CS181 Practice Questions: Max-Margin Classification

When maximizing the margin, we seek to learn linear functions of the form

$$f(\boldsymbol{x}, \boldsymbol{w}, b) = \boldsymbol{w}^{\top} \boldsymbol{\phi}(\boldsymbol{x}) + b$$

where $\boldsymbol{w}$ is an M-dimensional column vector of weights and $\boldsymbol{\phi}(\boldsymbol{x})$ is a collection of feature maps (like the regression and neural network case). The training data set comprise of N input vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ with the corresponding target labels $t_1, t_2, \ldots, t_N$, where $t_n \in \{-1, +1\}$.

1. **Computing the Margin**

   What is the perpendicular distance from a data point $\mathbf{x_n}$ to the decision boundary $y_w(x)$?

$$\frac{t_n(\boldsymbol{w}^{\top} \boldsymbol{\phi}(\boldsymbol{x}_n) + b)}{||\boldsymbol{w}||}$$

2. **Basic Maximization Problem**

What is the optimization problem we write for maximizing the margin?

$$\boldsymbol{w}^\star, b^\star = \arg\max_{\boldsymbol{w}, b} \left\{ \frac{1}{||\boldsymbol{w}||} \min_n \left[t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) \right] \right\}$$

3. **Constrained Minimization**

What is the corresponding constrained quadratic minimization problem for maximizing the margin?

$$\boldsymbol{w}^\star, b^\star = \arg\min_{w,b} \left\{ \frac{1}{2} ||\boldsymbol{w}||_2^2 \right\} \text{ s.t. } t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) \geq 1$$

4. **Equivalence**

Explain (at a high level) why the constrained quadratic minimization of question (3) is equivalent to the unconstrained maximization in (2)

From above, we have that the max-margin problem is given by

$$\boldsymbol{w}^\star, b^\star = \arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{||\boldsymbol{w}||} \min_n \left[t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) \right] \right\}$$

and that the corresponding minimization problem is

$$\boldsymbol{w}^\star, b^\star = \arg\min_{\boldsymbol{w},b} \left\{ \frac{1}{2} ||\boldsymbol{w}||_2^2 \right\} \text{ s.t. } t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) \geq 1$$

Minimizing $\frac{1}{2}||\boldsymbol{w}||^2$ is equivalent to maximizing $\frac{1}{||\boldsymbol{w}||}$ because $||\boldsymbol{w}|| \geq 0$. Furthermore, we note that the normalized orthogonal distance from a point to the decision boundary is invariant under scalar multiplication. To see this, we have

$$\frac{t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b)}{||\boldsymbol{w}||^2} = \frac{\beta}{\beta} \cdot \frac{t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b)}{||\boldsymbol{w}||^2} = \frac{t_n(\beta \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + (\beta b))}{||\beta \boldsymbol{w}||^2}$$

Thus, since the data is linearly separable, so there exists a decision boundary with non-zero, positive margin for each example, we do not lose any generality in imposing the restraint $t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) \geq 1$ (because we can just scale $\boldsymbol{w}$ until the minimal value is ≥ 1).

5. **Tightness**

What happens to the inequalities $t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) \geq 1$ for the optimal solution?

At the optimal solution, some of the inequalities must be tight, i.e., $t_n(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n) + b) = 1$. If it wasn't, we could always choose a $\boldsymbol{w}$ with a smaller norm that still satisfied the constraint.

6. **Kernels**

What is kernel function?

A kernel function is a scalar product on two vectors mapped by basis functions into a feature space. In general, we use kernels to map into higher dimensional feature spaces, using them to circumvent costly computations in high dimension spaces.