\nabla_R L = 0$. This implies

$$\|x_i - a\|_2^2 \leq s_i \quad \forall i$$

Also, from $\nabla_s L = \mathbf{0}$, we have that $\mu = C\mathbf{1}$. This again implies by complementary slackness that $s_i = 0$. Hence, this implies that

$$\|x_i - a\|_2^2 \leq 0 \quad \forall i$$

However, this is only possible if all the x_i are equal. But we have that there are $d > 1$ distinct data points and thus the last conclusion is a contradiction. Thus, we conclude that $\alpha^* \neq \mathbf{0}$.

We now show the second part. We do this by contraposition. That is, we will show that $R^{2^*} = 0 \implies C \leq 1/(d-1)$. If we assume that $R^2 = 0$, we get from $\nabla_R L = 0$ and $\alpha \geq \mathbf{0}$ that

$$\sum_{i=1}^{m} \alpha_i \leq 1$$

We also have from $\nabla_s L = 0$ and $\mu \geq \mathbf{0}$ that

$$\alpha \leq C\mathbf{1}$$

Now, we argue that the worst case is that $(d-1)$ components of α matter in the worst case. If we have that $x_i = a$ for $m - d + 1$ data points (to maintain the assumption that there are d distinct points). Then we have that $\|x_i - a\|_2^2 = 0 = s_i$. Recall that $R^2 = 0$ by assumption. In this case picking $s_i = 0$ clearly minimizes the objective. Also, note that if $x_i = a$, the then equality in $\nabla_a L = \mathbf{0}$ only depends on the components that $x_i \neq a$. Thus, we can set α_i to whatever we want for the components that $x_i = a$. In the worst case, we set $\alpha_i = 0$. Thus, our previous inequality become

$$\sum_{i=1}^{d-1} \alpha_i \leq 1$$

Combining this with $\alpha \leq C\mathbf{1}$ yields

$$(d-1)C \leq 1 \implies C \leq 1/(d-1)$$

Thus, we have shown that $R^{2^*} = 0 \implies C \leq 1/(d-1)$ whose contrapositive is $C > 1/(d-1) \implies R^{2^*} > 0$.

d) The support vectors are as follows.

   1) When $0 < \alpha_i^* < C$, then we have that $\mu_i > 0$ from $\nabla_s L = 0$ which implies by complementary slackness that $s_i = 0$. Thus, these support vectors sits directly on the border of the sphere centered at a with radius R.

   2) If $\alpha_i^* = C$, then we have that $\mu_i = 0$ and $s_i \geq 0$. These vectors lie outside the sphere centered at a with radius R.

e) If we assume that $C > 1/(d-1)$, then we have that $R^2 > 0$ and that $\sum_{i=1}^m \alpha_i = 1$ from $\nabla_R = 0$. Which means that $a = \sum_{i=1}^m \alpha_i x_i$ from $\nabla_a L = \mathbf{0}$. Making this substitution into the primal problem and canceling out the terms that result from the derivatives in the KKT conditions (the same steps we've done in previous questions), the dual problem is

$$\begin{aligned} g(\alpha, \mu) &= \min_{R\in\mathbb{R}, a\in\mathbb{R}^n, s\in\mathbb{R}^m} L(R, a, s, \alpha, \mu) \\ &= \sum_{i=1}^m \alpha_i(\|x_i - \sum_{j=1}^m \alpha_j x_j\|_2^2) \\ &= \sum_{i=1}^m \left[x_i^T x_i \alpha_i + \sum_{j=1}^m \left(-2x_i^T x_j \alpha_j \alpha_i + x_j^T x_j \alpha_j^2 \alpha_i \right) \right] \end{aligned}$$

Bringing in the outside sum and using the fact that $\sum_{i=1}^m \alpha_i = 1$, we can change the last term to be

$$g(\alpha, \mu) = \sum_{i=1}^m \left[x_i^T x_i \alpha_i + \sum_{j=1}^m \left(-2x_i^T x_j \alpha_j \alpha_i + x_j^T x_j \alpha_j^2 \right) \right]$$

It is easy to show that the first term can be written as $\mathrm{diag}(X^T X)^T \alpha$ (**note:** the class notes incorrectly say it is $\mathbf{1}^T \mathrm{diag}(X^T X)\alpha$ which does not make sense) and the second and third terms can be written as $-\alpha^T X^T X \alpha$ simply by writing these as summations and canceling the right terms. Thus, we get a compact formulation of the dual

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^m, \mu \in \mathbb{R}^m} \quad & \mathrm{diag}(X^T X)^T \alpha - \alpha^T X^T X \alpha \\ \text{s.t.} \quad & 1 = \sum_{i=1}^m \alpha_i \\ & C\mathbf{1} - \alpha - \mu = \mathbf{0} \\ & \alpha \geq \mathbf{0} \\ & \mu \geq \mathbf{0} \end{aligned}$$

Using the same trick as before to get rid of μ and writing the first constraint in vector form,

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^m, \mu \in \mathbb{R}^m} \quad & \mathrm{diag}(X^T X)^T \alpha - \alpha^T X^T X \alpha \\ \text{s.t.} \quad & \mathbf{1}^T \alpha = \mathbf{1} \\ & \alpha \leq C\mathbf{1} \\ & \alpha \geq \mathbf{0} \end{aligned}$$

f) Assuming that $C > 1/(d-1)$. Given the dual solution, we have that $a = \sum_{i=1}^m \alpha_i x_i$. Once you know a, then it is trivial to find R^2 and s as the primal problem no longer has a quadratic constraint (the terms involving a become constants). It becomes a linear program which is readily solved. Alternatively, without the linear program, complementary slackness ensures that there are m equations involving s and R^2. As $1 = \sum_{i=1}^{m} \alpha_i$, we know that $\|x_i - a\|_2^2 = R^2 + s_i$ for some i. If $\mu_i > 0$ and $\alpha_i > 0$ then we are done as R^2 is immediately available as $s_i = 0$. Suppose that $\mu_i = 0$ whenever $\alpha_i > 0$ and $\alpha_i = 0$ whenever $\mu_i > 0$, this is the worst case. Then we have m equations but $m + 1$ unknowns. However, as it is a minimization problem, we write out all the equations with $\|x_i - a\|_2^2 = R^2 + s_i$ and start out with $R^2 = 0$. Record the objective for this value of R^2. Then we increase R^2 while the objective keeps decreasing and stop (or decrease step size) when the objective increases. Binary search would provide quick convergence as we know the largest R^2 can become is the smallest s_i value when $R^2 = 0$.