Solution SVM question

2022 Exam

Hi all!

Sadly, I have not been able to find someone with an Ipad or so to record the solution to this live. However, I will try and explain the question and solution in this document :)

Understanding the problem So, what is this question about? We assume we have some boundary M with some predefined shape given by function f, in this case, something resembling one farfalla pasta. We will assume that our data is separable by this shape if we find the right scale σ for M.

We see that the 'signed distance' is defined by $d(\mathbf{x}, \sigma) := ||\mathbf{x}||_2 - \sigma f(\mathbf{x})$. Let's unpack this first. Here, $\sigma f(\mathbf{x})$ is simply the point on the scaled boundary which is in the same 'radial' direction as some point $\mathbf{x}$. As such, suppose the point $\mathbf{x}$ lies closer to the origin than the point $\sigma f(\mathbf{x})$, i.e. we have that $||\mathbf{x}|| \leq \sigma f(\mathbf{x})$. Then, we have that $d(\mathbf{x}, \sigma) = ||\mathbf{x}||_2 - \sigma f(\mathbf{x}) \leq 0$. Similarly, we have that if $\mathbf{x}$ lies further from the origin than $\sigma f(\mathbf{x})$, we have that $d(\mathbf{x}, \sigma) = ||\mathbf{x}||_2 - \sigma f(\mathbf{x}) \geq 0$.

Let us now choose to classify all points further than $\sigma f(\mathbf{x}_n)$ from the origin as $t_n = +1$, and all points closer to the origin as $t_n = -1$. Then, it follows that $t_n d(\mathbf{x}_n, \sigma) \geq 0$.

Notice, however, that this inequality still holds if we scale the entire thing with some positive number α, i.e.

$$t_n d(\mathbf{x}_n, \sigma) \geq 0 \iff \alpha \cdot t_n d(\mathbf{x}_n, \sigma) \geq 0.$$

We will now pick our α – as typical – in a way such that the closest point to the boundary is of distance 1, i.e. we ensure that

$$\alpha t_n d(\mathbf{x}_n, \sigma) \geq 1,$$

hence all points being at a minimum of distance 1. We can also write this as

$$\alpha \cdot t_n(||\mathbf{x}||_2 - \sigma f(\mathbf{x})) \geq 1 \iff t_n(\alpha||\mathbf{x}||_2 - \beta f(\mathbf{x})) \geq 1,$$

where we define $\beta := \alpha\sigma$.

By $\alpha t_n d(\mathbf{x}_n, \sigma) \geq 1$ and $|t_n| = 1$, clearly maximizing the boundary corresponds to having as small a value for α, as the smaller α is, the larger $d(\mathbf{x}, \sigma)$ has to be to ensure that $\alpha \cdot t_n d(\mathbf{x}_n, \sigma) \geq 1$. As such, we can define a new (convex) problem:

$$\begin{aligned} \arg\min_{\alpha} \quad & \frac{1}{2}\alpha^2 \\ \text{s.t.} \quad & t_n(\alpha||\mathbf{x}_n||_2 - \beta f(\mathbf{x}_n)) \geq 1 \\ & \alpha\beta \geq 0, \end{aligned} \tag{1}$$

for all n.

Notice that ensuring that $\alpha\beta \geq 0$, implies that $\sigma = \frac{\beta}{\alpha} \geq 0$, i.e. our scaling factor is indeed positive.

Before we start Please notice that even though the setup for this question is very elaborate, in the end it really resembles that standard SVM question a lot, try and compare to our standard object:

$$\begin{aligned} \arg\min_{\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{s.t.} \quad & t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \end{aligned} \tag{2}$$

13a To find the size of the boundary, we can simply look at our definition. Let's just pick some positively classified point $\mathbf{x}, t$ which is exactly at our scaled-signed distance 1 (per definition, of course, we have this point as we build our boundary in this way), i.e. we have that

$\alpha t d(\mathbf{x}, \sigma) = 1$.

Now, since $t = 1$, we find that $d(\mathbf{x}, \sigma) = \frac{1}{\alpha}$. Of course, this answer makes sense as we exactly chose α to be the number that scales $d(\mathbf{x}, \sigma)$ to be 1. That was not too bad!

13b Now, we introduce some slack variables. This means that for each datapoint, we will add some slack variable ξ_n which allows the model to be more flexible with the boundary, i.e. we now simply will ensure that

$$t_n(\alpha||\mathbf{x}_n||_2 - \beta f(\mathbf{x}_n)) \geq 1 - \xi_n.$$

We know that our objective will change, i.e. we will now minimize the old thing plus a penality for the slack variables. Lastly, we need to ensure as always that the slack variables are positive, giving us:

$$\begin{aligned} \arg\min_{\alpha,\{\xi_n\}} \quad & \frac{1}{2}\alpha^2 + C\sum_n \xi_n \\ \text{s.t.} \quad & t_n(\alpha||\mathbf{x}_n||_2 - \beta f(\mathbf{x}_n)) \geq 1 - \xi_n, \\ & \alpha\beta \geq 0 \\ & \xi_n \geq 0 \end{aligned} \tag{3}$$

for all n.

13c If you get this far, you can really go on auto-pilot mode when writing down the lagrangian. A bit of silly advice, but just don't mess it up by being super careful. We do a minimization procedure, so we subtract the constraints. It really isn't so hard, you just copy the function you minimize $f(\alpha, \{\xi_n\})$, and every time you have some constrain $g(\alpha, \{\xi_n\}) \geq c$, we subtract $\lambda(g(\alpha, \{\xi_n\}) - c)$ for some Lagrange multiplier λ.

We have three types of constraints, we have N Lagrange multipliers for the first one, then simply 1 Lagrange multiplier for the second one, and again N Lagrange multiplier for the third one, say $\{\lambda_n\}, \gamma, \{\mu_n\}$ respectively. I will swap constraints 2 and 3 giving:

$$
\begin{aligned}
\mathcal{L}(\alpha, \{\xi_n\}, \{\lambda_n\}, \{\mu_n\}, \gamma) = &\frac{1}{2}\alpha^2 + C\sum_n \xi_n \\
&- \sum_n \lambda_n(t_n(\alpha||\mathbf{x}_n|| - \beta f(\mathbf{x}_n)) - 1 + \xi_n) \\
&- \sum_n \mu_n \xi_n \\
&- \gamma\alpha\beta.
\end{aligned}
$$

13d Now, how many KKT conditions do we get from this? Again, let's just go per constraint.

For the constraint that $\sum_n \lambda_n(t_n(\alpha||\mathbf{x}_n|| - \beta f(\mathbf{x}_n)) - 1 + \xi_n)$, we get $3N$ conditions, i.e. for each n we have that

- $\lambda_n \geq 0$
- $t_n(\alpha||\mathbf{x}_n|| - \beta f(\mathbf{x}_n)) - 1 + \xi_n) \geq 0$
- $\lambda_n(t_n(\alpha||\mathbf{x}_n|| - \beta f(\mathbf{x}_n)) - 1 + \xi_n) = 0$

Of course, similarly we will have $3N$ constraint for $\sum_n \mu_n \xi_n$:

- $\xi_n \geq 0$
- $\mu_n \geq 0$
- $\xi_n \mu_n = 0$

Last, we have the constraint $\gamma\alpha\beta$, giving us

- $\alpha\beta \geq 0$
- $\gamma \geq 0$
- $\gamma\alpha\beta = 0$

Hence, in total we have $3N + 3N + 3 = 6N + 3$ constraints.

13e Well, what are our primal variables, i.e. the non-Langrian multiplier guys? Here, it is α, β, and our $\{\xi_n\}$. We see that

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha} = 0 &\iff \alpha - \sum_n [\lambda_n t_n ||\mathbf{x}_n||] - \gamma\beta = 0 \\
&\iff \alpha = \sum_n \lambda_n t_n ||\mathbf{x}_n|| + \gamma\beta.
\end{aligned}
$$

Similarly, we see that

$$
\frac{\partial \mathcal{L}}{\partial \beta} = 0 \iff \sum_n [\lambda_n t_n f(\mathbf{x}_n)] - \gamma\alpha = 0
$$

and

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = 0 \iff C - \lambda_n \mu_n = 0$$
$$\iff C = \lambda_n + \mu_n$$

As you can see, even though the Lagrangian is a bit big, the derivates are actually quite simple.

13f It's not so hard to see that the first two options don't make much sense. If $\beta = 0$, we'd have $\sigma = 0$ and hence everything would be $+1$. Also, if $\beta, \gamma \geq 0$, it must be (by complementary slackness that $\gamma\alpha\beta = 0$) that $\alpha = 0$, and thus that $\sigma = \frac{\beta}{\alpha}$ tend to infinity, making everything be -1.

13g If we write out $\tilde{\mathcal{L}}$, it's not so hard to see that the way in which any two points $\mathbf{x}_n, \mathbf{x}_m$ as 'multiplied' is by $||\mathbf{x}_n||||\mathbf{x}_m||$, hence forming our kernel $k(\mathbf{x}_n, \mathbf{x}_m) = ||\mathbf{x}_n||||\mathbf{x}_m||$.

13h So our kernel does not depend on the shape M. Does this make sense? Well, yeah, it does in the sense that a kernel should only measure similarity, and then the can use a specific problem setting – i.e. with our butterfly boundary – to **use** this similarity. They should really be seen as two different things. Of course, M still affects our constraints in the optimization, so it still affects our solution.

Final words As you can see, this question was not as bad as it looked! Just stay calm and reapply the knowledge you have about SVMs. Cheers :D